PIER
PENN INSTITUTE for ECONOMIC RESEARCH
UNIVERSITY of PENNSYLVANIA

The Ronald O. Perelman Center for Political
Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297

pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper

22-030

Learning Through Repetition? A Dynamic Evaluation of Grade Retention in Portugal

EMILIO BORGHESAN
University of Pennsylvania

HUGO REIS
Banco de Portugal
Catolica Lisbon SBE and IZA

PETRA E. TODD
University of Pennsylvania
NBER, HCEO and IZA

November 11, 2022

Learning Through Repetition? A Dynamic Evaluation of Grade Retention in Portugal

Emilio Borghesan, Hugo Reis, and Petra E. Todd [1]

November 11, 2022

[1] Emilio Borghesan is a graduate student in the University of Pennsylvania PhD program. Hugo Reis is a research economist at the Banco de Portugal and a Professor of Economics at Catolica Lisbon SBE and a member of IZA. Petra Todd is Edmund J. and Louise W. Kahn Term Professor of Economics at the University of Pennsylvania and a member of the U Penn Population Studies Group, NBER, HCEO and IZA. This paper was presented in seminars at U Penn and Brandeis University and as a keynote lecture at the Brazilian Econometric Society meetings and the Barcelona School of Economics Summer Forum. We thank participants for many helpful comments. Hugo Reis acknowledges the financial support of Fundação para a Ciência e Tecnologia through the project number PTDC/EGE-ECO/6055/2020.

Abstract

High rates of grade retention are a matter of much controversy and debate worldwide. Although some students may learn more with extended classroom time, other students get discouraged and drop out of school. This paper develops and implements a dynamic value-added modeling approach for estimating grade retention effects in Portuguese high schools where over 40% of students were retained. The statistical model is derived from an education production function that describes how knowledge cumulates with sequential years of school attendance, including repeated grades. Model parameters are obtained using simulated method of moments applied to nationwide administrative test score data. The estimated model is used to simulate achievement in math and Portuguese under the existing grade retention and compulsory schooling policies and under alternative policies. Results show that the average impact of the current policy on 12th grade test scores of retained students is positive, 0.2 standard deviations in math and 0.5 s.d. in Portuguese. However, we find that the test score impacts are heterogeneous and roughly one third of students experience learning loss. Retention also significantly increases school dropout, especially for male youth and older students. We compute policy-relevant treatment effects for retention's effects on lifetime earnings, taking into account retention's simultaneous effects on educational attainment, knowledge, and age of labor market entry, and we solve for the optimal retention policy that maximizes average lifetime earnings in the population.

1 Introduction

Grade retention patterns differ markedly across countries. In most Scandinavian countries and also in Korea, Japan and Malaysia, grade repetition is very rare or in some cases not allowed (e.g. Norway, Iceland). However, in other countries, such as the Netherlands, France, Portugal and Brazil, repetition is common and more than 20% of students repeat at least one grade by age 15.[1] Portugal, the focus of this paper, has some of the highest rates of grade retention in the world. More than 40% of students in the cohort we study – students who started 9th grade in 2008/2009 – are held back at least one year in high school alone. Despite the vast differences in grade retention policies across countries and their possible effects on schooling and labor market outcomes, there is little research that considers the question of optimal retention policy design, as we do in this paper.

Whether grade retention provides benefits for lower achieving students is a matter of much controversy and debate, with some governments passing laws that aim to preempt schools' ability to retain students (Colombia, for example). Proponents of grade retention argue that the practice provides students with the opportunity to master the curriculum before moving on to more advanced material. Under this view, academic achievement is a cumulative process and mastering the material in one grade facilitates learning in the next. On the other hand, retained students may be stigmatized, have a hard time adjusting to a new peer group and suffer from reduced self-esteem. If students get discouraged, then high retention rates, particularly in secondary school, could increase dropout. Grade retention policies are also expensive, because they entail additional per pupil expenditures as well as opportunity costs of delayed labor market entry for students. For these reasons, it is important to understand whether the benefits outweigh the costs.

A large literature aims to estimate grade retention effects on retained students in elementary school, middle school, and high school. Randomized control trials are not feasible in this context, so the existing research is based on observational data, which poses two distinct methodological challenges. The first is that retained students tend to perform less well than their peers and are often negatively selected on both observable and unobservable dimensions, such as ability, socioe-

[1] For example, based on the 2015 PISA student questionnaire, the percentage of 15 year-old students who had been retained were: 7.1% in Australia, 11.0% in the United States, 20.1% in the Netherlands, 22.1% in France, 31.2% in Portugal, 31.3% in Spain, and 34% in Belgium.

conomic background, intrinsic motivation, and emotional maturity. The second is that retention may increase the likelihood of dropout, causing dynamic sample selection bias in comparisons of retained and nonretained students at different grades. Controlling for the endogeneity of the retention decision and nonrandom selection is requisite to infer the causal effects of retention on educational attainment and academic achievement.

Research studies have addressed these challenges using a variety of methodological approaches, including regression-discontinuity, instrumental variables, control function, factor-analytic methods, difference-in-differences, and matching-on-observables approaches. Sometimes retained students are compared to non-retained students of the same age and, at other times, they are compared to students of the same grade but of different ages. The empirical evidence reported in the literature on whether grade retention is harmful or beneficial is mixed. Holmes and Matthews (1984) and Jimerson (2001) present two different meta-analyses that focus on the frequency of positive and negative estimated effects across studies without accounting for study differences in research designs. Both papers conclude that the preponderance of the evidence is that grade retention is harmful. Allen et al. (2009) carry out a meta-analysis of 22 studies that they deem to have well-matched comparison groups. They explore how estimated retention effect sizes vary with the study design quality, with the grade in which the student is retained and with the number of grades since retention. Their findings challenge the view that grade retention is harmful. We describe how this study relates to and builds on the extensive literature in section two.

The goals of this paper are (i) to develop and estimate a dynamically linked value-added model for analyzing the effects of grade retention on academic achievement, (ii) to use the estimated model to evaluate the mean and distributional effects of grade retention in Portuguese high schools, and (iii) to explore questions related to optimal retention policy design. Our model aims to capture how learning in one grade depends on the extent of learning in prior grades, including any repeated grades, as well as on school and family investments. Specifically, we specify a model where knowledge in a given year and subject (e.g. 12th grade math) depends on the previous year's knowledge, intervening family and school inputs, unobserved factors and random shocks.[2]

[2] See Todd and Wolpin (2003) for discussion of methods for estimating educational production functions and

There are several key differences between our modeling approach and existing approaches (described in detail in section two). One is that we analyze retention effects within a value-added framework that controls for lagged test scores. Our focus is on academic achievement for youth attending high school, with math and Portuguese test scores at the end of middle school (in ninth grade) serving as initial conditions. The initial test scores along with family background characteristics are highly predictive of subsequent high school test score performance. Our modeling framework also allows for unobservable types that may jointly affect test scores, dropout and grade retention, although we find that selection into retention is primarily on the basis of observable characteristics.[3] A second key difference between our framework and existing ones is that we introduce a separate achievement production function for the learning that takes place when students take a grade for the second time. Students who repeat grades have multiple test score observations in the same grade, and we explicitly model the dynamic process by which students accumulate additional knowledge when they repeat a grade. Lastly, we analyze the effects of retention policies on three outcomes - test scores, dropout, and earnings. Much of the retention literature focuses on children in kindergarten or primary school when retention-induced dropout is unlikely. With older youth, dropout and its implications for educational attainment and labor market earnings potential are important factors that need to be considered.

As previously noted, retained students differ in observable dimensions and there is typically a negative selection pattern. Our model incorporates rich observed heterogeneity by allowing student gender and family demographics to affect schooling preferences as well as the achievement production function. The model also includes observable dimensions of school quality, such as class size, school size, teacher age and gender. Lastly, it incorporates unobserved heterogeneity in the form of unobserved discrete multinomial types.[4] As discussed in Todd and Wolpin (2003) and Rivkin et al. (2005), in estimating value-added models, it is important to control for inherent student abilities, corresponding to cognitive ability, personality traits, or motivation, that may

assumptions that underlie value-added models.

[3] In the context of grade retention, self-selection is not a key concern as students do not usually desire to be retained.

[4] The inclusion of discrete unobserved types is common in the dynamic discrete choice literature, see, e.g., Eckstein and Wolpin (1999), Keane and Wolpin (1997), Arcidiacono et al. (2007).

affect a child's achievement growth. These types enter multiple model equations to control for unobservable attributes and, in doing so, allow for correlated error structures across equations.

We analyze a large administrative database from the Ministry of Education in Portugal that contains information on annual school enrollment and standardized test scores in math and Portuguese for approximately fifty thousand students for the years 2008-2013. Enrollment and grade retention are measured in every year, but the standardized tests are high-stakes tests that are only administered in grades 9 and 12.[5] If a student is retained in 9th or 12th grade, we observe multiple test scores corresponding to the different years when the student took the same-grade test. As described below, our model and estimation approach accommodates the fact that some students have multiple test scores in the same grade as well as the fact that enrollment, dropout, and retention are observed annually whereas test scores are observed only in the years and grades when students take the tests. Educational researchers commonly encounter such data complications, because standardized tests are often not available for every grade. Our estimation strategy addresses this differential timing in a way that is consistent with our theoretical knowledge accumulation model. Our estimation strategy derives moments from the model that correspond to the mean and variance of test scores for subgroups of students with different schooling trajectories. These moments form the basis for efficient parameter estimation via Simulated Method of Moments (SMM).

The estimated model is used to study how grade retention affects learning in math and Portuguese, as reflected in test scores, as well as dropout behavior. The analysis yields a number of insights about grade retention impacts and provides guidance as to how Portugal's current retention policy could be modified to achieve greater benefits. First, we find that the current policy increases 12th grade test scores on average by about 0.20 standard deviations (s.d.) in math and about 0.50 s.d. in Portuguese for students experiencing retention. Second, the largest grade retention benefits are observed for lower performing students. Third, grade retention significantly increases dropout, particularly for male youth and for youth that are older than their peers at the time of entering high school. Also, retention in the 10th grade (the first grade in high school) increases dropout more than retention in later grades. Fourth, retention impacts are heterogeneous, with more than 30%

[5] The 9th grade test is partly used as a basis for deciding whether the student passes the grade. The 12th grade test score is used both as a basis for completing high school as well as a college entrance examination.

of students experiencing negative test score impacts. In analysis similar to the marginal treatment effects (MTE) approach of Heckman and Vytlacil (2005) and Heckman et al. (2006b), we estimate we estimate several nonparametric regressions of treatment effects on the probability of being retained, where the outcomes in the treated and untreated states are obtained from dynamic model simulations. We find that the current retention policy targets students with the highest potential test score benefits but that these students also experience the greatest increase in dropout risk.

Lastly, we perform a cost-benefit analysis to assess whether the lifetime earnings benefits of grade retention exceed the costs, incorporating the estimated heterogeneous impacts, namely that grade retention induces earlier dropout for some students but increases cognitive skills for others who stay in school. The cost-benefit analysis trades off the benefits of increased skill with the costs of reduced educational attainment and delayed labor market entry. Another relevant consideration is that grade retention policies interact with compulsory schooling policies, because prohibiting dropout mitigates some of the negative grade retention impacts. In 2009, Portugal increased its compulsory schooling age from 15 to 18, which affected cohorts that entered high school a few years after the one we study.

Using our estimated model, we evaluate how retention policies together with compulsory schooling laws affect each student's skill level, educational attainment, and predicted lifetime earnings. Predicted earnings are obtained from Mincer earnings regressions estimated on a matched employee-employer data set (the *Quadros de Pessoal*).[6] We then compute average policy relevant treatment effects (PRTE) corresponding to a range of counterfactual policies that reduce the percentage of students retained.[7] We find that the actual policy experienced by the 2008 cohort that we analyze, which was a 40% retention rate, passes a cost-benefit test only if the labor market returns to math and language skills are high (over 13%).

The paper develops as follows. Section two describes how our study relates to and builds on the extensive grade retention literature. Section three provides background on the Portuguese education system. Section four develops the modeling framework, and section five discusses esti-

[6] The *Quadros de Pessoal* data are used frequently in analysis of the Portuguese labor market (Campos and Reis (2017)).

[7] See Heckman and Vytlacil (2001) and, more recently, Mogstad et al. (2018) for discussion of PRTE parameters.

mation. Section six describes the data and estimated model parameters. Section seven presents simulation-based estimates of retention impacts on test scores, dropout and educational attainment, and provides evidence on impact heterogeneity. Section eight performs cost-benefit analysis. Section nine concludes.

2 Related literature

Studies in the literature use a variety of methods to evaluate the effects of grade retention in primary, lower secondary, and upper secondary grade levels. Below, we focus on more recent studies, grouped by the methodologies used to control for potential endogeneity bias. Existing studies typically do not control for dynamic selection bias arising from school drop out, which is a key concern with older-age youth. Also, most studies focus on test score and education impacts and do not consider impacts on lifetime earnings streams and/or the question of optimal retention policy design. The methods we develop and implement in this paper address these concerns, which are particularly important in the context of high school retention.

Several studies in the literature use regression-discontinuity (RD) designs, exploiting discontinuities in the rules determining which children are retained. For example, Jacob and Lefgren (2004) and Jacob and Lefgren (2009) use a RD estimator that exploits the implementation of an accountability policy in Chicago Public Schools. They find a modest benefit of third grade retention on achievement scores but no effect of sixth grade retention. When Jacob and Lefgren (2009) look at longer-term impacts, they find that eighth grade retention decreases high school completion but earlier retention does not have this effect. Manacorda (2012) analyzes grade retention effects in junior high school using administrative data from Uruguay and also exploiting a discontinuity in the retention rules. He shows that grade failure leads to an increase in dropout in the year the retention occurs. Eren et al. (2017) use an RD design to analyze the net effects of summer school remediation and test-based promotion policies in Louisiana on high school completion and juvenile crime. For eighth grade students, they find that retention decreases crime but increases dropout.[8] A well known limitation of RD designs is that they identify retention impacts only for

[8] There are several papers using an RD approach to analyze retention in earlier grades. Schwerdt et al. (2017)

the subgroup of children near/at the margin of being retained. In section 7 of this paper, we show that the retention impacts in Portuguese high schools are larger for students with a high probability of retention than for the average student or for students near the margin of being retained.

Another group of studies use instrumental variables (IV) estimators to analyze the effects of grade retention, usually within a regression framework. When impacts are heterogeneous across students, IV estimates are interpretable as a local average treatment effect (LATE), which is the average impact of retention for children who were retained because of the value of the instrument (the so-called complier group).[9] An early study by Eide and Showalter (2001) uses the High School and Beyond data set to analyze the effects of high school grade retention on dropout and on subsequent earnings. The instruments are derived from kindergarten school entry rules and correspond to various functions of the difference between the child's birthday and the cut-off for starting kindergarten. Their OLS estimates suggest that grade retention increases the probability of dropping out of school and negatively affects earnings, but their IV estimates tend not to be statistically significantly different from zero. Pereira and Reis (2014) and Garcia-Pérez et al. (2014) study the impact of grade retention in Portugal and Spain using the PISA data set, also using birth date as an instrumental variable. Pereira and Reis (2014) find that grade retention has small positive impacts on educational outcomes in the short term, while Garcia-Pérez et al. (2014) find negative effects. Lastly, Gary-Bobo et al. (2016) use quarter-of-birth as an instrument to estimate the effects of grade retention in French junior high schools. They find that the IV estimates are not robust and instead adopt a factor analytical model (described below).

Another branch of literature uses factor analytic dynamic models (FADM) to analyze grade

exploit a discontinuity in retention probabilities under Florida's test-based promotion policy to study third grade retention impacts on student outcomes through high school. They find large positive effects on achievement through grade 10 but that the effects fade out if students are compared to same-age peers. They also find that third grade retention increases students' grade point averages and leads them to take fewer remedial courses in high school but does not affect students' graduation probability. Figlio and Özek (2020) use an RD design to study effects of third grade retention plus instructional support on English language learners in 12 Florida school districts. They find that grade retention increases language proficiency and increases the likelihood that students take more advanced coursework in middle school and high school. Winters and Greene (2012) also use a RD strategy to study the effects of remediation and grade retention in Florida. Students who were retained in the third grade were first required to attend summer school and then were assigned to a high quality teacher during the retention year. The study finds a statistically significant positive impact on student achievement in math, reading, and science that is sustained for some years after the treatment but then dissipates.

[9] See Angrist and Imbens (1995).

retention impacts.[10] For example, Fruehwirth et al. (2016) analyze the effects of grade retention in kindergarten and other elementary grades using ECLS-K data. The authors develop a potential outcomes framework in which retention's effects can vary with the grade in which the student was retained as well as the number of years since retention. Along with the outcomes model, they specify a probabilistic grade retention equation, where a low-dimensional set of unobservable factors can affect both the outcome and retention equations. Their impact analysis compares retained and nonretained children at the same age, which is possible because the ECLS-K tests are comparable across years. As they note, with age held constant, retained children were exposed to less curricula. They find that grade retention has significant negative effects on retained children. Saltiel and Sarzosa (2020) also use the ECLS-K data to analyze the effects of grade retention, focusing on retention in kindergarten and first grade. They estimate a dynamic model of cognitive and noncognitive skill formation, which incorporates children's latent abilities, parental skills, and investment choices.[11] Their results show that retention lowers cognitive skills for retained students, slightly increases noncognitive skills (by 0.02 s.d.) and increases parental investments.[12]

Gary-Bobo et al. (2016) use a factor analytic framework to study the effects of grade retention for French junior high school students, grades 6-9. They find on average significant positive but small effects of retention on the test scores of those who are retained but negative impacts on the probability of completing grade 9. Cockx et al. (2019) develop and estimate a FADM to estimate retention effects in Flemish secondary schools on test scores, dropout, downgrading of schooling track and delayed graduation. They find neutral effects on short-term academic achievement but longer-term adverse effects on the other schooling outcomes, particularly for less able students. The framework we develop differs from the FADM models in the literature in its use of dynamically linked value-added models, its incorporation of a switching regression to capture differences in the learning process during retention years, and in providing an estimation method that addresses the

[10] Early discussion of these methods include Carneiro et al. (2003) and Heckman and Navarro (2007)

[11] They build on previous work developing such models (e.g. Cunha et al. (2006), Cunha et al. (2010), Agostinelli and Wiswall (2016)).

[12] A study by Dong (2010) also uses the ECLS-K to analyze the impact of kindergarten grade retention. She implements a control function estimator that jointly models the choice of enrolling in a school that allows kindergarten retention, the decision for a child to repeat kindergarten, and academic performance in subsequent grades (through grade three). She finds that retention improves academic performance, but the positive effects diminish from 1st to the 3rd grades. Her study compares retained and non-retained students holding grade constant.

common problem of test scores only being observed for a subset of grades.

A few studies exploit policy changes to analyze retention impacts. For example, Ferreira et al. (2018) study the effects of a 2010 policy change in Colombia that allowed schools to increase their retention rates above 5%, which led to a positive impact on Spanish test scores but no effect on math scores. Battistin and Schizzerotto (2012) study a remedial education reform in Italy that changed the promotion criteria in upper secondary schools. They exploit geographical variation in the reform's implementation and find heterogeneous impacts, with students in lower educational tracks experiencing negative impacts.

3 The Portuguese education system

3.1 Organization and governance

The Ministry of Education defines, coordinates, and implements national education policies at the pre-school, basic, and upper secondary levels. The school network is divided into regional school clusters, which have some autonomy in terms of pedagogy, scheduling, teaching, and managing non-teaching staff. The public education system is divided in pre-school education (from ages three to five), basic education (grades 1 to 9), and upper secondary education (grades 10 to 12). Public education is free and universally available from the age of five (the final year of pre-school). For the cohort of students that we study, school attendance was mandatory for nine years or until age 15. After 2009, upper secondary school was made mandatory (up through age 18) with the passage of Law no. 85/2009, but the law did not take effect immediately. The first affected cohort were students who entered high school in 2012/2013, three years after the cohort we study.

Basic education has a common curriculum, with the goal of providing a general educational background. It is divided into three cycles. The first cycle corresponds to the first four years (grades one to four) and the second cycle to the next two years (grades five and six). The third cycle corresponds to lower secondary education (grades seven to nine).

Upper secondary education lasts for three years and provides students with different pathways to match their vocational interests and/or to prepare them for post-secondary studies. Admission to alternative secondary tracks is open to everyone, under the view that all individuals should be able to choose from all of the educational and training options available. Students may choose from three tracks: a) science-humanities courses; b) vocational courses; c) other education and training. Track (a) is most oriented towards further studies, although other tracks also offer dual certification (academic and vocational). Access to post-secondary education in Portugal is through competitive national exams taken in the 12th grade. In principle, students who attend alternative tracks can take the exams and obtain access to higher education. The administrative data that we analyze pertain to students pursuing the science and humanities track, which includes four subtracks with different course and examination requirements: science and technology, socio-economic science, languages and humanities, and visual arts.

3.2 Assessment

Evaluation in the Portuguese educational system is based on both course grades and external national exams. At the end of basic education (9th grade), students take national exams in Portuguese and math. At the end of upper secondary education (12th grade), students in the science-humanities course track take four national exams.[13] Portuguese is a required subject for all students; however, math is only a required subject for 12th grade students in the science and technology and socio-economic science subtracks. The other required exams vary depending on the secondary track pursued and the requirements of universities for admission to different majors/programs.

We focus our analysis on learning in math and Portuguese courses and pool students in the science and technology and socio-economic science subtracks into a single category, which we call STEM. Students in the languages and humanities as well as visual arts subtracks are grouped together into a single non-STEM category. Students in STEM constitute 76% of our sample.

3.3 Retention in upper secondary school

Retention in Portugal depends on both course marks and national exam scores. At the end of each academic year, students receive a mark in each subject ranging from 1 to 20. In grades 10 and 11, a student fails in a specific subject if the mark is below 10, and failure results in retention. At

[13] A small number of courses, not including math or Portuguese, allow students to take final exams following 11th grade.

the end of 12th grade, students take the national examinations. In subjects that are covered by these national exams, the final course mark is calculated as a weighted average of the internal mark (70%) and the exam score (30%), which is rescaled to be between 0 and 20. A student is retained in a 12th grade course if this weighted average is below 10.

One reason why retention is so common is that the national examinations at the end of high school serve the dual purpose of high school exit exams and college entrance exams. They are often more challenging than the assessments students complete in their high school courses. The average student fails the national math exam and narrowly passes the Portuguese exam (as described in Table 3). The only factor keeping retention rates from being even higher is that internal grades frequently raise students' weighted average above the minimum passing threshold.

Our data set includes information on exam scores, dropping out, retention and some limited information on course grades. We do not incorporate grades in our analysis for two main reasons. First, the grades are not available in 10th and 11th grades. Second, grading criteria vary across schools and across teachers within a school, and students take different sets of courses depending on their course track. We focus instead on the administrative test score outcomes, because the tests are comparable across all students. The tests are also highly relevant to students' future success, because they are the primary determinants of admission to universities and admission to majors within universities.

4 Model

4.1 The value-added model

In this section, we exposit the value-added model of knowledge acquisition in two subjects (math and Portuguese) that we estimate. We first describe how the model can be derived from a general cumulative educational production function. Let K_{ia} denote knowledge of individual i at age a. Let μ_{i0} denote the individual's unobserved endowment. We assume that a young person's accumulated knowledge depends on the history of family inputs, school inputs, and on the endowment. Let $F_i(a)$ denote family inputs applied up until age a and $S_i(a)$ school inputs applied through age a.

A general specification for knowledge at any age a can be written as

$$K_{ia} = K_a[F_i(a), S_i(a), \mu_{i0}, \xi_{ia}] \tag{1}$$

where ξ_{ia} represents random components (such as a random illness the day of the test). Let $I_i(a)$ denote the vector of family and school inputs, $I_i(a) = (F_i(a), S_i(a))$. A linear regression analogue of the previous equation can be written as

$$K_{ia} = I_{ia}\alpha_1 + I_{ia-1}\alpha_2 + I_{ia-2}\alpha_3 + ... + I_{i1}\alpha_a + \beta_a\mu_{i0} + \xi_{ia} \tag{2}$$

As discussed in Todd and Wolpin (2003), the major challenges in estimating educational production function models are that the entire history of family and school inputs is almost never observed and endowments are not observed. However, it is possible to impose some restrictions on the cumulative model coefficients to derive an implementable specification. For example, we obtain a value-added model if we multiply K_{ia-1} by γ and subtract:

$$K_{ia} - \gamma K_{ia-1} = I_{ia}\alpha_1 + I_{ia-1}(\alpha_2 - \gamma\alpha_1) + ... + I_{i1}(\alpha_a - \gamma\alpha_{a-1}) + (\beta_a - \gamma\beta_{a-1})\mu_{i0} + \xi_{ia} - \gamma\xi_{ia-1}$$

If we further assume that $\alpha_a = \gamma\alpha_{a-1}$, then only the lagged test score, contemporaneous input measures, and unobserved endowment terms remain in the equation.[14]

4.2 Incorporating multiple grades and grade retention

In the data, we observe student-level school enrollment patterns in grades 9-12, including any grade retentions, and standardized test scores for national tests administered in grades 9 and 12. As previously described, these are high stakes tests. The 9th grade tests, in conjunction with course grades, are used to determine whether a student passes the grade. The 12th grade test is also used, in part, to determine passing and, in addition, is one of the criteria colleges use in deciding whether to admit students and in determining the set of college majors for which the student is eligible.

[14] This assumption implies that the input coefficients decline geometrically with distance, as measured by age, from the achievement measurement and the rate of decline is the same across input measures. If one is willing to impose a similar assumption on μ_{i0}, namely that $\beta_a = \gamma\beta_{a-1}$, then we obtain a value-added model of the kind commonly estimated in the literature on test score determinants:

$$K_{ia} = \gamma K_{ia-1} + I_{ia}\alpha_1 + \xi_{ia} - \gamma\xi_{ia-1}$$

Without this assumption, the value-added equation includes an unobserved endowment effect.

Our school enrollment model begins at the start of high school (10th grade). Students can begin high school at different ages, depending on the age at which they started kindergarten and on whether they were retained in prior grades. We describe the academic achievement model in terms of time t rather than age, although age evolves deterministically with time. In the empirical analysis, we will control for age variation across students within a grade.

Let K_t^G denote the knowledge measure at grade $G \in \{9, 10, 11, 12\}$ in time t. Our sample consists of all students who enter grade 10 after completing grade 9. Students can pass or fail (be retained) in grades 10, 11, and 12. They can also drop out of school in any grade. Table 1 describes the different possible high school trajectories and the knowledge measures associated with each trajectory for students that only fail the same grade at most once. Additional paths in which students fail the same grade multiple times are relatively infrequent in our data and are not shown in Table 1 (but are used in estimation). The third column lists the knowledge measures that are observed in the data for each high school trajectory. These observed knowledge measures are a subset of all the potential grade-specific knowledge measures, because the national exams are only taken in the 9th and 12th grades, and we observe test scores over a period of five years.[15]

Knowledge in a particular year in a certain subject (math or Portuguese) depends on the knowledge level in the prior year and the levels of intervening family and school inputs. Let I_t^g denote the vector of inputs (family and school) applied at time t in grade g. In theory, inputs into the achievement production function could be modeled as endogenous.[16] However, the family inputs that we use in our analysis, parental education and socioeconomic status, vary little over time. School inputs also vary little in response to grade retention. Students who repeat a grade experience an average decline in class size of 0.1 students in comparison to a 0.7 decrease in class size for promoted students. Similarly, the year-to-year change in teacher age is 0.40 (0.49) years increase in math (Portuguese) for retained students, while that for promoted students is 0.80 (0.90). School size and the gender of the teacher differ minimally over time for both groups of students. The fact that school inputs are very similar for retained and promoted students is consistent with

[15] We observe individual's test scores over a five-year period starting in 2008/2009. In the sixth year, we see whether the student is still enrolled in high school (relevant for students who are retained two or more times), but we do not see their end-of-year scores on the national exams.

[16] See related discussion in Todd and Wolpin (2003), Todd and Wolpin (2007).

Table 1: Common high school trajectories and knowledge measures

Grades in High School	*All Knowledge measures*	*Observed Knowledge measures*
10,11,12	K_t^9, K_{t+1}^{10},K_{t+2}^{11},K_{t+3}^{12}	K_t^9, K_{t+3}^{12}
10,10,11,12	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11},K_{t+4}^{12}	K_t^9, K_{t+4}^{12}
10,11,11,12	K_t^9, K_{t+1}^{10},K_{t+2}^{11},K_{t+3}^{11},K_{t+4}^{12}	K_t^9, K_{t+4}^{12}
10,11,12,12	K_t^9, K_{t+1}^{10},K_{t+2}^{11},K_{t+3}^{12},K_{t+4}^{12}	K_t^9, K_{t+3}^{12},K_{t+4}^{12}
10,10,11,11,12	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11},K_{t+4}^{11},K_{t+5}^{12}	K_t^9
10,11,11,12,12	K_t^9, K_{t+1}^{10},K_{t+2}^{11},K_{t+3}^{11},K_{t+4}^{12},K_{t+5}^{12}	K_t^9,K_{t+4}^{12}
10,10,11,11,12,12	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11},K_{t+4}^{11},K_{t+5}^{12},K_{t+6}^{12}	K_t^9
10,10d	K_t^9, K_{t+1}^{10}	K_t^9
10,11d	K_t^9, K_{t+1}^{10}	K_t^9
10,11,11d	K_t^9, K_{t+1}^{10},K_{t+2}^{11}	K_t^9
10,10,11,11d	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11}	K_t^9
10,11,12d	K_t^9, K_{t+1}^{10},K_{t+2}^{11}	K_t^9
10,10,11,12d	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11}	K_t^9
10,10,11,11,12d	K_t^9, K_{t+1}^{10},K_{t+2}^{10},K_{t+3}^{11},K_{t+4}^{11}	K_t^9

Note: The table lists the knowledge measures available for students with the most common paths in the data. Subscripts refer to the year, while supercripts refer to the grade. Observed knowledge measures correspond to measures in the data.

there being no official policy in Portugal to target additional resources towards retained students.

As described in the introduction, we also incorporate into the model unobserved heterogeneity in the form of discrete types, which are assumed to be known to the individual but not to the econometrician. The types enter the model through the knowledge accumulation equations as well as through dropout and grade retention decisions (as described in detail below), which allows for correlated error structures across equations. The estimation accounts for the fact that retained students tend to be negatively selected by controlling for observable heterogeneity (e.g. initial ninth grade test scores, gender, family background, prior retention history) as well as for unobserved heterogeneity (types). Let $\mu_i^m = 1$ if individual i is type m ($m \in 1...M$), else $\mu_i^m = 0$.

Standardized tests measure mastery of the curriculum in a particular grade. Retained students get tested on the same curriculum twice (the test is comparable although not identical across years). We therefore specify two separate value-added models depending on whether a student advanced to the next grade or was retained to allow the learning process to differ for students seeing the material for the first or second time. Let $R_{i,t}^g = 1$ if student i repeats grade g, else $= 0$. The value added model for a student in grade g who is not retained in the prior year ($R^{g-1}_{i,t-1} = 0$) is given by

$$K^g_{i,t} = \gamma^g K^{g-1}_{i,t-1} + \beta^g I^g_{i,t} + \phi^g X^g_i + \sum_{m=1}^{M} \alpha^m \mu^m_i + \varepsilon^g_{i,t} \ , \tag{3}$$

whereas the equation for a student who is retained ($R^g_{i,t-1} = 1$) is given by

$$K^g_{i,t} = \gamma^{gR} K^g_{i,t-1} + \beta^{gR} I^g_{i,t} + \phi^{gR} X^g_i + \sum_{m=1}^{M} \alpha^{mR} \mu^m_i + \varepsilon^{gR}_{i,t} \ , \tag{4}$$

where X_i denote time-invariant individual characteristics, including the student's gender and their age at secondary school entry relative to the modal age for secondary school entry (e.g. 0 years behind, 1 year behind, etc, which usually reflects retentions prior to high school).[17] In implementation, we estimate equations (3) and (4) jointly for math and Portuguese, as retention and dropout depend on both subjects. For ease of exposition, for now, assume that no one drops out of school but later we will incorporate dropout.

We can substitute the value-added equations for grades 10 and 11 to obtain a model that expresses 12th grade knowledge as a function of 9th grade knowledge. The required equation substitutions depend on the specific high-school trajectory pursued by each student. For example, consider a student who had no retentions and for whom we observe knowledge in 12th grade (time $t+3$) and 9th grade (time t), intervening family and school inputs at grade g, $I^g_{i,t}$, and time-invariant controls, X_i. Through substitution, we obtain

$$\begin{aligned} K^{12}_{i,t+3} = {} & \gamma^{12}\gamma^{11}\gamma^{10} K^9_{i,t} + \gamma^{12}\gamma^{11}\beta^{10} I^{10}_{i,t+1} + \gamma^{12}\beta^{11} I^{11}_{i,t+2} + \beta^{12} I^{12}_{i,t+3} + \gamma^{12}\gamma^{11}\phi^{10} X_i + \gamma^{12}\phi^{11} X_i \\ & + \phi^{12} X_i + \sum_{m=1}^{M} \mu^m_i (\gamma^{12}\gamma^{10}\alpha^{10}_m + \gamma^{12}\alpha^{11}_m + \alpha^{12}_m) + \{\varepsilon^{12}_{i,t+3} + \gamma^{12}\varepsilon^{11}_{i,t+2} + \gamma^{12}\gamma^{11}\varepsilon^{10}_{i,t+1}\} \end{aligned}$$

Now, consider a student who was retained in grade 10 and whose high school grade trajectory is 9, 10, 10, 11, 12. After making the appropriate substitutions, one obtains that knowledge in grade

[17] Our specification assumes additive separability for ease of making the required substitutions to address differential data timing. Alternatively, one could assume a Cobb-Douglas production function model that is additively separable in logs, similar to the kinds of models considered in Todd and Wolpin (2018).

12 (time $t+4$, given this particular trajectory) can be written as

$$\begin{aligned} K_{i,t+4}^{12} = \gamma^{12}\gamma^{11}\gamma^{10R}\gamma^{10}K_{i,t}^{9} + \gamma^{12}\gamma^{11}\gamma^{10R}\beta^{10}I_{i,t+1}^{10} + \gamma^{12}\gamma^{11}\beta^{10R}I_{i,t+2}^{10} + \gamma^{12}\beta^{11}I_{i,t+3}^{11} + \beta^{12}I_{i,t+4}^{12} \\ +\gamma^{12}\gamma^{11}\gamma^{10R}\phi^{10}X_i + \gamma^{12}\gamma^{11}\phi^{10R}X_i + \gamma^{12}\phi^{11}X_i + \phi^{12}X_i + \\ +\{\varepsilon_{i,t+4}^{12} + \gamma^{12}\varepsilon_{i,t+3}^{11} + \gamma^{12}\gamma^{11}\varepsilon_{i,t+2}^{10R} + \gamma^{12}\gamma^{11}\gamma^{10R}\varepsilon_{i,t+1}^{10}\} \end{aligned}$$

Comparing this equation with the previous one shows that a student who attends school longer because of retentions will have additional terms in the value-added model to reflect the additional year(s) of investment. Also, the coefficient on the lagged knowledge measure, $K_{i,t}^{9}$, will differ compared to a student who was not retained. We use these knowledge accumulation equations to form moments of the 12th grade test score distribution for students with different schooling trajectories and estimate the model parameters using SMM.

4.3 Special cases

It is useful to consider how knowledge accumulates for some special cases to explore potential retention effects within this framework. Consider the case where $\gamma^g = 1$ for all g. In that case, prior knowledge does not depreciate from year to year. Students who are retained in grade 10 have an additional year of investment. Assuming the residuals have mean zero and that family and school investments have positive coefficients ($\beta^{10R}I_{i,t+2}^{10} > 0$), the model implies that retention would be beneficial on average (that is, $E(K_{i,t+4}^{12} > K_{i,t+3}^{12})$).

In general, though, we would expect $\gamma^g < 1$, as value-added coefficients on lagged test scores are typically estimated to be less than one. In that case, a low level of initial knowledge in grade 9 is less consequential in grade 12 (the depreciation will be greater) if the student is retained. If investment levels were zero in every year, then retention would be harmful for end-of-schooling knowledge levels, as retention implies an extra year of knowledge depreciation. However, in the general case when knowledge depends not only on lagged knowledge but also on additional family and school inputs (investments), the effect of retention will depend on the relative benefit of an additional year of investment and on the lag coefficients determining how knowledge accumulates from year to year.

In principle, separate knowledge equations could be estimated for each grade level. However, we only observe test scores (knowledge measures) in grades 9 and 12 when the tests were administered, so we impose some restrictions on the stability of the value-added equation coefficients across grades. In particular, we assume $\gamma^g = \gamma$, $\gamma^{gR} = \gamma^R$, $\beta^g = \beta$, $\beta^{gR} = \beta^R$ for all grades g. We estimate separate parameters for math and Portuguese.

4.4 Dropout and grade retention

In grades 10 and 11, students are either retained in all subjects or in none, but in grade 12 they may be retained in one subject (e.g. math) but not the other (e.g. Portuguese). We model retention in grades 10 and 11 as functions of both math ($K_{i,t}^{g,M}$) and Portuguese ($K_{i,t}^{g,P}$) knowledge levels:

$$R_{i,t}^{10} = \mathbb{1}(\lambda_0^{10} + \lambda_1^{10} K_{i,t}^{10,M} + \lambda_2^{10} K_{i,t}^{10,P} + \lambda_3^{10} H_{i,t-3}^{R} + \lambda_4^{10} I_{i,t} + \lambda_5^{10} X_i + \sum_{m=1}^{M} \alpha_{RET}^{m,10} \mu_i^m + \nu_{i,t}^{10} > 0) \ , \tag{5}$$

$$R_{i,t}^{11} = \mathbb{1}(\lambda_0^{11} + \lambda_1^{11} K_{i,t}^{11,M} + \lambda_2^{11} K_{i,t}^{11,P} + \lambda_3^{11} H_{i,t-3}^{R} + \lambda_4^{11} I_{i,t} + \lambda_5^{11} X_i + \sum_{m=1}^{M} \alpha_{RET}^{m,11} \mu_i^m + \nu_{i,t}^{11} > 0) \ , \tag{6}$$

where $\nu_{i,t}^{g}$ is a normally-distributed error term, and $I_{i,t}$ denotes family investment, as before. We also assume that retention potentially depends on an additional variable that captures the "culture of retention" in the *concelho* (similar to a municipality) where the student attends high school. This variable, $H_{i,t-3}^{R}$, is the average retention rate in the *concelho* three years earlier. This variable provides an exclusion restriction generating variation in retention rates across students. The term $\sum_{m=1}^{M} \alpha_{RET}^{m,11} \mu_i^m$ is the effect of unobserved heterogeneity (discrete multinomial types).

As noted, retention in grade 12 is subject-specific, so we allow for two subject-specific retention equations in that grade:

$$R_{i,t}^{12,M} = \mathbb{1}(\lambda_0^{12,M} + \lambda_1^{12,M} K_{i,t}^{12,M} + \lambda_2^{12,M} H_{i,t-3}^{R} + \lambda_3^{12,M} I_{i,t} + \lambda_4^{12,M} X_i + \sum_{m=1}^{M} \alpha_{RET}^{m,12M} \mu_i^m + \nu_{i,t}^{12,M} > 0) \ , \tag{7}$$

$$R_{i,t}^{12,P} = \mathbb{1}(\lambda_0^{12,P} + \lambda_1^{12,P} K_{i,t}^{12,P} + \lambda_2^{12,P} H_{i,t-3}^{R} + \lambda_3^{12,P} I_{i,t} + \lambda_4^{12,P} X_i + \sum_{m=1}^{M} \alpha_{RET}^{m,12P} \mu_i^m + \nu_{i,t}^{12,P} > 0) \ . \tag{8}$$

Retention is observed in every year.

Similarly, we observe which students drop out in every year. We assume that the dropout decision depends on investment variables, $I_{i,t}$, knowledge in both subjects in the prior year, $K^M_{i,t-1}$ and $K^P_{i,t-1}$, an individual's unobserved type, and four variables describing labor market conditions. The first and second variables measure gender-specific employment probabilities for high school dropouts and high school graduates, respectively, in the region of the country where the student attends school. The third and fourth variables are gender-specific monthly salaries (Euros/mo) for high-school dropouts and high school graduates in the district where the student attends high school. There are a total of five regions and eighteen districts in Portugal.[18]

Dropping out in grades 10, 11 or 12 is modeled as a binary outcome. In deciding whether to drop out, individuals compare their expected future lifetime earnings stream if they stay in school versus if they drop out.[19] Let $d^g_{i,t}$ denote whether a student drops out at grade g. Let Y_{1t} and Y_{0t} denote the earnings at time t for a student who drops out versus stays in school. The dropout decision is made sequentially and is based on maximization of expected future earnings net of any psychic costs (C) of attending school:

$$d_{it} = 1 \text{ if } E[\sum_{j=1}^{T=k} \frac{Y_{1,t+j}}{(1+r)^j} - \sum_{j=0}^{T=k} \frac{Y_{0,t+j}}{(1+r)^j} - C|\Omega_t] \geq 0, \text{ else} = 0. \tag{9}$$

The first term is the earnings stream if the person drops out, the second term is the earnings stream if he/she does not drop out, which reflects the foregone earnings cost of attending an extra year of school, k is the year of retirement, and Ω_t is the information set at time t used to form expectations about future earnings.[20] We assume individuals forecast their labor market earnings prospects using their own characteristics and using information on the earnings of current labor market participants in the locality where they reside. The dropout decision will also depend on the costs of remaining in school, C, which we assume to be a function of knowledge levels, family

[18] Because we do not observe schooling information in the year the student drops out, we write the dropout model as a function of schooling variables measured in the prior academic year. We have in mind a scenario in which a student decides in August 2009 whether to enroll in school the next year, $2009/2010 = t + 1$, and bases his decision jointly on current labor market conditions and what transpired in the previous academic year – scores on exams as measured in May/June as well as family investment in the prior academic year.

[19] Our dropout model is based on a general framework described in Heckman et al. (1999).

[20] If the individual pursues additional years of schooling, then the earnings could be zero for some years, $Y_{0,t+j} = 0$ for $j = 0, 1, 2, ...$

background and the likelihood of future retentions (as reflected in prior retentions). The dropout model for grade g, which is a function of these variables, is specified as:

$$d^g_{i,t} = \mathbb{1}(\delta^g_0 + \delta^g_1 K^{g,M}_{i,t-1} + \delta^g_2 K^{g,P}_{i,t-1} + \delta^g_3 L_{i,t} + \delta^g_4 \sum_{k=1}^{t-1} R^{g(k)}_{i,k} + \delta^g_5 I_{i,t-1} + \delta^g_6 X_i + \sum_{m=1}^{M} \alpha^{g,m}_D \mu^m_i + \eta^g_{i,t} > 0) \; , \tag{10}$$

where $L_{i,t}$ is a vector containing the four labor market variables and $\sum_{k=1}^{t-1} R^{g(k)}_{i,k}$ is the total number of retentions in high school up through the prior year, $t-1$. The coefficient on this variable, δ^g_4, captures possible discouragement effects of retention in high school and is the main channel through which retention influences drop out. The other, indirect, channel through which retention influences dropout is through its effect on future test scores.

5 Estimation

We estimate the parameters of the model using simulated method of moments. We adopt an unconditional simulation approach (Gourieroux et al. (1996)), meaning that we simulate each student's path of educational outcomes throughout high school starting from ninth grade initial conditions. We chose an unconditional simulation approach, because test scores are not available in all grades. The model parameters minimize the distance between the model simulations and data moments.

Table 1 lists common high school trajectories in the data. As previously noted, students are sometimes retained in one subject but not the other. We define a variable, h_i, that represents an individual's history, which is the Cartesian product of an individual's math path and Portuguese path, $h_i \equiv PathMath_i \times PathPT_i$. We denote the set of possible histories by $\mathcal{H}$. $\mathcal{H}$ contains more histories than are shown in Table 1, which lists only the most common histories in the data.

To simplify notation, let $y_{i,t} = (K^{g,M}_{i,t}, K^{g,P}_{i,t}, h_i)$ denote the vector of endogenous variables, $z_i = \{I_{i,t}, X_i, L_{i,t}, H^R_{i,t}\}_{t=1}^T$ the vector of exogenous variables, $\epsilon_{\mathbf{i}} := \{\{\epsilon^s_{i,t}\}_{t=1}^T\}_{s=1}^S = \{\{\nu^s_{i,t}, \varepsilon^{g,s}_{i,t}, \varepsilon^{gR,s}_{i,t}, \eta^s_{i,t}\}_{t=1}^T\}_{s=1}^S$ the set of simulated errors, and θ the vector of model parameters. A history includes the dropout and retention decisions, so we do not separately include $R^g_{i,t}$ and $d^g_{i,t}$, in $y_{i,t}$.

We simulate the full path of endogenous variables as a function of exogenous variables and shocks. A given simulation, s, can be written as $y^s_{i,t} = r(y^s_{i,t-1}(\theta), z_i, \epsilon^s_i)$ where $y^s_{i,t-1}(\theta)$ is a simulated value that depends on the parameter vector θ as well as prior realizations of z_i and ϵ_i^s. A path simulation allows us to write $y_{i,t}^s$ as a function of only the exogenous variables, z_i, the shocks, $\epsilon_{\mathbf{i}}^{\mathbf{s}}$, and the initial value of the process, $y_{i,0}$: $y_{i,t}^s = r(y_{i,0}, z_{i,1}, \ldots, z_{i,t}, \epsilon_{i,1}^s, \ldots, \epsilon_{i,t}^s)$.

The Simulated Method of Moments estimator minimizes the following objective function:

$$J(\theta) = \{\frac{1}{N}\sum_{i=1}^{N} z_i[F(y_i) - f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta)]\}'\Omega^*(\theta)\{\frac{1}{N}\sum_{i=1}^{N} z_i[F(y_i) - f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta)]\}\ , \tag{11}$$

where $F(y_i)$ is a function only of the data and $f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta)$ is its corresponding simulated value. $f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta)$ is an unbiased simulator of $F(y_i)$, meaning that $\mathbb{E}[f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta) \mid z_i] = \mathbb{E}[F(y_i) \mid z_i]$.

Recall that our model includes unobserved discrete types. Our simulation-based estimator integrates over the unobserved types by drawing the types from a multinomial distribution and then simulating the rest of the outcomes conditional on type. The type probabilities are estimated along with the other parameters (i.e. included in θ). For ease of notation, we do not write $f(z_i, \epsilon_{\mathbf{i}}, y_{i,0}; \theta)$ as a function of the unobserved types, μ_i^m. Once drawn for each individual and each simulation, types are treated in the same way as observed covariates.[21]

There are three basic functions in $F(y_i)$ and $f(z_i, \epsilon_i, y_0; \theta)$: $\mathbb{1}_{(h_i=h)}$, $\mathbb{1}_{(h_i=h)}K_{i,t}^g$, and $\mathbb{1}_{(h_i=h)}{K_{i,t}^{g}}^{2}$. The moments we target, $m_i = z_i \cdot F(y_i)$, are covariances between exogenous variables and a student's endogenous history, covariances between exogenous variables and observed endogenous test scores, and the second moments of test scores. A full list of the moments is provided in Appendix A.

[21] The types appear jointly in the value-added, dropout, and retention equations. Identification of the distribution of these types follows an identification-in-the-limit argument of Heckman and Navarro (2007). Their theorem, which requires only that errors be independent across individuals and of regressors in the initial time period, includes our multinomially distributed types as a special case. This argument requires transition-specific exclusion restrictions, which in our setting are the local labor market variables and the historical retention rates in each concelho for dropout and retention, respectively. Appendix D presents plots of the SMM objective at the minimum as a function of each type-specific parameter, holding constant all other parameters. The plots reveal that the function is responsive to changes in the parameter values.

We use the optimal weight matrix, $\Omega^*(\theta)$, which can be estimated by the following formula:

$$
\begin{aligned}
\hat{\Omega}^*(\theta) = \Bigg\{ \frac{1}{N} \sum_{i=1}^{N} z_i & \left[F(y_i) - \frac{1}{S_2} \sum_{s_2=1}^{S_2} f\left(z_i, \epsilon_i^{s2}, y_{i,0}; \theta\right) \right] \\
& \times \left[F(y_i) - \frac{1}{S_2} \sum_{s_2=1}^{S_2} f\left(z_i, \epsilon_i^{s2}, y_{i,0}; \theta\right) \right]' z_i' \\
& + \frac{1}{S} \frac{1}{N} \sum_{i=1}^{N} z_i \left[f\left(z_i, \epsilon_i^{s1}, y_{i,0}; \theta\right) - \frac{1}{S_2} \sum_{s_2=1}^{S_2} f\left(z_i, \epsilon_i^{s2}, y_{i,0}; \theta\right) \right] \\
& \times \left[f\left(z_i, \epsilon_i^{s1}, y_{i,0}; \theta\right) - \frac{1}{S_2} \sum_{s_2=1}^{S_2} f\left(z_i, \epsilon_i^{s2}, y_{i,0}; \theta\right) \right]' z_i' \Bigg\}^{-1}
\end{aligned}
\tag{12}
$$

As discussed in Gourieroux et al. (1996), obtaining this optimal weighting matrix requires doing additional simulations.[22] The optimal weighting matrix depends on simulations used in the estimation, ϵ_i^{s1}, as well as a separate set of simulations used only for computing the weight matrix, ϵ_i^{s2}.[23] We first optimize $J(\theta)$ with the identity weight matrix to obtain an estimate of θ. Then, we calculate $\hat{\Omega}^*(\theta)$ and re-optimize with this new weight matrix.

Because $J(\theta)$ is discontinuous in θ, it can be time-consuming to optimize the function. We therefore pursue an strategy proposed by Ackerberg (2009), which uses importance sampling to produce a simulator that is continuous in θ and which permits the use of gradient-based optimization techniques. Appendix B provides a detailed explanation of how we implement this method.

6 Empirical results

6.1 Data description

We estimate the model using an administrative data set obtained from the Portuguese Ministry of Education (MISI) that tracks all public school students throughout their schooling trajectory during the academic years 2008/09 - 2013/14. It follows 48,697 individuals enrolled in the 9th grade in 2008/09 that choose the general track (rather than the vocational track) in the beginning of secondary school (10th grade).[24]

[22] Although the main simulation for SMM requires S simulations, the calculation of the optimal covariance matrix requires a larger number of simulations, $S_2 > S$.

[23] Gourieroux et al. (1996) suggest using one of the S values of ϵ_i^{s1}.

[24] The initial data set comprises all individuals, regardless of age, enrolled in a public school in the 9th grade in the academic year 2008/09 representing a total of 82,412 individuals. In the following year, students may be repeating

We have information on the schooling trajectory of all students, starting with 9th grade and then for up to five years after. For each student, we have information on whether they were promoted or retained at the end of each academic year and if they dropped out. Some observations lack complete records, but the quality of the panel is extremely good. Discarding observations with obvious coding errors and missing data, we end up with a sample of 46,103 individuals: 36,496 who enter high school at the modal age, 6,952 lagging behind one year for their age, and 2,655 lagging behind two or more years. During the time period of our data, all students take the standardized national exams in math and Portuguese at the end of the 9th grade. We observe 9th grade test scores and, for students who finish high school, we also have 12th grade national exam scores.[25] Figure 1 shows the distribution of grade retentions during the three years of high school by whether a student graduates from high school or drops out. 26,693 of students do not repeat any grade, which represents 58%. One-year repeaters amount to 26%, while the remaining 16% of students are retained more than one year. A total of 16,944 students drop out, of whom 7,818 (46%) were retained one year and 7,521 (44%) were retained two or more years while in high school.

Figure 1: Retention Distribution by High School Completion/Dropout Status

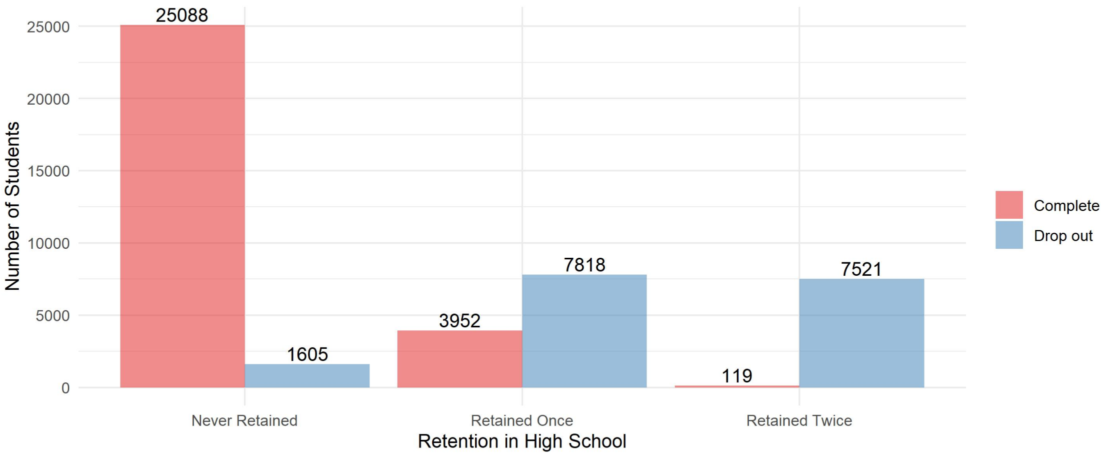


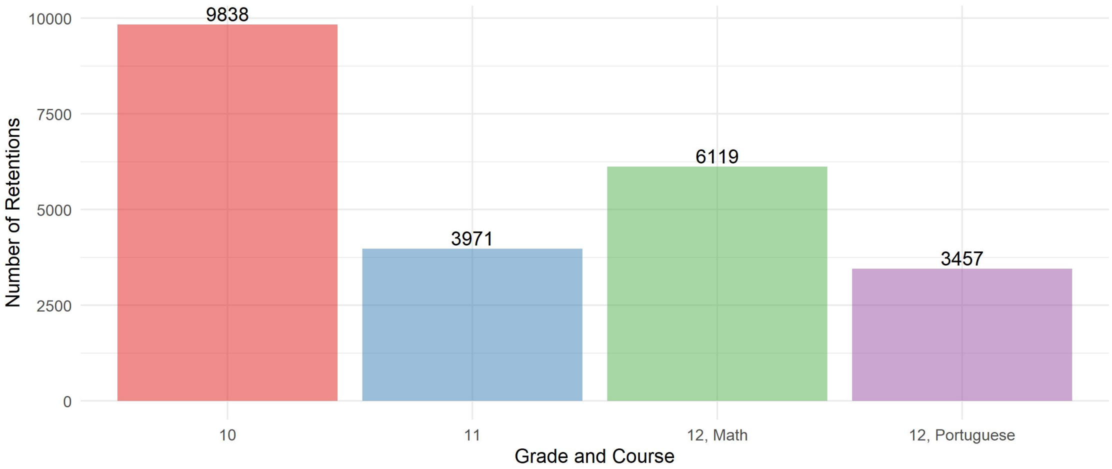
Figure 2 displays the fraction of students retained in each grade. Retention rates vary consid-

9th grade, be in the 10th grade general track or 10th grade vocational track, or have dropped out. We analyze the 48,697 students who transition to the 10th grade general track.

[25] For students who were retained two times while in high school, we see whether the student is still enrolled in the final year but we do not see the end-of-year exam scores.

erably by course subject level and grade level. 21.3% of students are retained in the 10th grade, whereas only 9.7% are retained in the 11th grade. In the 12th grade retention is subject-specific, and students are more commonly retained in mathematics than in Portuguese.

Figure 2: Fraction retained in each grade



6.2 Summary statistics

The data provide detailed information on individual and family characteristics for each student, including their grade, track of studies, gender, age, and nationality. We also have information on family background, including whether the family receives governmental income support (a proxy for lower income), parents' education. and information on the student's class and school, including class size, school size, and their teacher's age and gender.[26]

Table 2 shows summary statistics for the analysis sample. Girls are slightly overrepresented (55%) but are underrepresented among retained students. Most students (79%) are 16 or younger upon entering 10th grade. Among youth who experience retention, 33% are 17 or older upon entry. One-third of the sample is in the "low SES" category, with a higher proportion (38%) for retained students. Mothers of children in the sample tend to be better educated than fathers: 45% of mothers have either less than basic education (grade 9) or have their education listed as

[26] Students attend classes with a particular group of students, so class size is the same in both Portuguese and math.

Table 2: Descriptive statistics

Variable	Category	All	Never Retained	Retained
Gender	Female	0.55	0.60	0.47
Age at High	16 or Under	0.79	0.89	0.66
School Entry	17	0.15	0.09	0.23
	18 or Older	0.06	0.02	0.10
SES	Low	0.33	0.29	0.38
	High	0.67	0.71	0.62
Mother's Ed	Less than Basic/Unknown	0.45	0.39	0.53
	Basic	0.18	0.17	0.19
	High School	0.19	0.20	0.18
	More than High School	0.18	0.24	0.10
Father's Ed	Less than Basic/Unknown	0.51	0.45	0.59
	Basic	0.18	0.18	0.19
	High School	0.18	0.19	0.15
	More than High School	0.13	0.17	0.07

Note: Low socioeconomic status (SES) corresponds to children of parents whose income qualifies them for a public subsidy. A basic education corresponds to nine years of schooling.

"unknown" compared with 51% for fathers, and 18% of mothers have more than a high school education compared with 13% for fathers.

Table 3 shows summary statistics pertaining to test scores, school characteristics, and teacher characteristics. It also shows the average values of the local labor market conditions (wages, unemployment rates) and historical grade retention rates in each *concelho* (municipality), that affect the dropout and retention decisions. Both the 9th grade and 12th grade tests are measured on a 0 to 100 scale, but the means differ; the average test score in 9th grade is 61.6 in math and 59.5 in Portuguese in comparison to 46.6 in math and 50.6 in Portuguese in 12th grade. Students who experience retention have test scores below average and the gap between retained and nonretained students widens from 9th to 12th grade. Moreover, the gap in average test scores between retained and non-retained students is greater in math than in Portuguese in both 9th and 12th grade. The largest gap of 25 points occurs for 12th grade math.

The second panel in Table 3 shows average school and teacher characteristics. Overall, school characteristics are similar for retained and nonretained students. The average school size is around

Table 3: Test scores, school/teacher characteristics, labor market, and historical retention variables

Variable	Category	All	Never Retained	Retained
Test Scores				
Math	9th	61.6	68.8	51.7
		(20.7)	(19.6)	(17.9)
	12th	46.6	57.2	21.6
		(23.8)	(18.7)	(13.5)
Portuguese	9th	59.5	64.8	52.2
		(14.8)	(13.8)	(12.9)
	12th	50.6	55.5	40.0
		(16.7)	(15.4)	(14.6)
School/Teacher Characteristics				
School Size		517	525	507
		(272)	(272)	(271)
Class Size		24.8	24.7	24.9
		(4.26)	(4.33)	(4.17)
Female Teacher	Math	0.72	0.73	0.71
		(0.451)	(0.442)	(0.459)
Female Teacher	Portuguese	0.79	0.80	0.79
		(0.405)	(0.404)	(0.406)
Teacher Age	Math	42.6	43.7	41.3
		(13.9)	(12.8)	(15)
Teacher Age	Portuguese	44.2	44.3	44.0
		(13.9)	(13.7)	(14.1)
Local Market/Retention History Variables				
Local Wage (Euros/mo)	HS Dropout	607	596	622
		(93.8)	(90.6)	(95.9)
	HS Graduate	750	733	775
		(147)	(140)	(151)
Local Emp. Rate	HS Dropout	0.88	0.88	0.88
		(0.023)	(0.023)	(0.024)
	HS Graduate	0.90	0.90	0.90
		(0.022)	(0.021)	(0.022)
Hist. Retention Rate		24.1	23.6	24.7
		(6.11)	(6.06)	(6.12)

Note: Standard deviations are shown in parentheses. The first observed 12th grade test score is used to compute the mean for retained students.

517 and the average class size is 25. The average age of teachers is between 42 and 44, and over 70% of teachers are female. A lower proportion of math than Portuguese teachers is female.

The third panel of Table 3 shows the means of the local labor market variables. High school dropouts make on average 607 euros per month and high school graduates 750 euros per month, a 24% premium. The employment rate is 90% for high school graduates and 88% for high school dropouts, which does not differ for the subset of students who were retained or not. Lastly, we include the historical retention rate in the *concelho* as a potential determinant of the probability of retention. This variable is measured three years prior, at time $t-3$, and captures the extent to which the municipality has a "culture" of grade retention. The average value of this variable across all regions and all grades is 24%.

Table 4 shows grade progression patterns in math and Portuguese by grade and gender. All students are required to take the Portuguese exam, but only students pursuing the STEM subtrack take the mathematics exam. Therefore, the sample sizes differ by track, which explains why the dropout rates can also differ. In grade 10, in the STEM track, 88% of students complete the grade and 12% drop out of school. Girls generally exhibit better performance than boys in terms of grade progression. Girls have lower dropout rates and higher grade completion in every grade, and, as shown in Table 2, lower retention rates. In grade 12, the completion rate in mathematics drops substantially, to 59% for girls and 50% for boys. This is likely due to the high rates of retention in mathematics in the 12th grade, as shown in Figure 2, and the strong connection between retention and dropout (recall Figure 1), which we will investigate further in the next section.

6.3 Estimated model parameters

Table 5 shows the parameter estimates for the value-added knowledge accumulation models. The first and third columns show the parameters for the first-time grade enrollees for math and Portuguese. The second and fourth columns show the corresponding estimates for grade repeaters. For the math test score, the lagged math score in the previous grade is highly statistically significant, with a coefficient 0.94 for first-time enrollees and 0.80 for repeaters. Students who are older than their peers at the time of entering high school (lagging behind, likely due to retentions in earlier

Table 4: High school completion and dropout rates by gender, track, and grade

STEM						
	All		Girls		Boys	
Grade	Complete	Dropout	Complete	Dropout	Complete	Dropout
10	0.88	0.12	0.90	0.10	0.85	0.15
11	0.78	0.09	0.82	0.08	0.75	0.10
12	0.55	0.24	0.59	0.22	0.50	0.25
All Tracks						
	All		Girls		Boys	
Grade	Complete	Dropout	Complete	Dropout	Complete	Dropout
10	0.90	0.10	0.93	0.07	0.88	0.13
11	0.84	0.07	0.87	0.06	0.79	0.08
12	0.73	0.10	0.79	0.08	0.67	0.13

grades) have significantly lower test score gains. With regard to teacher characteristics, having a female teacher is associated with a lower test score gain and having an older teacher a higher test score gain among first-time enrollees. The data do not contain information on teachers' exact experience levels, but age likely indicates greater experience. Larger class size lowers test scores for first time enrollees, but the effect size is small (0.36 points per increase of 10 students). Youth whose mothers have higher education levels show larger gains and having a father with the highest education level also significantly increases test scores. Females show substantially larger test score gains than males in both math and Portuguese.

For the Portuguese knowledge accumulation equation, the estimated coefficient on the lagged test score is also highly statistically significant, although the magnitude is lower than for math - 0.85 for first grade enrollees and 0.70 for repeaters. Individuals who are older relative to peers again have significantly lower test score gains. Teacher age and gender are not statistically significant. Children who come from low SES families exhibit lower test score gains. As with math, test scores gains increase with the mother's education level, and females have greater gains. However, father's education is not significant for Portuguese.

The last five rows of the table provide information on the estimated distribution of the unobserved types and the estimated type-specific intercepts. The model included three types and the estimated type proportions are roughly equal (33%, 35%, and 32%). The type-specific intercepts are not statistically different from zero, suggesting that the included observables capture much of the heterogeneity across students in test score gains.

Table 5: Estimated value-added equation parameters

	Math				*Portuguese*			
	First time		Repeating		First time		Repeating	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Lagged Score, $K_{i,t-1}$	0.94	0.00	0.80	0.02	0.85	0.00	0.76	0.03
Constant	-5.93	0.29	4.93	2.85	4.17	0.37	19.88	2.84
Relative Age at Gr 10	-4.35	0.14	-0.74	0.72	-1.99	0.08	-1.40	0.58
Missing School Info	-1.18	0.34	5.96	3.23	0.73	0.42	-3.30	3.58
Female Teacher	-1.62	0.24	2.00	2.40	0.26	0.28	0.51	2.55
Teacher Age	0.46	0.06	0.92	0.53	-0.03	0.06	-0.40	0.41
School size	0.33	0.23	-0.96	1.60	-0.05	0.16	-2.05	1.29
Class size	-0.36	0.10	0.59	1.01	0.00	0.11	2.29	0.69
Low SES	-0.88	0.18	0.07	1.47	-0.46	0.15	-1.69	1.31
Mother Basic Educ	1.30	0.20	4.21	1.33	0.35	0.13	-5.79	1.08
Mother HS Educ	1.90	0.18	1.41	1.48	0.44	0.12	0.79	1.26
Mother $>$ HS Educ	2.01	0.21	3.82	1.89	1.48	0.15	1.14	1.21
Father Basic Educ 1	-0.45	0.18	-5.12	1.21	0.38	0.12	-1.52	1.05
Father HS Educ 2	-1.36	0.18	0.44	1.59	-0.07	0.13	-0.75	1.26
Father $>$ HS Educ 3	1.51	0.22	2.35	2.09	0.26	0.16	-1.63	1.25
Female	3.00	0.12	1.60	0.91	0.67	0.08	2.27	0.73
Intercept type 1 (α_1)	0.21	1.72	-0.33	2.69	0.16	1.20	-1.50	3.63
Intercept type 2 (α_2)	0.00	1.74	0.50	3.51	1.42	1.25	1.06	3.98
σ^2	85.09	0.98	138.89	9.33	85.09	0.98	138.89	9.33
$P(Type1)$	0.33	0.02	0.33	0.02	0.33	0.02	0.33	0.02
$P(Type2)$	0.35	0.01	0.35	0.01	0.35	0.01	0.35	0.01

Note: The table presents parameter estimates for the four value-added equations. The omitted education category is less than basic. Missing School Info is an indicator for whether information was missing on the student's teacher variables or class size. School size is measured in thousands of students, class size is measured in tens of students, and teacher age is measured in tens of years.

Table 6 shows the dropout parameter estimates for different grades. Students with higher test scores have significantly lower dropout propensities. Local labor market conditions are important determinants of dropout decisions. Specifically, a higher average income for dropouts increases the dropout probability whereas a higher average income for high school graduates lowers it. Being retained in a prior year of high school substantially increases the dropout probability, as indicated by the variable *Yrs Retained in HS*. Dropout is also higher for youth attending schools with larger classes. With regard to parental background, higher mother's education levels are protective against dropout. Father's education is often statistically significant as well, although the pattern of estimated coefficients is not monotonic in education levels. Lastly, females have a lower probability of dropping out in 10th grade but a higher probability in the 11th or 12th grade. The type intercepts are not statistically significantly different from zero. Table 7 shows the estimated model coefficients for the grade-specific retention equations. We estimate two models for grade 12, because students can fail in math or Portuguese (or both) in this grade. We know that test scores are one of the factors that schools use along with grades in making retention decisions. As expected, the retention

Table 6: Dropout equation parameters

	Grade 10		*Grade 11*		*Grade 12*	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	3.54	0.61	-1.23	0.21	-1.21	0.14
Math	-0.57	0.21	-0.79	0.14	-1.38	0.12
Portuguese	-1.04	0.25	-0.79	0.16	-1.59	0.16
Local Employment: Dropout	-1.10	0.38	-0.1	0.13	-0.37	0.09
Local Employment: Graduate	-0.12	0.40	-0.64	0.13	-0.09	0.09
Local Income: Dropout	0.52	0.14	0.82	0.06	0.41	0.06
Local Income: Graduate	-0.32	0.06	-0.39	0.04	-0.15	0.04
Relative Age at Gr 10	0.65	0.08	0.42	0.07	0.36	0.05
Yrs Retained in HS	2.58	0.29	2.47	0.12	1.38	0.09
School size	-0.44	0.15	-0.05	0.12	0.49	0.12
Class size	1.41	0.15	1.33	0.08	0.94	0.06
Low SES	0.12	0.08	0.21	0.07	0.06	0.09
Mother Basic Educ 1	1.70	0.13	1.71	0.10	-1.47	0.21
Mother HS Educ 2	-0.38	0.15	0.15	0.13	-0.14	0.10
Mother > HS Educ 3	-0.86	0.22	-0.20	0.18	-0.26	0.19
Father Basic Educ 1	-0.40	0.12	-0.63	0.09	-0.50	0.12
Father HS Educ 2	-0.18	0.12	-0.50	0.10	-0.17	0.12
Father > HS Educ 3	-0.53	0.21	-0.91	0.19	-0.82	0.39
Female	-0.38	0.24	0.23	0.11	0.19	0.10
Intercept type 1 (α_1)	0.09	2.38	0.33	0.51	0.42	0.29
Intercept type 2 (α_2)	0.30	2.50	0.35	0.45	0.39	0.31
Non-STEM	...	...	...	...	0.00	0.11

Note: The omitted education category is less than basic. The coefficients and standard errors on math and Portuguese knowledge have been scaled so that they represent the effects of a 100-point increase in these scores. School size is measured in thousands of students and class size is measured in tens of students. Local income is measured in 100s of Euros and the employment rate has been multiplied by 10.

probability depends significantly on knowledge (test scores). Students are also more likely to be retained if a *concelho* has a culture of retention (as indicated by the retention rate in the three years prior, which serves as an exclusion restriction in the model). Students who are older relative to peers are much more likely to be retained. The retention probability depends positively on school size and class size in 10th grade but negatively on class size in 11th and 12th grade. The family's SES status is not a significant determinant of retention in most equations. Youth whose mothers have only a basic education are more likely to be retained in the 10th grade. Lastly, females have a lower retention probability in every grade and subject. Finally, we allow for non-STEM students to have different baseline probability of dropout in grade 12, although this turns out to not be significantly different from zero.

Table 7: Retention equation parameters

	Grade 10		*Grade 11*		*Grade 12*			
					Math		*Portuguese*	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.44	0.06	-1.12	0.08	-0.33	0.07	-1.02	0.06
Math	-2.26	0.06	-0.57	0.09	-1.41	0.05	...	...
Portuguese	-2.43	0.09	-0.95	0.10	...	...	-1.41	0.07
Historical Retention Rate	0.42	0.13	2.28	0.20	-0.17	0.16	-0.19	0.11
Relative Age at Gr 10	0.06	0.02	0.02	0.03	0.82	0.02	0.34	0.03
School size	0.25	0.04	0.25	0.05	0.07	0.06	0.05	0.05
Class size	0.22	0.02	-0.08	0.03	-0.17	0.02	-0.13	0.02
Low SES	-0.01	0.02	0.06	0.04	-0.40	0.05	-0.01	0.04
Mother Basic Educ	0.36	0.03	-0.10	0.05	0.13	0.05	-0.03	0.05
Mother HS Educ	-0.24	0.03	-0.28	0.05	-0.14	0.05	-0.12	0.05
Mother > HS Educ	0.13	0.04	-0.17	0.05	-0.59	0.09	-0.13	0.07
Father Basic Educ	-0.21	0.03	0.11	0.05	-0.12	0.05	0.10	0.04
Father HS Educ	0.00	0.03	0.19	0.05	-0.21	0.06	0.05	0.05
Father > HS Educ	-0.02	0.04	0.17	0.05	-0.56	0.08	-0.08	0.07
Female	-0.31	0.02	-0.24	0.03	-0.41	0.04	-0.12	0.03
Intercept type 1 (α_1)	-0.40	0.27	0.13	0.34	0.75	0.56	0.46	0.15
Intercept type 2 (α_2)	-0.03	0.29	0.10	0.37	-0.09	0.57	0.15	0.16
Non-STEM	-1.98	0.04	-0.65	0.06	...	....	...	...

Note: Retention is grade-specific in grades 10 and 11 but grade-subject-specific in grade 12. Retention in grade 12 in Math depends on math scores but not Portuguese scores and vice versa for retention in grade 12 in Portuguese. The omitted education categories are less than basic. The coefficients and standard errors on math and Portuguese knowledge have been scaled so that they represent the effects of a 100-point increase in these scores. The historical retention rate is measured on a 0-1 scale

6.4 Goodness-of-fit

Our estimation targets over nine hundred moments. Tables C-1 and C-2 in Appendix C show the model fit for the proportion of students with each history. The tables show that the model does a good job of matching the most common histories for both the STEM and non-STEM students. For example, the STEM history 10-11-12X10-11-12 is followed by 50.3% of individuals in the data and 54.3% of individuals in model simulations, while the path 10-10dX10-10d constitutes 8.77% of individuals in the data and 11.2% in the model. The model fit is not quite as good for paths that feature multiple retentions in which students stay in school for some time after the retention, such as 10-11-12-12-12X10-11-12, 10-10-11-11dX10-10-11-11d, and 10-11-11-11dX10-11-11-11d. The next section analyzes treatment effects for several different subgroups, including those students who graduate high school within four years and for whom the model fit is particularly good.

7 Grade retention impacts

7.1 Average impacts

Our main goal in estimating the above dynamic educational production function model is to evaluate grade retention effects, on both test scores and educational attainment, by simulating these outcomes in a counterfactual scenario in which retention is eliminated and comparing it with the status quo. Table 8 shows the effects of grade retention on 12th grade average test scores for three different subgroups: those who were retained in 10th or 11th grades, those who were retained in 12th grade, and those who were retained in any year and graduated high school within four years (rather than 3). The fourth column shows the proportion of students with a 12th grade test score in each of these categories. The table shows both the impact on the raw test score and in terms of standard deviations (S.D.). Retention has positive impacts in both math and Portuguese on average, ranging from 0.17-0.5 s.d. The largest effects for math are observed for 12th grade retention. The retention effects on Portuguese test scores are approximately equal regardless of the grade in which the student is retained.

We also consider the effect of eliminating grade retention on the probability of dropping out. Table 9 shows that retention's effect on dropout is large. Among students who are retained (the

Table 8: Retention impacts on 12th grade test scores

	Math		*Portuguese*		
Subsample of students	Raw	S.D.	Raw	S.D.	Proportion
Retained in 10th/11th Grades	3.93	0.17	8.20	0.48	0.11
	(1.27)	(0.05)	(0.87)	(0.05)	(0.01)
Retained in 12th Grade	5.11	0.22	8.42	0.50	0.08
	(2.79)	(0.12)	(1.44)	(0.08)	(0.03)
Graduate with one extra year	4.68	0.20	8.53	0.50	0.13
	(1.05)	(0.04)	(0.71)	(0.04)	(0.02)

Note: Retention impacts are estimated for the subset of students who enroll in 12th grade under both the status quo and counterfactual simulations. The standard deviations on the 12th grade math and Portuguese exams are 23.6 and 17.0 points. The column labeled proportion indicates the proportion in the simulation who take an exam in the 12th grade. Standard errors, obtained from 100 parametric bootstrap replications, are shown in parentheses.

treated), retention increases the probability of dropping out by 56 percentage points for boys and 49 percentage points for girls. The high dropout rates observed in the data among retained students, 80% for all students, can be attributed both to causal effects and selection, but the causal effects are most salient ($0.53/0.80 > 0.5$).

Table 9: Retention impacts on dropout proportions

	Dropout proportion			*Impact of retention*		
	All	Boys	Girls	All	Boys	Girls
All students	0.37	0.45	0.30	0.18	0.24	0.12
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Retained students (TT)	0.80	0.83	0.76	0.53	0.56	0.49
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)

Note: Results are shown for all students and for retained students (treatment on the treated). Dropout proportions are directly from the data, while retention impacts are calculated on the basis of model simulations.

In Table 10 we examine how 12th grade average retention effects vary with students' initial ranks in the 9th grade test score distribution (prior to entering high school). In both math and Portuguese, students who begin the ninth grade with low test scores benefit more from retention. For math, students in the bottom tercile in the ninth grade experience about a 0.3 s.d. increase in test scores and students in the middle tercile a 0.2 s.d. increase, while students in the top tercile benefit little from retention. In Portuguese, the gains are more pronounced. Students scoring in the bottom tercile in Portuguese in the ninth grade experience about a 0.6 s.d. increase in test scores, while students in the middle and top terciles experience about a 0.5 and 0.4 s.d. increase.

Table 10: Retention impacts on 12th grade scores by 9th grade test score tercile

	Bottom Tercile		*Middle Tercile*		*Top Tercile*	
Subsample	Raw	S.D.	Raw	S.D.	Raw	S.D.
			Math			
Retained in 10th/11th Grades	7.50	0.32	3.75	0.16	1.42	0.06
	(1.46)	(0.06)	(1.25)	(0.05)	(1.42)	(0.06)
Retained in 12th Grade	8.45	0.36	4.68	0.20	1.89	0.08
	(2.41)	(0.10)	(2.73)	(0.12)	(3.04)	(0.13)
Graduate with one extra year	8.55	0.36	4.73	0.20	1.99	0.08
	(1.15)	(0.05)	(1.05)	(0.04)	(1.20)	(0.05)
			Portuguese			
Retained in 10th/11th Grades	10.02	0.59	7.87	0.46	5.71	0.34
	(0.80)	(0.05)	(0.88)	(0.05)	(1.12)	(0.07)
Retained in 12th Grade	9.72	0.57	8.19	0.48	6.91	0.41
	(1.35)	(0.08)	(1.46)	(0.09)	(1.60)	(0.09)
Graduate with one extra year	10.38	0.61	8.24	0.49	6.17	0.36
	(0.69)	(0.04)	(0.71)	(0.04)	(0.92)	(0.05)

Note: Retention impacts are estimated for the subset of students who enroll in 12th grade under both the status quo and counterfactual simulations. The standard deviations on the 12th grade math and Portuguese exams are 23.6 and 17.0 points.

We also examine the retention impacts on dropping out by initial 9th grade math exam in Table 11. The table shows that students in the bottom tercile are much more likely to drop out of high school. This is especially true if the students are retained: 87% of retained students who score in the bottom third in 9th grade drop out of high school. Retention causes 55% of these retained students to drop out, meaning that 87% - 55% = 32% of of them drop out in both the retention and no-retention simulations. Retention is therefore a significant cause of dropout for these students.

In Table 12, we disaggregate the retention impacts by gender. The last column shows the proportion of the sample within each category. Impacts are larger in Portuguese than math for both girls and boys. The impacts for girls are also slightly larger than for boys.

Appendix C presents additional evidence of impact heterogeneity. Table C-3 shows that the effects of retention on test scores are similar by SES status. However, Table C-4 shows that impacts vary by the student's age relative to the majority age at his/her grade level. Students who begin

Table 11: Retention impacts on dropout by 9th grade math test score tercile

	Proportion			*Impact of retention*		
	Bottom	Middle	Top	Bottom	Middle	Top
All Students	0.60	0.38	0.10	0.27	0.18	0.08
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Retained Students	0.87	0.75	0.61	0.55	0.54	0.45
	(0.01)	(0.02)	(0.03)	(0.01)	(0.02)	(0.03)

Note: The table presents estimates of the effect of grade retention on test scores for three treated subgroups disaggregated by initial (grade 9) test score in math. Dropout proportions come directly from the data, while retention impacts are calculated on the basis of model simulations.

Table 12: Retention impacts on 12th grade test scores by gender

	Math		*Portuguese*		
Boys	Raw	S.D.	Raw	S.D.	Proportion
Retained in 10th/11th Grades	3.66	0.15	7.89	0.46	0.13
	(1.25)	(0.05)	(0.87)	(0.05)	(0.01)
Retained in 12th Grade	4.58	0.19	7.75	0.46	0.10
	(2.72)	(0.12)	(1.42)	(0.08)	(0.02)
Graduate in 4 Years	4.38	0.19	8.17	0.48	0.15
	(2.72)	(0.12)	(1.42)	(0.08)	(0.02)
Girls					
Retained in 10th/11th Grades	4.38	0.19	8.62	0.51	0.08
	(1.47)	(0.06)	(0.96)	(0.06)	(0.01)
Retained in 12th Grade	6.00	0.25	9.06	0.53	0.06
	(3.03)	(0.13)	(1.56)	(0.09)	(0.01)
Graduate in 4 Years	5.13	0.22	8.98	0.53	0.10
	(1.27)	(0.05)	(0.8)	(0.05)	(0.01)

Note: Retention impacts are estimated for the subset of students who enroll in 12th grade under both the status quo and counterfactual simulations. The standard deviations on the 12th grade math and Portuguese exams are 23.6 and 17.0 points. The column labeled proportion indicates the proportion in the simulation who take an exam in the 12th grade.

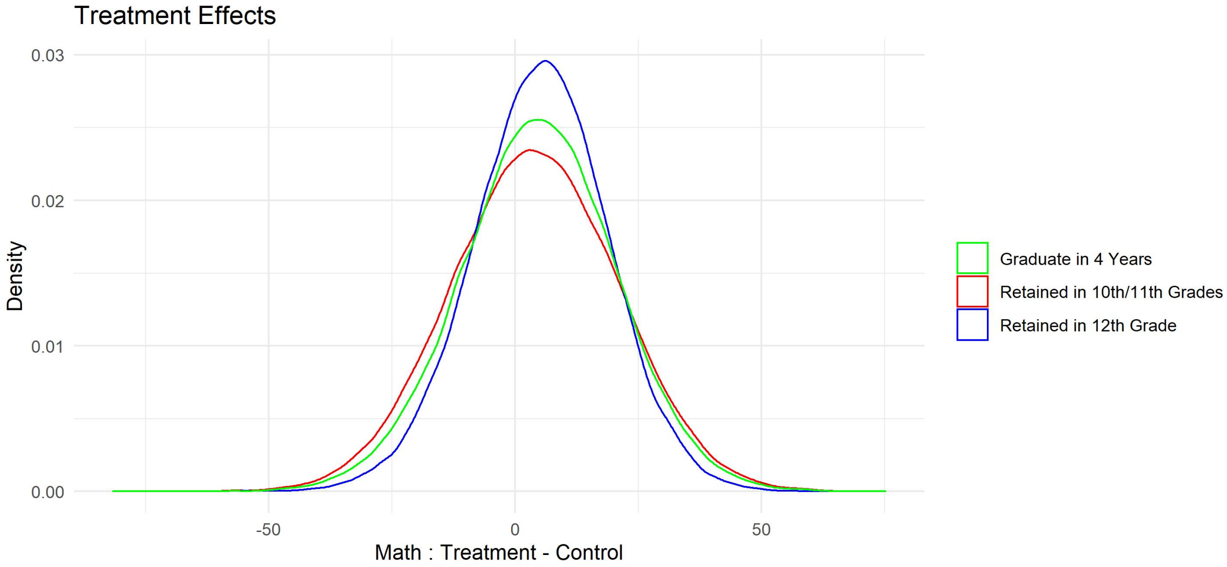
high school at ages older than the modal age for their grade experience greater test score gains in math from retention, but the effect on Portuguese scores varies little with age.

Given our finding in Table 9 that retention substantially increases dropping out, we would expect to see impacts on high school completion rates. According to the estimates reported in Table 13, retention reduces high school completion rates by 21 percentage points, with larger effects for boys (29 pp) than for girls (14 pp).

Table 13: Effect of Retention on High School Completion

	All	*Boys*	*Girls*
Status Quo	0.74	0.65	0.81
	(0.01)	(0.02)	(0.01)
No Retention	0.94	0.94	0.95
	(0.00)	(0.01)	(0.00)
Effect of Retention	-0.21	-0.29	-0.14
	(0.01)	(0.02)	(0.01)

Figure 3: Distributional impacts on math scores, by subgroup

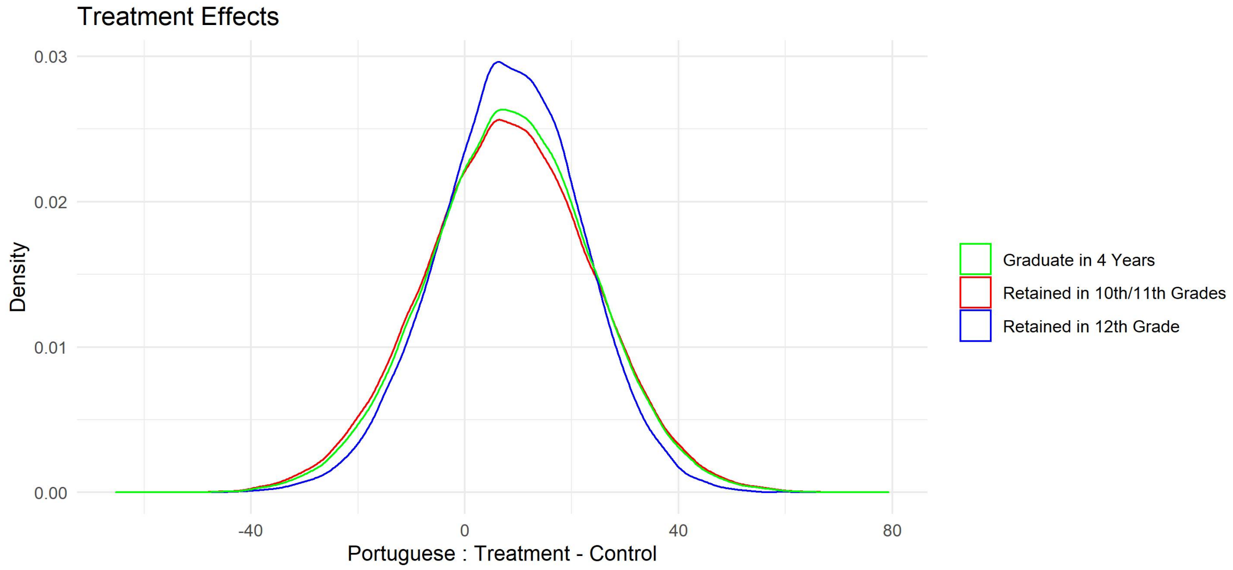


7.2 Retention Impact Heterogeneity

We next further examine impact heterogeneity across students. Figures 3 and 4 graph the distribution of 12th grade test score impacts for the same three subsamples of retained students that were considered in the last section and for math and Portuguese. Although the overall average impacts are positive, the figures reveal a wide variance with a substantial fraction of students experiencing negative impacts. The variance is higher and the incidence of negative impacts greater for 10th\11th grade retention than for 12th grade.

Table 14 shows the proportion of students who experience a negative 12th grade test score impact as a result of being retained, which ranges from 26% to 41% across the groups. Despite the

Figure 4: Distributional impacts on Portuguese scores, by subgroup



overall average test score impacts being positive, substantial numbers of students do not experience gains, even among the students who do not drop out prior to taking the 12th grade test.

We next investigate whether schools are retaining the "right" students, that is, those students who are most likely to benefit. To examine this question, we simulate outcomes for the case where everyone is retained in the 10th grade versus a scenario in which no one is retained in any grade, and compare the resulting 12th grade test scores across the two scenarios.[27] We repeat this exercise for 11th grade and 12th grade (in both math and Portuguese). In each case, we derive individual treatment impacts and nonparametrically regress these treatment impacts on the predicted probability of retention for that student as derived from the retention model. Those with the highest propensity scores are the ones at greatest risk of grade retention. Figure 5 displays the results of these nonparametric regressions. In panels (a), (b) and (c), the change in scores in both math and Portuguese attributable to retention are plotted against the probability of retention in grades 10, 11, and 12 respectively.[28]

[27] Specifically, we compare test scores after 4 years in high school in the first setting with the scores obtained after three years in the second setting.

[28] We use local linear regression with an Epanechnikov kernel and a bandwidth of 0.10 for all plots, and the 95% confidence bands depicted in the picture are based on asymptotic standard errors. We plot the regressions only where there is positive density for the retention probabilities, as indicated in Figure 5-(d).

Table 14: Proportion of students with a negative impact by subject

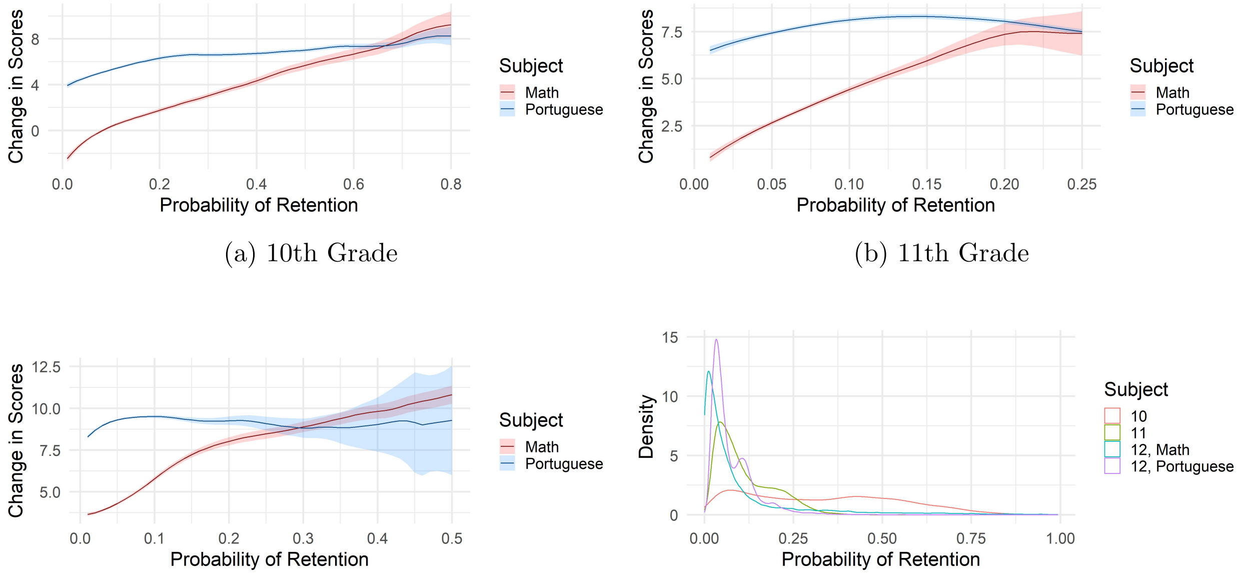
	Math	*Portuguese*
Retained in 10th/11th Grades	0.41	0.30
	(0.03)	(0.02)
Retained in 12th Grade	0.35	0.26
	(0.07)	(0.04)
Graduate in 4 Years	0.38	0.29
	(0.03)	(0.02)

The relationship between retention probability and the treatment impact on math exam scores is upward-sloping in all grades, indicating that students who are most likely to be retained benefit the most. For Portuguese test scores, the same pattern of greater gains at higher probabilities of retention holds in 10th grade, but there is little evidence of sorting, either positive or negative, in later grades.

Figure 6 displays the estimated nonparametric regressions of the dropout treatment effect on the probability of being retained in each grade and subject. As before, for each grade, the model compares a world without retention to a world in which every student is retained in that grade. The treatment effect is computed as the difference in indicator variables for whether the student dropped out at any point in high school across the two settings. The figures show that, in every grade and subject, students who are more likely to be retained are more likely to drop out as a result of retention. The relationship is also concave, with marginal increases in retention probability from initially low levels affecting dropout more than increases from high levels.

In addition to examining how test scores and dropout vary with the probability of retention, we also computed marginal treatment effects (MTE). The MTE (Heckman and Vytlacil 1999; 2005; 2007) expresses the expected treatment effect as a function of the quantiles of an unobserved variable determining selection into treatment. Given the dynamic model used in this paper, obtaining treatment effects beyond grade 10 requires simulating test score outcomes and dropout outcomes with and without retention over multiple time periods. In Appendix E, we plot the MTE curves as functions of the retention equation unobservables. The MTE curves tend to be fairly flat, suggesting that selection into retention is primarily on the basis of observed characteristics rather

Figure 5: How 12th grade test score impacts vary by the retention year and the probability of retention



(a) 10th Grade

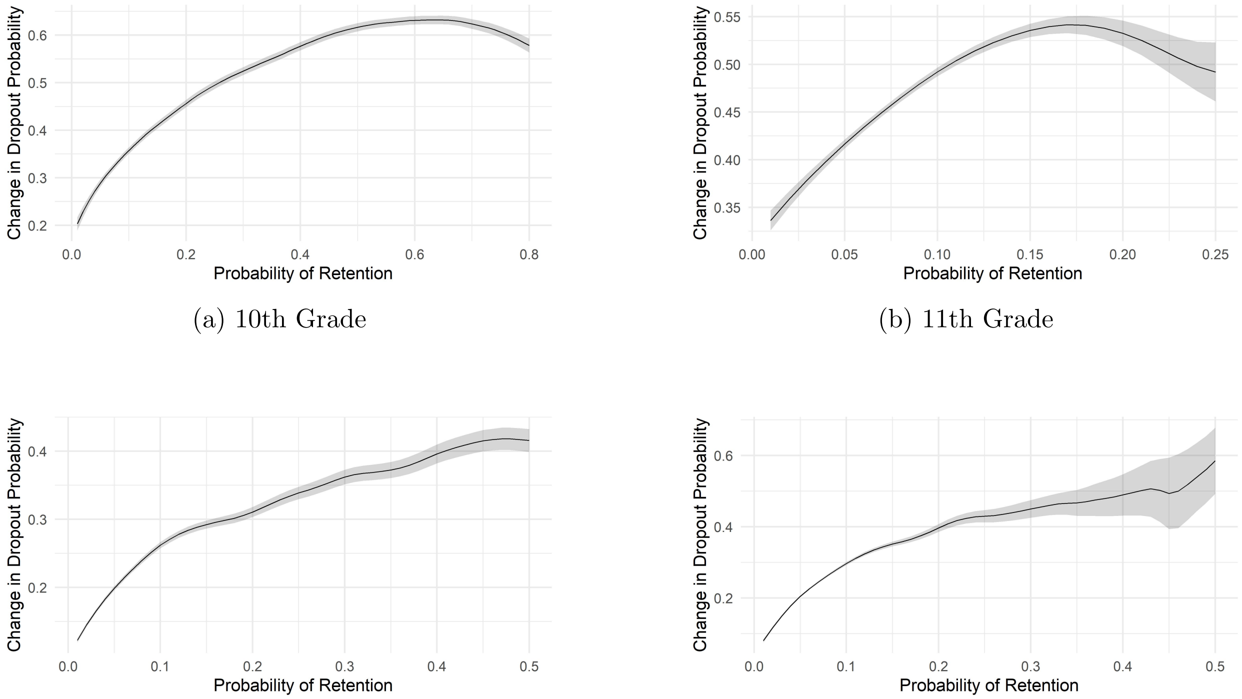
(b) 11th Grade

(c) 12th Grade

(d) Density

Note: Panels (a), (b), and (c) depict nonparametric regressions of the retention test score effect on the probability of retention in each grade for the sample of students who take the 12th grade test. The regressions are estimated using local linear regression with an Epanechnikov kernel and a bandwidth of 0.10. 95% confidence intervals are indicated by shaded regions. Panel (d) depicts the density of retention probabilities by grade and subject.

Figure 6: How dropout impacts vary by the retention year and the probability of retention



(a) 10th Grade

(b) 11th Grade

(c) 12th Grade, Math

(d) 12th Grade, Portuguese

Note: The figures depict nonparametric regressions of the dropout effect of retention on the model-derived probability of retention by grade and subject. The regressions are estimated using local linear regression with an Epanechnikov kernel and a bandwidth of 0.10. 95% confidence intervals are shaded in grey.

than unobservables.[29] Two possible explanations for why unobservables play a less important role in this context are that there is virtually no self-selection into treatment, because students do not usually want to be retained, and the initial conditions include ninth grade test scores, which are powerful predictors of later high school performance.

To summarize, the students being retained are the ones who experience the greatest potential test score benefits from retention, on average. However, grade retention also substantially increases the dropout risk. Our analysis corroborates findings from RDD studies, which typically conclude that retention improves academic performance but also increases dropout, particularly in high

[29] Recall that the coefficients on the unobserved types in the retention equations (Table 7) are mostly statistically insignificant.

school. We have also shown, though, that an RD estimator applied to Portugal would understate the test score benefits from grade retention for retained students, because the impact estimates are greatest for those with the highest predicted retention probabilities, e.g. those who would not be near the threshold for retention. A complete analysis of the costs and benefits of retention policies must simultaneously account for the benefits accruing from increased cognitive skills and the costs due to increased dropout and reduced educational attainment, a task to which we now turn.

8 Towards an Optimal Retention Policy

We have seen that retention raises test scores conditional on staying in school but also increases the probability of dropping out. An additional factor to consider is that retained students who stay in school will typically enter the labor market later and forego a year's salary (or more). Any notion of an optimal retention policy must trade off the possible earnings benefits accruing from increased cognitive ability with the costs incurred because of delayed labor market entry and potentially reduced educational attainment. Compulsory schooling laws can mitigate the costs to some extent if they are effective in preventing dropout. In this section, we consider grade retention's multiple effects on skill accumulation and educational attainment within a single framework and show how the optimal retention policy varies depending on the labor market returns to cognitive skills. Our analysis assumes that Portugal's current compulsory schooling law, which requires that individuals stay in school until age 18, is enforced.

Performing the cost benefit calculations requires predicting lifetime earnings under different grade retention policies, where retention potentially affects individuals' educational attainment, age of labor market entry, and cognitive skill levels. As shown in section 7, there is substantial individual heterogeneity in retention impacts. Below, we use a standard Mincer log wage model to predict wages, which we then convert to present discounted lifetime earnings streams, taking into account retention policy effects, which can vary at the individual level, on educational attainment, cognitive skills, and the age of labor market entry. Unfortunately, there are no data sets available for Portugal that include measures of wages and of math and Portuguese skills. However, we consider a range of plausible estimates for returns to these skills, drawing on evidence reported in the literature using data sets for other countries, and we compute the optimal retention rate as the one that maximizes average expected lifetime earnings.[30] Our findings enable us to answer the question of how high the return to cognitive skills would need to be to justify the policy that was in place at the time of our data collection of retaining 44% of students.

To estimate the returns to education in Portugal, we use employee-employer matched data derived from the *Quadros de Pessoal* data set. We estimate a standard Mincer-type log wage regression:

$$lnw_{i,t} = \alpha + \beta * S_i + \delta_0 exp_{i,t} + \delta_1 exp_{i,t}^2 + \varepsilon_{i,t} \ , \tag{13}$$

where $lnw_{i,t}$ is the log hourly wage measured in Euros for individual i in year t, S_i is educational attainment, and $exp_{i,t}$ is Mincer experience (age − years of schooling − 6).[31]

We simulate the wage for individual i at age a under retention policy P as follows:

$$w_{i,t}(a, S_i, K_i^M, K_i^P, \varepsilon_{i,t}, P) = \exp(ln\hat{w}_{i,t}(a, S_i(P)) + \phi_1 K_i^M(P) + \phi_2 K_i^P(P) + \varepsilon_{i,t}) \ , \tag{14}$$

where we vary ϕ_1 and ϕ_2 according to a range of values found in the literature. The retention policy simultaneously affects educational attainment, $S_i(P)$, and knowledge in math and Portuguese, $K_i^M(P)$ and $K_i^P(P)$, which we standardize by their subject-specific means and standard deviations in the data. After simulating earnings for each individual in each year, we then compute annual earnings by multiplying the wage by 170 hours per month and 14 months of pay per year (workers in Portugal receive both holiday pay and a Christmas bonus each worth one month of pay), and discount future years using an interest rate of 2%.[32] We calculate lifetime discounted earnings using the following expression:

$$Y_i(P) = \sum_{a=a_P}^{a=65} (\frac{1}{1+r})^{a-19} w_{i,t(a)}(a, S_i, K_i^M, K_i^P, \varepsilon_{i,t}, P) \times 170 \times 14 \tag{15}$$

[30] See, for example, Chetty et al. (2011), Dougherty (2003), Heckman et al. (2006a), Cawley et al. (2001), Murnane et al. (2000). A recent study by Watts (2020) uses a longitudinal UK data set, the National Child Development Study, to estimate associations between math and reading skills measured at age 16 and subsequent earnings at four ages between 33-50. The effects reported in the paper of a standard deviation increase in test scores on yearly earnings range over the lifecycle from 3% to 11% in math and from 4% to 12% in English.

[31] Years of schooling includes any retention years. Parameter estimates for equation (13) are given in Appendix table C-5.

[32] We assume in our cost-benefit calculation that individuals work at every age after high school graduation.

where a_P is the age of labor market entry, which may vary with the policy, thereby accounting for the foregone earnings due to delayed labor market entry. The net benefit in terms of lifetime earnings of the status quo policy, $\tilde{P}$, relative to a world with no retention is

$$\Delta = \frac{1}{N} \sum_{i=1}^{N} Y_i(\tilde{P}) - Y_i(0).$$

Δ corresponds to a policy-relevant treatment effect (PRTE), as defined in Heckman and Vytlacil (2005).[33] In what follows, we choose the policy $\tilde{P}$ to maximize the PRTE.

Our calculations assume that the returns to knowledge in math and Portuguese are equivalent: $\phi_1 = \phi_2 = \phi$. If, say, $\phi = 0.05$ and we set $\tilde{P}$ equal to the status quo retention policy, then $\Delta = -4,781$, indicating that the costs of policy $\tilde{P}$ exceed the benefits by 4,781 Euros over the course of the average student's lifetime.[34] In the ensuing analysis, we vary ϕ between 0.05 and 0.15 and solve for the retention policy that maximizes Δ, assuming that compulsory schooling laws are enforced. Specifically, for a given value of ϕ, we add a constant term to each retention equation (equations 5 through 8) and search over the value of this constant that maximizes Δ. We then simulate the model under this optimal policy and report the fraction of students retained in any year. We repeat this exercise, varying the value of ϕ.

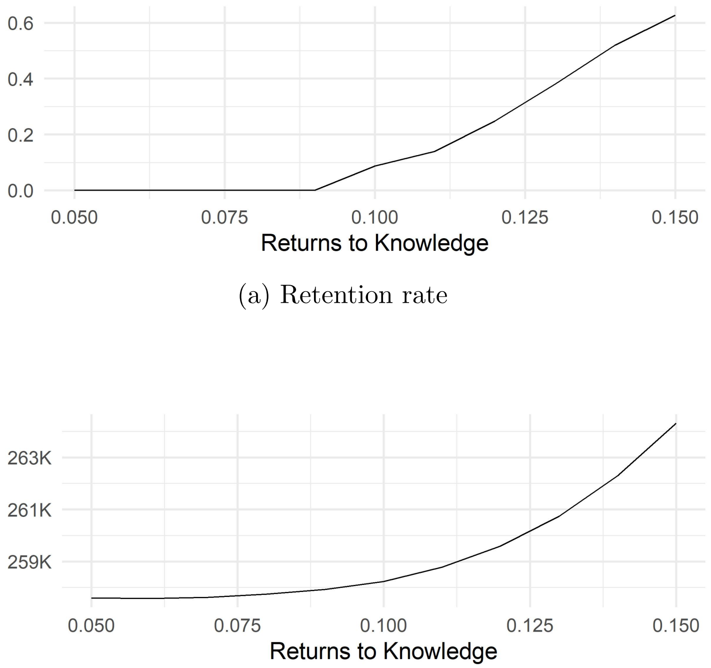
Figure 7 plots the implied fraction of high school students who are retained in any grade under the optimal retention policy for ϕ ranging between 0.05 and 0.15. For $\phi = 0.05$, the optimal retention policy is to not retain any students. As the returns increase, however, the cognitive benefits of retention begin to outweigh the costs of reduced educational attainment and delayed labor market entry. For returns of about 12%, which is the maximum return to skill estimated by Watts (2020) in the UK when controlling for family background and health, it is optimal to retain 24.8% of students. This retention rate is far below the rate of 44% observed in the Portuguese high school data we use in our analysis. The returns to cognitive skills would need to be high, about 14% per standard deviation, to justify the existing retention rate. Overall, the figure shows that

[33] Subsequent definitions of PRTEs, as in Mogstad et al. (2018), normalize Δ by the change in treatment probabilities across the two policy regimes. We work with the non-normalized PRTE, as we are interested in solving for the retention policy that maximizes the overall benefits for society.

[34] If compulsory schooling laws were not enforced and students were permitted to drop out before they reach age 18, then the costs would be even larger and Δ would equal $-6,234$ Euros.

whether retention is beneficial or costly in the long run depends on how much the labor market rewards math and language skills.

Figure 7: Optimal retention rates as a function of returns to knowledge



(a) Retention rate

(b) Discounted lifetime income

Note: The figures depicts the relationship between the returns to a standard deviation increase in knowledge along the x-axis and the fraction of students who are retained under the optimal retention policy in panel (a) and the expected discounted lifetime income under this policy in Euros (2013) in panel (b). The assumptions used to compute expected lifetime income are detailed in section 8.

Our cost-benefit analysis is subject to a few caveats. First, the calculations assume that individuals are continuously employed. There is a small literature, including studies in economics, sociology and psychology, that analyzes whether retained students experience adverse labor market consequences in terms of a higher probability of unemployment and lower levels of job security. Baert and Picchio (2021) summarizes much of this literature and describes some studies finding evidence of adverse effects on employment.[35] Second, our calculations only consider private earnings

[35] Baert and Picchio (2021) also present the results of an RCT that they carried out in Belgium that experimentally varied grade retention information on fictitious resumes. Their results showed that grade retention did not significantly

returns to schooling and cognitive skills. If there are substantial social returns, then the rankings of the different policy scenarios could be altered. Third, our cost-benefit analysis also did not account for possible positive impacts on future generations occurring through the inter-generational transmission of human capital. As seen in section 5, mother's and father's education levels are important determinants of youth's academic achievement. Allowing for inter-generational transmission of human capital would likely raise the benefits to the retention policy analyzed in this paper. Finally, we consider the costs and benefits of retention for the student only. Taking into account the costs of educating retained students for an additional year will lower the societal benefits of retention and require still higher returns to knowledge to justify the existing policy.

9 Conclusions

In recent decades, Portugal has made great strides in improving its population's skill levels. Between 2000 and 2020, the fraction of employees with a ninth grade education or less fell from nearly 80% of the total workforce to 40%. Over the same time period, employees with higher education increased from 9% to 30% of the workforce. In addition, there was significant improvement in Portuguese students' relative performance on international tests. Scores in three PISA subjects – Math, Reading, and Science – have been above the OECD average since 2015. Yet, Portugal continues to face significant challenges, including unusually high rates of grade retention and dropout.

This paper developed and implemented a dynamic value-added modeling approach for analyzing grade retention impacts. The model captures the cumulative nature of the educational production process including the additional learning that takes place during years when students repeat grades. To our knowledge, this type of switching value-added model, used to explicitly model how learning for retained children takes place over two separate years, has not been considered in the prior retention literature. The model incorporates rich observable heterogeneity on student, parent, and school characteristics as well as unobserved heterogeneity (in the form of unobserved types) that simultaneously affect achievement, retention and dropout.

affect positive call-back by employers but made it less likely that individuals got called back for jobs with a large training component.

The model estimates show fairly large positive average effects of retention on test scores for retained students, about 0.2 s.d.in math and 0.5 s.d.in Portuguese. However, we also find that retention discourages students from finishing high school. The causal effect of retention on dropout is a 53 pp increase in the probability of dropping out among retained students. The increase in dropout is greater for boys than girls and for students who are older than their same-grade peers. We uncover a pattern whereby students who entered high school with lower test scores experience the greatest cognitive gains from grade retention but also the greatest increase in dropout.

Our distributional analysis of retention impacts reveals substantial individual-level heterogeneity. Despite the overall average retention impacts being positive, about 30% of students experience negative impacts on 12th grade test scores. Some students finish high school with greater skill levels in math and Portuguese as a result of grade retention, but other students get discouraged and experience lower test scores or drop out of school altogether.

We conduct a cost-benefit analysis that compares the life-time earnings streams under different retention policies, accounting for individual impact heterogeneity as estimated. For typical values of the wage returns to knowledge that have been estimated in the literature, we find that the current retention policy is not, on average, beneficial to students, as the foregone earnings cost due to delayed labor market entry and reduced educational attainment (for those induced to drop out because of retention) exceeds the expected gain accruing from greater knowledge. A modified policy that retains a lower proportion of students and strictly enforces compulsory schooling laws to mitigate the dropout problem would lead to greater lifetime earnings gains. Portugal has moved in the direction of this optimal policy in recent years, with retention rates falling steadily over the last decade. The lessons derived from the Portuguese context may be of interest to other countries with high retention rates.

References

Daniel A Ackerberg. A new use of importance sampling to reduce computational burden in simulation estimation. *QME*, 7(4):343–376, 2009.

Francesco Agostinelli and Matthew Wiswall. Estimating the technology of children's skill formation. Technical report, National Bureau of Economic Research, 2016.

Chiharu S Allen, Qi Chen, Victor L Willson, and Jan N Hughes. Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, 31(4):480–499, 2009.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

Peter Arcidiacono, Holger Sieg, and Frank Sloan. Living rationally under the volcano? an empirical analysis of heavy drinking and smoking. *International Economic Review*, 48(1):37–65, 2007.

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

Erich Battistin and Antonio Schizzerotto. Threat of grade retention, remedial education and student achievement: Evidence from upper secondary schools in italy. 2012.

M Campos and Hugo Reis. Revisiting the returns to schooling in the portuguese economy. *Banco de Portugal Economic Studies*, 3(2):1–28, 2017.

Pedro Carneiro, Karsten T Hansen, and James J Heckman. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college, 2003.

John Cawley, James Heckman, and Edward Vytlacil. Three observations on wages and measured cognitive ability. *Labour economics*, 8(4):419–442, 2001.

Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, 126(4):1593–1660, 2011.

Bart Cockx, Matteo Picchio, and Stijn Baert. Modeling the effects of grade retention in high school. *Journal of Applied Econometrics*, 34(3):403–424, 2019.

Flavio Cunha, James J Heckman, Lance Lochner, and Dimitriy V Masterov. Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1:697–812, 2006.

Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

Christopher Dougherty. Numeracy, literacy and earnings: evidence from the national longitudinal survey of youth. *Economics of education review*, 22(5):511–521, 2003.

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

Eric R Eide and Mark H Showalter. The effect of grade retention on educational and labor market outcomes. *Economics of Education review*, 20(6):563–576, 2001.

Ozkan Eren, Briggs Depew, and Stephen Barnes. Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, 153:9–31, 2017.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018.

Maria Ferreira, Bart Golsteyn, and Sergio Parra-Cely. The effect of grade retention on secondary school performance: Evidence from a natural experiment. 2018.

David Figlio and Umut Özek. An extra year to learn english? early grade retention and the human capital development of english learners. *Journal of Public Economics*, 186:104184, 2020.

Jane Cooley Fruehwirth, Salvador Navarro, and Yuya Takahashi. How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, 34(4):979–1021, 2016.

Robert J Gary-Bobo, Marion Goussé, and Jean-Marc Robin. Grade retention and unobserved heterogeneity. *Quantitative Economics*, 7(3):781–820, 2016.

Monfort Gourieroux, Christian Gourieroux, Alain Monfort, Director Alain Monfort, et al. *Simulation-based econometric methods*. Oxford university press, 1996.

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

James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111, 2001.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

James J Heckman and Edward J Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734, 1999.

James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics*, 6:4875–5143, 2007.

James J Heckman, Robert J LaLonde, and Jeffrey A Smith. The economics and econometrics of active labor market programs. In *Handbook of labor economics*, volume 3, pages 1865–2097. Elsevier, 1999.

James J Heckman, Jora Stixrud, and Sergio Urzua. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24(3): 411–482, 2006a.

James J Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3):389–432, 2006b.

C Thomas Holmes and Kenneth M Matthews. The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of educational research*, 54(2):225–236, 1984.

Brian A Jacob and Lars Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, 86(1):226–244, 2004.

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

Shane R Jimerson. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School psychology review*, 30(3):420–437, 2001.

Michael P Keane and Kenneth I Wolpin. The career decisions of young men. *Journal of political Economy*, 105(3):473–522, 1997.

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

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619, 2018.

Richard J Murnane, John B Willett, Yves Duhaldeborde, and John H Tyler. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, 19(4):547–568, 2000.

M.C. Pereira and H. Reis. Grade retention during basic education in portugal: determinants and impact on student achievement. *Economic Bulletin*, pages 61–82, 2014.

Steven G Rivkin, Eric A Hanushek, and John F Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

Tom Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, Department of Computer Sciences, University of Texas at Austin, 1990.

Fernando Saltiel and Miguel Sarzosa. Grade retention and multidimensional skill formation in young children. 2020.

Guido Schwerdt, Martin R West, and Marcus A Winters. The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. *Journal of Public Economics*, 152:154–169, 2017.

Petra Todd and Kenneth I Wolpin. Accounting for mathematics performance of high school students in mexico: Estimating a coordination game in the classroom. *Journal of Political Economy*, 126 (6):2608–2650, 2018.

Petra E Todd and Kenneth I Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33, 2003.

Petra E Todd and Kenneth I Wolpin. The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital*, 1(1):91–136, 2007.

Tyler W Watts. Academic achievement and economic attainment: Reexamining associations between test scores and long-run earnings. *AERA Open*, 6(2):2332858420928985, 2020.

Marcus A Winters and Jay P Greene. The medium-run effects of florida's test-based promotion policy. *Education Finance and Policy*, 7(3):305–330, 2012.

Appendix A: Moments used in estimation

This appendix describes the moments used in estimation. As discussed in the main text, the estimation is based on three types of moments for students with different schooling trajectories, as summarized by their history h_i, which was defined in section 5. The moments are formed from three types of simulated dependent variables: $\mathbb{1}_{h_i=h}$, $\mathbb{1}_{h_i=h}K^g_{i,t}$, and $\mathbb{1}_{h_i=h}{K^g_{i,t}}^2$. That is, the moments depend on the proportion of students with each history, the mean knowledge levels (test scores) and the second moment of test scores for both math and Portuguese for students with different histories. Table 15 lists the histories that form the basis of the SMM estimation for the STEM track students together with the percentage of students in the estimation sample that have each history.

Table 15: Targeted Histories, STEM

History	Percentage
10-11-12X10-11-12	50.3%
10-10dX10-10d	8.77%
10-11-12-12-12X10-11-12 or 10-11-12-12-12dX10-11-12	5.20%
10-11-12-12X10-11-12	4.78%
10-11-12dX10-11-12d	2.09%
10-10-10dX10-10-10d	2.52%
10-11-12-12dX10-11-12	3.07%
10-11-11dX10-11-11d	2.12%
10-10-11-12X10-10-11-12 or 10-10-11-12dX10-10-11-12	3.36%
10-10-11-11dX10-10-11-11d	2.11%
10-11-11-11dX10-11-11-11d	1.66%
10-11dX10-11d	1.53%
10-11-11-12X10-11-11-12	1.20%
10-11-12-12X10-11-12-12	0.11%
10-11-12-12-12dX10-11-12-12-12d or 10-11-12-12-12X10-11-12-12-12	0.25%
	SUM: 89.1 %

Altogether, we target fifteen histories, which represent 89.1 percent of the STEM students in our estimation sample. For those histories in which twelfth grade math and Portuguese test scores are observed (histories 1, 3, 4, 7, 9, 13, 14, and 15), we additionally target moments involving the 12th grade test scores.

Moments are formed by multiplying the three types of endogenous variables (histories, test scores, and squared test scores) by exogenous variables and summing over all individuals. The moments, m_i, for those individuals enrolled in the STEM track are as follows. The histories $h_1, \ldots, h_{15}$ correspond to the fifteen targeted STEM histories in Table 15.

$$m_i = \begin{pmatrix} \mathbf{M^1} \\ \mathbf{M^2} \\ \mathbf{M^3} \end{pmatrix} \tag{16}$$

where

$$\mathbf{M^1} = \begin{pmatrix}
\mathbb{1}_{h_i=h_1} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,0})' \\
\mathbb{1}_{h_i=h_2} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,0})' \\
\mathbb{1}_{h_i=h_3} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,0})' \\
\mathbb{1}_{h_i=h_4} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,3})' \\
\mathbb{1}_{h_i=h_5} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,3})' \\
\mathbb{1}_{h_i=h_6} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,1})' \\
\mathbb{1}_{h_i=h_7} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,3})' \\
\mathbb{1}_{h_i=h_8} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,2}, X'_i, H^R_{i,2-3}, L'_{i,2})' \\
\mathbb{1}_{h_i=h_9} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,1})' \\
\mathbb{1}_{h_i=h_{10}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,3})' \\
\mathbb{1}_{h_i=h_{11}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,3})' \\
\mathbb{1}_{h_i=h_{12}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,1}, X'_i, H^R_{i,1-3}, L'_{i,1})' \\
\mathbb{1}_{h_i=h_{13}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})' \\
\mathbb{1}_{h_i=h_{14}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})' \\
\mathbb{1}_{h_i=h_{15}} \cdot (1, K_{i,0}^{9,M}, K_{i,0}^{9,P}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})'
\end{pmatrix},$$

The moments in $\mathbf{M^1}$ represent covariances between the variables in the retention and dropout equations and the history indicators, which identify the parameters of the dropout and retention equations. The moments in $\mathbf{M^2}$ are covariances between the covariates in the value added equations and the 12th grade math and Portuguese test scores for both retained and non-retained students. These moments identify the parameters in the value added equations. The moments in $\mathbf{M^3}$ are squared test scores for students with each history in which test scores are observable. These moments identify the variances of the error terms in the value added equations.

$$
\mathbf{M^2} = \begin{pmatrix}
\mathbb{1}_{h_i=h_1} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_1} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,1}, X'_i)' \\
\mathbb{1}_{h_i=h_3} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_3} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_3} K_{i,4}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_4} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_4} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_4} K_{i,4}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_7} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_7} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_9} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_{13}} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,1}, X'_i)' \\
\mathbb{1}_{h_i=h_{13}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,1}, X'_i)' \\
\mathbb{1}_{h_i=h_{14}} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_{14}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_{14}} K_{i,4}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_{14}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_{15}} K_{i,3}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_{15}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,3}, X'_i)' \\
\mathbb{1}_{h_i=h_{15}} K_{i,4}^{12,M} \cdot (1, K_{i,0}^{9,M}, I'_{i,4}, X'_i)' \\
\mathbb{1}_{h_i=h_{15}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I'_{i,4}, X'_i)'
\end{pmatrix}, \quad
\mathbf{M^3} = \begin{pmatrix}
\mathbb{1}_{h_i=h_1} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_1} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_3} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_3} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_3} {K_{i,4}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_4} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_4} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_4} {K_{i,4}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_7} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_7} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_9} {K_{i,4}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_{13}} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_{13}} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_{14}} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_{14}} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_{14}} {K_{i,4}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_{14}} {K_{i,4}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_{15}} {K_{i,3}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_{15}} {K_{i,3}^{12,P}}^2 \\
\mathbb{1}_{h_i=h_{15}} {K_{i,4}^{12,M}}^2 \\
\mathbb{1}_{h_i=h_{15}} {K_{i,4}^{12,P}}^2
\end{pmatrix}.
$$

Our vector of moments used in estimation also includes moments for students in the non-STEM track. For non-STEM students, the test score moments only relate to Portuguese scores, because these students do not take math exams. Table 16 lists the histories that we target for the students enrolled in the non-STEM track together with the percentage of the students in the estimation sample that have each history.

Table 16: Targeted Histories, non-STEM

History	Percentage
X10-11-12	75.8%
X10-10-11-12	10.7%
X10-11-12-12d	3.15%
X10-11-12-12-12d	2.46%
X10-11-12-12	2.08%
X10-11-11-12	1.77%
X10-10-11-12-12 or X10-10-11-12-12d	1.62%
X10-10-11-11-12	0.66%
X10-11-11-12-12 or X10-11-11-12-12d	0.54%
	SUM: 98.7 %

The nine histories in table 16 represent 98.7 percent of the non-STEM students in our estimation sample. For these non-STEM students, we include 12th grade Portuguese test score moments, $\mathbb{1}_{(h_i=h)}K^P_{i,t}$ and $\mathbb{1}_{(h_i=h)}{K^P_{i,t}}^2$ for those histories which permit a 12th grade Portuguese test score to be observed (all histories except history 8, in which test scores are only observed after the 5th year of high school enrollment and thus beyond the reach of our data, which covers ninth grade plus up to four years of high school).

Moments for these non-STEM students are formed as described above, by multiplying the endogenous variables (history indicators, test scores, and squared test scores) by the covariates in the value-added, retention, and dropout equations and summing over all individuals. These moments are given below, where the histories $h^{NS}_1, \dots h^{NS}_9$ correspond to the nine targeted non-STEM histories shown in Table 16.

$$m_i^{NS} = \begin{pmatrix} \mathbf{M^{1,NS}} \\ \mathbf{M^{2,NS}} \\ \mathbf{M^{3,NS}} \end{pmatrix}$$

where

$$\mathbf{M^{1,NS}} = \begin{pmatrix} \mathbb{1}_{h_i=h_1^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,1}, X'_i, H^R_{i,1-3})' \\ \mathbb{1}_{h_i=h_2^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,1}, X'_i, H^R_{i,1-3})' \\ \mathbb{1}_{h_i=h_3^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,4})' \\ \mathbb{1}_{h_i=h_4^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})' \\ \mathbb{1}_{h_i=h_5^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,3}, X'_i, H^R_{i,3-3}, L'_{i,4})' \\ \mathbb{1}_{h_i=h_6^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,2}, X'_i, H^R_{i,2-3}, L'_{i,2})' \\ \mathbb{1}_{h_i=h_7^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})' \\ \mathbb{1}_{h_i=h_8^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,3}, X'_i, H^R_{i,3-3})' \\ \mathbb{1}_{h_i=h_9^{NS}} \cdot (1, K^{9,P}_{i,0}, I'_{i,4}, X'_i, H^R_{i,4-3}, L'_{i,4})' \end{pmatrix},$$

$$\mathbf{M^{2,NS}} = \begin{pmatrix} \mathbb{1}_{h_i=h_1^{NS}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,1}', X_i')' \\ \mathbb{1}_{h_i=h_2^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,1}', X_i')' \\ \mathbb{1}_{h_i=h_3^{NS}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,3}', X_i')' \\ \mathbb{1}_{h_i=h_4^{NS}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,4}', X_i')' \\ \mathbb{1}_{h_i=h_4^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,4}', X_i')' \\ \mathbb{1}_{h_i=h_5^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,3}', X_i')' \\ \mathbb{1}_{h_i=h_6^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,2}', X_i')' \\ \mathbb{1}_{h_i=h_7^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,4}', X_i')' \\ \mathbb{1}_{h_i=h_9^{NS}} K_{i,3}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,4}', X_i')' \\ \mathbb{1}_{h_i=h_9^{NS}} K_{i,4}^{12,P} \cdot (1, K_{i,0}^{9,P}, I_{i,4}', X_i')' \end{pmatrix}, \qquad \mathbf{M^{3,NS}} = \begin{pmatrix} \mathbb{1}_{h_i=h_1^{NS}} K_{i,3}^{12,P^2} \\ \mathbb{1}_{h_i=h_2^{NS}} K_{i,4}^{12,P^2} \\ \mathbb{1}_{h_i=h_3^{NS}} K_{i,3}^{12,P^2} \\ \mathbb{1}_{h_i=h_4^{NS}} K_{i,3}^{12,P^2} \\ \mathbb{1}_{h_i=h_4^{NS}} K_{i,4}^{12,P^2} \\ \mathbb{1}_{h_i=h_5^{NS}} K_{i,4}^{12,P^2} \\ \mathbb{1}_{h_i=h_6^{NS}} K_{i,4}^{12,P^2} \\ \mathbb{1}_{h_i=h_7^{NS}} K_{i,4}^{12,P^2} \\ \mathbb{1}_{h_i=h_9^{NS}} K_{i,3}^{12,P^2} \\ \mathbb{1}_{h_i=h_9^{NS}} K_{i,4}^{12,P^2} \end{pmatrix}.$$

As before, the moments in $\mathbf{M^{1,NS}}$ represent covariances between the covariates in the retention and dropout equations and the history indicators, the moments in $\mathbf{M^{2,NS}}$ are covariances between the covariates in the value added equations and the endogenous twelfth grade Portuguese test scores for both retained and non-retained students, and the moments in $\mathbf{M^{3,NS}}$ are squared test scores for students with each history in which test scores are observable.

Appendix B: Method used to smooth the objective function (online)

This section describes how we smooth our objective function to facilitate its minimization. In many Simulated Method of Moments problems, a simple frequency simulator is used to simulate the model dependent variables. A set of S errors, ϵ_i^s, are drawn for each individual, $i = 1, \ldots, N$ and $s = 1 \ldots S$, and, given a guess for the model parameters, the model is simulated S times and the frequency simulator is calculated as $\bar{f}(z_i, \epsilon_i, y_0; \theta) = \frac{1}{S}\sum_{s=1}^{S} f(z_i, \epsilon_i^s, y_0; \theta)$. Our model contains discrete dropout and retention decisions, and hence the frequency simulator produces an objective function, $J(\theta)$, that is discontinuous in θ. Due to difficulties associated with optimizing a discontinuous function with a large parameter space, we do not use a frequency simulator.

Instead, following suggestions in Ackerberg (2009), we combine importance sampling with a change of variable in integration to produce a simulator, and hence objective function, that is continuous in θ. Suppose, first, that we can write $f(z_i, \epsilon_i, y_0; \theta) = \tilde{f}(u(z_i, \epsilon_i, y_0; \theta))$, so that the simulator is a new function of an index vector, u. The objective function therefore depends on the parameter vector, θ, only through u.

We then make use of the following identities to obtain a simulator that is smooth in θ:

$$\begin{aligned}
\mathbb{E}[\tilde{f}(u_i)] &= \int \tilde{f}(u_i) p(u_i \mid z_i, y_{i,0}, \theta) du_i \\
&= \int \tilde{f}(u_i) \frac{p(u_i \mid z_i, y_{i,0}, \theta)}{g(u_i \mid z_i, y_{i,0})} g(u_i \mid z_i, y_0) du_i \\
&\approx \frac{1}{S}\sum_{s=1}^{S} \tilde{f}(u_{i,s}) \frac{p(u_{i,s} \mid z_i, y_{i,0}, \theta)}{g(u_{i,s} \mid z_i, y_{i,0})} \ , \qquad (17)
\end{aligned}$$

where the analog estimator in (17) depends on θ only through the density $p(u_{i,s} \mid z_i, y_{i,0}, \theta)$. The simulator works by initially drawing a large number deviates, $u_{i,s}$, for each individual from the density $g(u_{i,s} \mid z_i, y_{i,0})$ and precomputing $\tilde{f}(u_{i,s})$ at the simulated deviates. Then, as θ varies during optimization, only the weight on these precomputed values, given by $\frac{p(u_{i,s}|z_i, y_{i,0}, \theta)}{g(u_{i,s}|z_i, y_{i,0})}$, changes. The simulator is continuous for the purposes of optimization, because all the discontinuities of the original function are transferred to $\tilde{f}(u_{i,s})$, which is precomputed. These precomputed values are then reweighted during optimization according to weights derived from a continuous density function, $p(u_{i,s} \mid z_i, y_{i,0}, \theta)$.

The key to implementing this method is finding the function, $\tilde{f}(u_{i,s})$, that satisfies $\tilde{f}(u_{i,s}(z_i, \epsilon_i, y_0; \theta)) = f(z_i, \epsilon_i, y_0; \theta)$, where $u_{i,s}$ has a known and continuous density function. The dependent variables whose moments we target have the form of $\mathbb{1}_{(h_i=h)}$, $\mathbb{1}_{(h_i=h)} K_{i,t}^{g}$, or $\mathbb{1}_{(h_i=h)} {K_{i,t}^{g}}^2$. Noting that a particular history depends on a unique series of dropout and retention decisions, and that the dropout and retention decisions are themselves indicator functions of continuous indices, we can therefore let u be a vector of these continuous indices. $\tilde{f}(\cdot)$ then applies a series of indicator functions to the elements of u to simulate moments of the form $\mathbb{1}_{(h_i=h)}$. In addition, since we are also interested in the endogenous test scores, we append the formulas for these scores to the end u.

In what follows we show how u is constructed, and we present the function $f(\cdot)$ that transforms u into the endogenous variables that comprise our moments. First it is useful to break down the simulated test scores for each period into a part that depends on observed exogenous variables, $\overline{K}_{i,t}^{g}$, and a part that depends on unobserved shocks. We do this in equations (18) through (27) for math. The formulas for simulated Portuguese test scores are analogous.

$$K_{i,1}^{10,M} = \underbrace{\gamma^M K_{i,0}^{9,M} + \beta^M I_{i,1} + \sum_{m=1}^{M} \alpha^{m,M} \mu_i^m}_{\overline{K}_{i,1}^{10,M}} + \varepsilon_{i,1}^{M} \, , \tag{18}$$

$$K_{i,2}^{11,M} = \underbrace{{\gamma^M}^2 K_{i,1}^{9,M} + \beta^M I_{i,2} + \gamma^M \beta^M I_{i,1} + (1+\gamma^M) \sum_{m=1}^{M} \alpha^{m,M} \mu_i^m}_{\overline{K}_{i,2}^{11,M}} + \varepsilon_{i,2}^{M} + \gamma^M \varepsilon_{i,1}^{M} \, , \tag{19}$$

$$K_{i,3}^{12,M} = \underbrace{{\gamma^M}^3 K_{i,1}^{9,M} + \beta^M I_{i,2} + \gamma^M \beta^M I_{i,2} + {\gamma^M}^2 \beta^M I_{i,1} + (1+\gamma^M+{\gamma^M}^2) \sum_{m=1}^{M} \alpha^{m,M} \mu_i^m +}_{\overline{K}_{i,3}^{12,M}}$$

$$\varepsilon_{i,3}^{M} + \gamma^M \varepsilon_{i,2}^{M} + {\gamma^M}^2 \varepsilon_{i,1}^{M} \, , \tag{20}$$

$$K_{i,2}^{10,MR} = \underbrace{\gamma^{MR}\overline{K}_{i,1}^{10,M} + \beta^{MR} I_{i,2} + \sum_{m=1}^{M} \alpha^{m,MR}\mu_i^m}_{\overline{K}_{i,2}^{10,MR}} + \varepsilon_{i,2}^{MR} + \gamma^{MR}\varepsilon_{i,1}^{M} , \tag{21}$$

$$\begin{aligned} K_{i,3}^{11,M} &= \underbrace{\gamma^{M}\overline{K}_{i,2}^{10,MR} + \beta^{M} I_{i,3} + \sum_{m=1}^{M} \alpha^{m,M}\mu_i^m +}_{\overline{K}_{i,3}^{11,M}} \\ &\quad \varepsilon_{i,3}^{M} + \gamma^{M}\varepsilon_{i,2}^{MR} + \gamma^{M}\gamma^{MR}\varepsilon_{i,1}^{M} , \end{aligned} \tag{22}$$

$$\begin{aligned} K_{i,4}^{12,M} &= \underbrace{\gamma^{M}\overline{K}_{i,3}^{11,M} + \beta^{M} I_{i,4} + \sum_{m=1}^{M} \alpha^{m,M}\mu_i^m +}_{\overline{K}_{i,4}^{12,M}} \\ &\quad \varepsilon_{i,4}^{M} + \gamma^{M}\varepsilon_{i,3}^{M} + \gamma^{M^2}\varepsilon_{i,2}^{MR} + \gamma^{M^2}\gamma^{MR}\varepsilon_{i,1}^{M} , \end{aligned} \tag{23}$$

$$\begin{aligned} K_{i,4}^{11,MR} &= \underbrace{\gamma^{MR}\overline{K}_{i,3}^{11,M} + \beta^{MR} I_{i,4} + \sum_{m=1}^{M} \alpha^{m,MR}\mu_i^m +}_{\overline{K}_{i,4}^{11,MR}} \\ &\quad \varepsilon_{i,4}^{MR} + \gamma^{MR}\varepsilon_{i,3}^{M} + \gamma^{M}\gamma^{MR}\varepsilon_{i,2}^{MR} + \gamma^{M^2}\gamma^{MR}\varepsilon_{i,1}^{M} , \end{aligned} \tag{24}$$

$$K_{i,3}^{11,MR} = \underbrace{\gamma^{M}\overline{K}_{i,2}^{11,M} + \beta^{MR} I_{i,3} + \sum_{m=1}^{M} \alpha^{m,MR}\mu_i^m}_{\overline{K}_{i,3}^{11,MR}} + \varepsilon_{i,3}^{MR} + \gamma^{MR}\varepsilon_{i,2}^{M} + \gamma^{MR}\gamma^{M}\varepsilon_{i,1}^{M} , \tag{25}$$

$$\begin{aligned} \tilde{K}_{i,4}^{12,M} &= \underbrace{\gamma^{M}\overline{K}_{i,3}^{11,MR} + \beta^{M} I_{i,4} + \sum_{m=1}^{M} \alpha^{m,M}\mu_i^m +}_{\overline{\tilde{K}}_{i,4}^{12,M}} \\ &\quad \varepsilon_{i,4}^{M} + \gamma^{M}\varepsilon_{i,3}^{MR} + \gamma^{M}\gamma^{MR}\varepsilon_{i,2}^{M} + \gamma^{MR}\gamma^{M^2}\varepsilon_{i,1}^{M} , \end{aligned} \tag{26}$$

$$\begin{aligned} \tilde{K}_{i,4}^{12,MR} &= \underbrace{\gamma^{MR}\overline{K}_{i,3}^{12,M} + \beta^{MR} I_{i,4} + \sum_{m=1}^{M} \alpha^{m,MR}\mu_i^m +}_{\overline{\tilde{K}}_{i,4}^{12,MR}} \\ &\quad \varepsilon_{i,4}^{MR} + \gamma^{MR}\varepsilon_{i,3}^{M} + \gamma^{M}\gamma^{MR}\varepsilon_{i,2}^{M} + \gamma^{MR}\gamma^{M^2}\varepsilon_{i,1}^{M} , \end{aligned} \tag{27}$$

The deterministic parts of simulated knowledge, $\overline{K}_{i,t}^{g}$, will form part of the mean of u_i, while the shocks will be part of the error. We are now ready to write the vector u_i as $u_i = mean_i + error_i$, where $mean_i$ depends exclusively on observed variables and $error_i$ depends exclusively on unobserved shocks. The mean vector is as follows:

$$mean_i = \begin{pmatrix}
\delta_0^{11} + \delta_1^{11}\overline{K}_{i,1}^{10,M} + \delta_2^{11}\overline{K}_{i,1}^{10,P} + \delta_3^{11}L_{i,2} + \delta_4^{11}I_{i,1} + \delta_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\delta_0^{12} + \delta_1^{12}\overline{K}_{i,2}^{11,M} + \delta_2^{12}\overline{K}_{i,2}^{11,P} + \delta_3^{12}L_{i,3} + \delta_4^{12}I_{i,2} + \delta_5^{12}X_i + \sum_{m=1}^{M}\alpha_D^{12,m}\mu_i^m \\
\lambda_0^{10} + \lambda_1^{10}\overline{K}_{i,1}^{10,M} + \lambda_2^{10}\overline{K}_{i,1}^{10,P} + \lambda_3^{10}H_{i,1-3}^{R} + \lambda_4^{10}I_{i,1} + \lambda_5^{10}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,10}\mu_i^m \\
\lambda_0^{11} + \lambda_1^{11}\overline{K}_{i,2}^{11,M} + \lambda_2^{11}\overline{K}_{i,2}^{11,P} + \lambda_3^{11}H_{i,2-3}^{R} + \lambda_4^{11}I_{i,2} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,11}\mu_i^m \\
\lambda_0^{12,M} + \lambda_1^{12,M}\overline{K}_{i,3}^{12,M} + \lambda_2^{12,M}H_{i,3-3}^{R} + \lambda_3^{12,M}I_{i,3} + \lambda_4^{12,M}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12M}\mu_i^m \\
\lambda_0^{12,P} + \lambda_1^{12,P}\overline{K}_{i,3}^{12,P} + \lambda_2^{12,P}H_{i,3-3}^{R} + \lambda_3^{12,P}I_{i,3} + \lambda_4^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12P}\mu_i^m \\
\delta_0^{10} + \delta_1^{10}\overline{K}_{i,1}^{10,M} + \delta_2^{10}\overline{K}_{i,1}^{10,P} + \delta_3^{10}L_{i,2} + \delta_4^{10}I_{i,1} + \delta_5^{10}X_i + \sum_{m=1}^{M}\alpha_D^{10,m}\mu_i^m \\
\delta_0^{12} + \delta_1^{12}\overline{K}_{i,3}^{12,M} + \delta_2^{12}\overline{K}_{i,3}^{12,P} + \delta_3^{12}L_{i,4} + \delta_4^{12}I_{i,3} + \delta_5^{12}X_i + \sum_{m=1}^{M}\alpha_D^{12,m}\mu_i^m \\
\lambda_0^{12,M} + \lambda_1^{12,M}\overline{\tilde{K}}_{i,4}^{12,MR} + \lambda_2^{12,M}H_{i,4-3}^{R} + \lambda_3^{12,M}I_{i,4} + \lambda_4^{12,M}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12M}\mu_i^m \\
\delta_0^{10} + \delta_1^{10}\overline{K}_{i,2}^{10,MR} + \delta_2^{10}\overline{K}_{i,2}^{10,PR} + \delta_3^{10}L_{i,3} + \delta_4^{10}I_{i,2} + \delta_5^{10}X_i + \sum_{m=1}^{M}\alpha_D^{10,m}\mu_i^m \\
\lambda_0^{10} + \lambda_1^{10}\overline{K}_{i,2}^{10,MR} + \lambda_2^{10}\overline{K}_{i,2}^{10,PR} + \lambda_3^{10}H_{i,2-3} + \lambda_4^{10}I_{i,2} + \lambda_5^{10}X_i + \sum_{m=1}^{M}\alpha_{RET}^{10,m}\mu_i^m \\
\delta_0^{11} + \delta_1^{11}\overline{K}_{i,2}^{11,M} + \delta_2^{11}\overline{K}_{i,2}^{11,P} + \delta_3^{11}L_{i,3} + \delta_4^{11}I_{i,2} + \delta_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\delta_0^{11} + \delta_1^{11}\overline{K}_{i,2}^{10,MR} + \delta_2^{11}\overline{K}_{i,2}^{10,PR} + \delta_3^{11}L_{i,3} + \delta_4^{11}I_{i,2} + \delta_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\delta_0^{12} + \delta_1^{12}\overline{K}_{i,3}^{11,M} + \delta_2^{12}\overline{K}_{i,3}^{11,P} + \delta_3^{12}L_{i,4} + \delta_4^{12}I_{i,3} + \delta_5^{12}X_i + \sum_{m=1}^{M}\alpha_D^{12,m}\mu_i^m \\
\lambda_0^{11} + \lambda_1^{11}\overline{K}_{i,3}^{11,M} + \lambda_2^{11}\overline{K}_{i,3}^{11,P} + \lambda_3^{11}H_{i,3-3} + \lambda_4^{11}I_{i,3} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{11,m}\mu_i^m \\
\lambda_0^{12,M} + \lambda_1^{12,M}\overline{K}_{i,4}^{12,M} + \lambda_2^{12,M}H_{i,4-3}^{R} + \lambda_3^{12,M}I_{i,4} + \lambda_4^{12,M}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12M}\mu_i^m \\
\lambda_0^{12,P} + \lambda_1^{12,P}\overline{K}_{i,4}^{12,P} + \lambda_2^{12,P}H_{i,4-3}^{R} + \lambda_3^{12,P}I_{i,4} + \lambda_4^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12P}\mu_i^m \\
\delta_0^{11} + \delta_1^{11}\overline{K}_{i,3}^{11,M} + \delta_2^{11}\overline{K}_{i,3}^{11,P} + \delta_3^{11}L_{i,4} + \delta_4^{11}I_{i,3} + \delta_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\delta_0^{11} + \delta_1^{11}\overline{K}_{i,3}^{11,MR} + \delta_2^{11}\overline{K}_{i,3}^{11,PR} + \delta_3^{11}L_{i,4} + \delta_4^{11}I_{i,3} + \delta_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\lambda_0^{11} + \lambda_1^{11}\overline{K}_{i,3}^{11,MR} + \lambda_2^{11}\overline{K}_{i,3}^{11,PR} + \lambda_3^{11}L_{i,4} + \lambda_4^{11}I_{i,3} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_D^{11,m}\mu_i^m \\
\delta_0^{12} + \delta_1^{12}\overline{K}_{i,3}^{11,MR} + \delta_2^{12}\overline{K}_{i,3}^{12,PR} + \delta_3^{12}L_{i,4} + \delta_4^{12}I_{i,3} + \delta_5^{12}X_i + \sum_{m=1}^{M}\alpha_D^{12,m}\mu_i^m \\
\lambda_0^{12,M} + \lambda_1^{12,M}\overline{\tilde{K}}_{i,4}^{12,M} + \lambda_2^{12,M}\overline{\tilde{K}}_{i,3}^{12,P} + \lambda_3^{12}H_{i,4-3} + \lambda_4^{12}I_{i,3} + \lambda_5^{12}X_i + \sum_{m=1}^{M}\alpha_D^{m,12M}\mu_i^m \\
\lambda_0^{12,P} + \lambda_1^{12,P}\overline{\tilde{K}}_{i,4}^{12,P} + \lambda_2^{12,P}\overline{\tilde{K}}_{i,3}^{12,P} + \lambda_3^{12}H_{i,4-3} + \lambda_4^{12}I_{i,3} + \lambda_5^{12}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12P}\mu_i^m \\
\lambda_0^{12,P} + \lambda_1^{12,P}\overline{\tilde{K}}_{i,4}^{12,PR} + \lambda_2^{12,P}\overline{\tilde{K}}_{i,4}^{12,PR} + \lambda_3^{12,P}H_{i,4-3} + \lambda_4^{12,P}I_{i,3} + \lambda_5^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12P}\mu_i^m \\
\overline{K}_{i,3}^{12,M} \\
\overline{K}_{i,4}^{12,M} \\
\overline{K}_{i,4}^{12,M} \\
\overline{K}_{i,4}^{12,MR} \\
\overline{K}_{i,3}^{12,P} \\
\overline{K}_{i,4}^{12,P} \\
\overline{K}_{i,4}^{12,P} \\
\overline{K}_{i,4}^{12,PR}
\end{pmatrix}$$

The error vector is as follows:

$$error_i = \begin{pmatrix}
\eta_{i,2} + \delta_1^{11}\varepsilon_{i,1}^M + \delta_2^{11}\varepsilon_{i,1}^P \\
\eta_{i,3} + \delta_1^{12}(\varepsilon_{i,2}^M + \gamma^M\varepsilon_{i,1}^M) + \delta_2^{12}(\varepsilon_{i,2}^P + \gamma^P\varepsilon_{i,1}^P) \\
\nu_{i,1} + \lambda_1^{10}\varepsilon_{i,1}^M + \lambda_2^{10}\varepsilon_{i,1}^P \\
\nu_{i,2} + \lambda_1^{11}(\varepsilon_{i,2}^M + \gamma^M\varepsilon_{i,1}^M) + \lambda_2^{11}(\varepsilon_{i,2}^P + \gamma^P\varepsilon_{i,1}^P) \\
\nu_{i,3}^M + \lambda_1^{12,M}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^M + {\gamma^M}^2\varepsilon_{i,1}^M) \\
\nu_{i,3}^P + \lambda_1^{12,P}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + {\gamma^P}^2\varepsilon_{i,1}^P) \\
\eta_{i,2} + \delta_1^{10}\varepsilon_{i,1}^M + \delta_2^{10}\varepsilon_{i,1}^P \\
\eta_{i,4} + \delta_1^{12}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^M + {\gamma^M}^2\varepsilon_{i,1}^M) + \delta_2^{12}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + {\gamma^P}^2\varepsilon_{i,1}^P) \\
\nu_{i,4}^M + \lambda_1^{12,M}(\varepsilon_{i,4}^{MR} + \gamma^{MR}\varepsilon_{i,3}^M + \gamma^{MR}\gamma^M\varepsilon_{i,2}^M + {\gamma^M}^2\gamma^{MR}\varepsilon_{i,1}^M) \\
\eta_{i,3}^M + \delta_1^{10,M}(\varepsilon_{i,2}^{MR} + \gamma^{MR}\varepsilon_{i,1}^M) + \delta_1^{10,P}(\varepsilon_{i,2}^{PR} + \gamma^{PR}\varepsilon_{i,1}^P) \\
\nu_{i,2}^M + \lambda_1^{10}(\varepsilon_{i,2}^{MR} + \gamma^{MR}\varepsilon_{i,1}^M) + \lambda_1^{10}(\varepsilon_{i,2}^{PR} + \gamma^{PR}\varepsilon_{i,1}^P) \\
\eta_{i,3}^M + \delta_1^{11}(\varepsilon_{i,2}^M + \gamma^M\varepsilon_{i,1}^M) + \delta_1^{11}(\varepsilon_{i,2}^P + \gamma^P\varepsilon_{i,1}^P) \\
\eta_{i,3}^M + \delta_1^{11}(\varepsilon_{i,2}^{MR} + \gamma^{MR}\varepsilon_{i,1}^M) + \delta_1^{11}(\varepsilon_{i,2}^{PR} + \gamma^{PR}\varepsilon_{i,1}^P) \\
\eta_{i,4}^M + \delta_1^{12}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^{MR} + \gamma^M\gamma^{MR}\varepsilon_{i,1}^M) + \delta_2^{12}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,1}^P) \\
\nu_{i,3}^P + \lambda_1^{11}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^{MR} + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) + \lambda_1^{11}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^{MR} + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) \\
\nu_{i,4}^M + \lambda_1^{12,M}(\varepsilon_{i,4}^M + \gamma^M\varepsilon_{i,3}^M + {\gamma^M}^2\varepsilon_{i,2}^{MR} + \gamma^{MR}{\gamma^M}^2\varepsilon_{i,1}^M) \\
\nu_{i,4}^P + \lambda_1^{12,P}(\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^P + {\gamma^P}^2\varepsilon_{i,2}^{PR} + \gamma^{PR}{\gamma^P}^2\varepsilon_{i,1}^P) \\
\eta_{i,4} + \delta_1^{11}(\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^{MR} + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) + \delta_2^{11}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^{PR} + \gamma^{PR}\gamma^P\varepsilon_{i,1}^P) \\
\eta_{i,4} + \delta_1^{11}(\varepsilon_{i,3}^{MR} + \gamma^{MR}\varepsilon_{i,2}^M + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) + \delta_2^{11}(\varepsilon_{i,3}^{PR} + \gamma^{PR}\varepsilon_{i,2}^P + \gamma^{PR}\gamma^P\varepsilon_{i,1}^P) \\
\nu_{i,4}^P + \lambda_1^{11}(\varepsilon_{i,3}^{MR} + \gamma^{MR}\varepsilon_{i,2}^M + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) + \lambda_2^{11}(\varepsilon_{i,3}^{PR} + \gamma^{PR}\varepsilon_{i,2}^P + \gamma^{PR}\gamma^P\varepsilon_{i,1}^P) \\
\eta_{i,4} + \delta_1^{12}(\varepsilon_{i,3}^{MR} + \gamma^{MR}\varepsilon_{i,2}^M + \gamma^{MR}\gamma^M\varepsilon_{i,1}^M) + \delta_2^{12}(\varepsilon_{i,3}^{PR} + \gamma^{PR}\varepsilon_{i,2}^P + \gamma^{PR}\gamma^P\varepsilon_{i,1}^P) \\
\nu_{i,4}^M + \lambda_1^{12,M}(\varepsilon_{i,4}^M + \gamma^M\varepsilon_{i,3}^{MR} + \gamma^M\gamma^{MR}\varepsilon_{i,2}^M + {\gamma^M}^2\gamma^{MR}\varepsilon_{i,1}^M) \\
\nu_{i,4}^P + \lambda_1^{12,P}(\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + {\gamma^P}^2\gamma^{PR}\varepsilon_{i,1}^P) \\
\nu_{i,4}^P + \lambda_1^{12,P}(\varepsilon_{i,4}^{PR} + \gamma^{PR}\varepsilon_{i,3}^P + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + {\gamma^P}^2\gamma^{PR}\varepsilon_{i,1}^P) \\
\varepsilon_{i,3}^M + \gamma^M\varepsilon_{i,2}^M + {\gamma^M}^2\varepsilon_{i,1}^M \\
\varepsilon_{i,4}^M + \gamma^M\varepsilon_{i,3}^M + {\gamma^M}^2\varepsilon_{i,2}^{MR} + {\gamma^M}^2\gamma^{MR}\varepsilon_{i,1}^M \\
\varepsilon_{i,4}^M + \gamma^M\varepsilon_{i,3}^{MR} + \gamma^M\gamma^{MR}\varepsilon_{i,2}^M + {\gamma^M}^2\gamma^{MR}\varepsilon_{i,1}^M \\
\varepsilon_{i,4}^{MR} + \gamma^{MR}\varepsilon_{i,3}^M + \gamma^M\gamma^{MR}\varepsilon_{i,2}^M + {\gamma^M}^2\gamma^{MR}\varepsilon_{i,1}^M \\
\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + {\gamma^P}^2\varepsilon_{i,1}^P \\
\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^P + {\gamma^P}^2\varepsilon_{i,2}^{PR} + {\gamma^P}^2\gamma^{PR}\varepsilon_{i,1}^P \\
\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + {\gamma^P}^2\gamma^{PR}\varepsilon_{i,1}^P \\
\varepsilon_{i,4}^{PR} + \gamma^{PR}\varepsilon_{i,3}^P + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + {\gamma^P}^2\gamma^{PR}\varepsilon_{i,1}^P
\end{pmatrix}$$

Sampling u_i, therefore involves repeatedly drawing from a multivariate normal distribution with mean given by $mean_i$ and covariance matrix given by the outer product of $error_i$. A .csv file containing the covariance matrix is available from the authors upon request.

We have written the vector of means, $mean_i$, as depending on the type of the individual, μ_i^m. While these types are unobserved, incorporating them in the importance sampler is straightforward. The effect of the types is to shift the intercepts in each of the dropout, retention, and value added equations. The types therefore affect $mean_i$, but not $error_i$ and therefore not the variance of the sampling distribution. When incorporating the unobserved types, we sample $u_{i,s}$ from the multivariate normal distribution with mean equal to $p_1 \times mean_i(\mu_i = 1) + p_2 \times mean_i(\mu_i = 2) + (1 - p_1 - p_2) \times mean_i(\mu_i = 3)$ and variance given by the outerproduct of $error_i$, where $(p_1, p_2, 1 - p_1 - p_2)$ are the probabilities of types one, two, and three, respectively.

For each simulation s, the dependent variables $F(u_{i,s})$ are precomputed. The function $F(\cdot)$ that transforms the deviates into simulated dependent variables is as follows:

$$F(u_{i,s}) = \begin{pmatrix} \mathbf{h^1}(u_{i,s}) \\ \mathbf{h^2}(u_{i,s}) \\ \mathbf{h^3}(u_{i,s}) \end{pmatrix} \tag{28}$$

where

$$\mathbf{h^1}(u) = \big(h_1(u), h_2(u), ..., h_{14}(u), h_{15}(u)\big)' =$$

$$\begin{pmatrix}
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} < 0, u^{(6)} < 0) \\
\mathbb{1}(u^{(3)} > 0, u^{(7)} > 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} > 0, u^{(6)} < 0, u^{(8)} < 0, u^{(9)} > 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} > 0, u^{(6)} < 0, u^{(8)} < 0, u^{(9)} < 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} > 0, u^{(3)} < 0, u^{(4)} < 0) \\
\mathbb{1}(u^{(3)} > 0, u^{(7)} < 0, u^{(10)} > 0, u^{(11)} > 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} > 0, u^{(6)} < 0, u^{(8)} > 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(3)} < 0, u^{(4)} > 0, u^{(12)} > 0) \\
\mathbb{1}(u^{(3)} > 0, u^{(7)} < 0, u^{(11)} < 0, u^{(13)} < 0, u^{(14)} < 0, u^{(15)} < 0, u^{(16)} < 0, u^{(17)} < 0) \\
\mathbb{1}(u^{(3)} > 0, u^{(7)} < 0, u^{(11)} < 0, u^{(13)} < 0, u^{(15)} > 0, u^{(18)} > 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(3)} < 0, u^{(4)} > 0, u^{(12)} < 0, u^{(19)} > 0, u^{(20)} > 0) \\
\mathbb{1}(u^{(1)} > 0, u^{(3)} < 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(3)} < 0, u^{(4)} > 0, u^{(12)} < 0, u^{(20)} < 0, u^{(21)} < 0, u^{(22)} < 0, u^{(23)} < 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} > 0, u^{(6)} > 0, u^{(8)} < 0, u^{(9)} < 0, u^{(24)} < 0) \\
\mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0, u^{(4)} < 0, u^{(5)} > 0, u^{(6)} > 0, u^{(8)} < 0, u^{(9)} < 0, u^{(24)} < 0)
\end{pmatrix}$$

$$\mathbf{h^2}(u) = \begin{pmatrix}
h_1(u) \cdot u^{(25)} \\
h_1(u) \cdot u^{(29)} \\
h_3(u) \cdot u^{(25)} \\
h_3(u) \cdot u^{(29)} \\
h_3(u) \cdot u^{(28)} \\
h_4(u) \cdot u^{(25)} \\
h_4(u) \cdot u^{(29)} \\
h_4(u) \cdot u^{(28)} \\
h_7(u) \cdot u^{(25)} \\
h_7(u) \cdot u^{(29)} \\
h_9(u) \cdot u^{(30)} \\
h_{13}(u) \cdot u^{(27)} \\
h_{13}(u) \cdot u^{(31)} \\
h_{14}(u) \cdot u^{(25)} \\
h_{14}(u) \cdot u^{(29)} \\
h_{14}(u) \cdot u^{(28)} \\
h_{14}(u) \cdot u^{(32)} \\
h_{14}(u) \cdot u^{(25)} \\
h_{14}(u) \cdot u^{(29)} \\
h_{14}(u) \cdot u^{(28)} \\
h_{14}(u) \cdot u^{(32)}
\end{pmatrix}$$

and $\mathbf{h^3}(u) = \mathbf{h^2}(u_i)^2$. $\mathbf{h^1}(u_{i,s})$ represent indicators for individual i's history in simulation s, $\mathbf{h^2}(u_{i,s})$ indicate the twelfth grade test scores for student i when the history permits the test score to be observed, and $\mathbf{h^3}(u_{i,s})$ are the squared test scores, again when the student's history in that simulation allows test scores to be observed.

For individuals who are not enrolled in STEM, we repeat the sampling process, albeit with a smaller number of equations. Define the sampled deviates as $u_i^{NS} = mean_i^{NS} + error_i^{NS}$, where $mean_i^{NS}$ and $error_i^{NS}$ are defined as follows:

$$mean_i^{NS} = \begin{pmatrix} \lambda_0^{10} + \lambda_2^{10}\overline{K}_{i,1}^{10,P} + \lambda_3^{10}H_{i,1-3}^{R} + \lambda_4^{10}I_{i,1} + \lambda_5^{10}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,10}\mu_i^m + \lambda_{NS}^{10} \\ \lambda_0^{11} + \lambda_2^{11}\overline{K}_{i,2}^{11,P} + \lambda_3^{11}H_{i,2-3}^{R} + \lambda_4^{11}I_{i,2} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,11}\mu_i^m + \lambda_{NS}^{11} \\ \lambda_0^{12,P} + \lambda_1^{12,P}\overline{K}_{i,3}^{12,P} + \lambda_2^{12,P}H_{i,3-3}^{R} + \lambda_3^{12,P}I_{i,2} + \lambda_4^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12}\mu_i^m \\ \lambda_0^{10} + \lambda_2^{10}\overline{K}_{i,2}^{10,PR} + \lambda_3^{10}H_{i,2-3}^{R} + \lambda_4^{10}I_{i,2} + \lambda_5^{10}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,10}\mu_i^m + \lambda_{NS}^{10} \\ \lambda_0^{11} + \lambda_2^{11}\overline{K}_{i,3}^{11,P} + \lambda_3^{11}H_{i,2-3}^{R} + \lambda_4^{11}I_{i,2} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,11}\mu_i^m + \lambda_{NS}^{11} \\ \lambda_0^{12,P} + \lambda_1^{12,P}\overline{K}_{i,4}^{12,P} + \lambda_2^{12,P}H_{i,4-3}^{R} + \lambda_3^{12,P}I_{i,4} + \lambda_4^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12}\mu_i^m \\ \delta_0^{12} + \delta_2^{12}\overline{K}_{i,3}^{12,P} + \delta_3^{12}L_{i,4}^{R} + \delta_4^{12}I_{i,3} + \delta_5^{12,P}X_i + \sum_{m=1}^{M}\alpha_{D}^{m,12}\mu_i^m \\ \lambda_0^{12,P} + \lambda_1^{12,P}\overline{\tilde{K}}_{i,4}^{12,PR} + \lambda_2^{12,P}H_{i,4-3}^{R} + \lambda_3^{12,P}I_{i,4} + \lambda_4^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12}\mu_i^m \\ \lambda_0^{11} + \lambda_2^{11}\overline{K}_{i,3}^{11,PR} + \lambda_3^{11}H_{i,3-3}^{R} + \lambda_4^{11}I_{i,3} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,11}\mu_i^m + \lambda_{NS}^{11} \\ \lambda_0^{12,P} + \lambda_2^{12P}\overline{\tilde{K}}_{i,4}^{12,P} + \lambda_3^{12,P}H_{i,4-3}^{R} + \lambda_4^{12,P}I_{i,4} + \lambda_5^{12,P}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,12P}\mu_i^m \\ \lambda_0^{11} + \lambda_2^{11}\overline{K}_{i,4}^{11,PR} + \lambda_3^{11}H_{i,4-3}^{R} + \lambda_4^{11}I_{i,4} + \lambda_5^{11}X_i + \sum_{m=1}^{M}\alpha_{RET}^{m,11}\mu_i^m + \lambda_{NS}^{11} \\ \overline{K}_{i,3}^{12,P} \\ \overline{K}_{i,4}^{12,P} \\ \overline{K}_{i,4}^{12,P} \\ \overline{K}_{i,4}^{12,PR} \end{pmatrix}$$

$$error_i^{NS} = \begin{pmatrix} \nu_{i,1} + \lambda_2^{10}\varepsilon_{i,1}^P \\ \nu_{i,2} + \lambda_2^{10}(\varepsilon_{i,2}^P + \gamma^P\varepsilon_{i,1}^P) \\ \nu_{i,3} + \lambda_2^{12,P}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + \gamma^{P^2}\varepsilon_{i,1}^P) \\ \nu_{i,2} + \lambda_2^{10}(\varepsilon_{i,2}^{PR} + \gamma^{PR}\varepsilon_{i,1}^P) \\ \nu_{i,3} + \lambda_2^{12,P}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,1}^P) \\ \nu_{i,4} + \lambda_1^{12,P}(\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^P + \gamma^{P^2}\varepsilon_{i,2}^{PR} + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P) \\ \eta_{i,4} + \delta_2^{12}(\varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + \gamma^{P^2}\varepsilon_{i,1}^P) \\ \nu_{i,4} + \lambda_1^{12,P}(\varepsilon_{i,4}^{PR} + \gamma^{PR}\varepsilon_{i,3}^P + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P) \\ \nu_{i,3} + \lambda_2^{11}(\varepsilon_{i,3}^{PR} + \gamma^{PR}\varepsilon_{i,2}^P + \gamma^P\gamma^{PR}\varepsilon_{i,1}^P) \\ \nu_{i,4} + \lambda_1^{12,P}(\varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P) \\ \nu_{i,4} + \lambda_1^{12,P}(\varepsilon_{i,4}^{PR} + \gamma^{PR}\varepsilon_{i,3}^P + \gamma^P\gamma^{PR}\varepsilon_{i,2}^{PR} + \gamma^P\gamma^{PR^2}\varepsilon_{i,1}^P) \\ \varepsilon_{i,3}^P + \gamma^P\varepsilon_{i,2}^P + \gamma^{P^2}\varepsilon_{i,1}^P \\ \varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^P + \gamma^{P^2}\varepsilon_{i,2}^{PR} + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P \\ \varepsilon_{i,4}^P + \gamma^P\varepsilon_{i,3}^{PR} + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P \\ \varepsilon_{i,4}^{PR} + \gamma^{PR}\varepsilon_{i,3}^P + \gamma^P\gamma^{PR}\varepsilon_{i,2}^P + \gamma^{P^2}\gamma^{PR}\varepsilon_{i,1}^P \end{pmatrix}$$

The function $F^{NS}(\cdot)$ that transforms the deviates for non-STEM students into simulated dependent variables is as follows:

$$F^{NS}(u_{i,s}) = \begin{pmatrix} \mathbf{h^{1,NS}}(u_{i,s}) \\ \mathbf{h^{2,NS}}(u_{i,s}) \\ \mathbf{h^{3,NS}}(u_{i,s}) \end{pmatrix} \tag{29}$$

where

$$\mathbf{h^{1,NS}}(u) = \begin{pmatrix} h_1^{NS}(u) \\ h_2^{NS}(u) \\ h_3^{NS}(u) \\ h_4^{NS}(u) \\ h_5^{NS}(u) \\ h_6^{NS}(u) \\ h_7^{NS}(u) \\ h_8^{NS}(u) \\ h_9^{NS}(u) \end{pmatrix} = \begin{pmatrix} \mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} < 0) \\ \mathbb{1}(u^{(1)} > 0, u^{(4)} < 0, u^{(5)} < 0, u^{(6)} < 0) \\ \mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} > 0, u^{(7)} > 0) \\ \mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} > 0, u^{(7)} < 0, u^{(8)} > 0) \\ \mathbb{1}(u^{(1)} < 0, u^{(2)} < 0, u^{(3)} > 0, u^{(7)} < 0, u^{(8)} < 0) \\ \mathbb{1}(u^{(1)} < 0, u^{(2)} > 0, u^{(9)} < 0, u^{(10)} < 0) \\ \mathbb{1}(u^{(1)} > 0, u^{(4)} > 0, u^{(5)} < 0, u^{(6)} < 0) \\ \mathbb{1}(u^{(1)} > 0, u^{(4)} < 0, u^{(5)} > 0, u^{(11)} < 0) \\ \mathbb{1}(u^{(1)} < 0, u^{(2)} > 0, u^{(9)} < 0, u^{(10)} > 0) \end{pmatrix}$$

$$\mathbf{h^{2,NS}}(u) = \begin{pmatrix} h_1^{NS}(u) \cdot u^{NS,(12)} \\ h_2^{NS}(u) \cdot u^{NS,(13)} \\ h_3^{NS}(u) \cdot u^{NS,(12)} \\ h_4^{NS}(u) \cdot u^{NS,(12)} \\ h_4^{NS}(u) \cdot u^{NS,(15)} \\ h_5^{NS}(u) \cdot u^{NS,(12)} \\ h_5^{NS}(u) \cdot u^{NS,(15)} \\ h_6^{NS}(u) \cdot u^{NS,(14)} \\ h_7^{NS}(u) \cdot u^{NS,(13)} \\ h_9^{NS}(u) \cdot u^{NS,(14)} \end{pmatrix}$$

and $\mathbf{h^{3,NS}}(u) = \mathbf{h^{2,NS}}(u)^2$. As before, $\mathbf{h^{1,NS}}(u)$ represent indicators for the history of (nonSTEM) student i in simulation s, while $\mathbf{h^{2,NS}}(u)$ represent the twelfth grade test scores for this student when they can be observed, and $\mathbf{h^{3,NS}}(u)$ are the square of these test scores, whenever it is possible to observe them.

We now provide a few examples to facilitate comprehension of this approach. Consider a student in the STEM track. We set an initial value for the parameter, θ^{init}, and let $g(u_{i,s} \mid z_i, y_{i,0}) := p(u_{i,s} \mid z_i, y_{i,0}, \theta^{init})$, where $p(u_{i,s} \mid z_i, y_{i,0}, \theta^{init})$ is a multivariate normal distribution with mean equal to $mean_i(\theta^{init})$ and covariance matrix given by $E[error_i(\theta^{init})error_i(\theta^{init})']$. For each individual i, S total random deviates are drawn from this distribution. We evaluate $f(u_{i,s})$ at every deviate. Suppose the first six components of $u_{i,s}$ are positive. $\mathbf{h^1}(u_{i,s})$ shows that individual i in simulation s ends up with the first history, 10-11-12X10-11-12. This individual will then have an observed math and Portuguese test score in the twelfth grade in the third year of high school. These scores are the 25th and 29th, respectively, elements of $u_{i,s}$ and these test scores form the first two elements of $\mathbf{h^2}(u_{i,s})$. Suppose, instead, that in simulation t the third and and seventh element of $u_{i,t}$ are positive. Then individual i in simulation t has history 2, $10 - 10dX10 - 10d$. This person will not have an observed test score.

The first twenty-four elements of $u_{i,s}$ uniquely characterize the fifteen STEM histories. In addition, there are 4 different realizations of twelfth grade test scores for each of math/Portuguese, depending on whether the student was retained in 10th, 11th, or 12th grade (if a student is retained in the 12th grade we observe two test scores: before she is retained, and after the year of retention).

Therefore, $u_{i,s}$ for a student enrolled in the STEM track is a vector of length $24 + 2 \times 4 = 32$.

Similarly, the first eleven elements of $u_{i,s}$ uniquely characterize the nine non-STEM histories. Because we observe only Portuguese scores for these histories, u for these students is a vector of length $11 + 1 \times 4 = 15$.

After pre-computing $f(u_{i,s})$ at every deviate, we optimize over the objective function by reweighting $f(u_{i,s})$ by $\frac{p(u_{i,s}|z_i,y_{i,0},\theta)}{g(u_{i,s}|z_i,y_{i,0})}$ as in equation (17) during optimization. We optimize $J(\theta)$ using the Broyden–Fletcher–Goldfarb–Shanno gradient-based algorithm. We supply a gradient that is approximated using finite differences with a step size of 10^{-8}. In order for the algorithm to work well, a good initial guess is crucial. For parameter values far from the truth, small numbers of individuals appear in the moments that we target, which causes the objective function to be relatively unresponsive to perturbations of the parameters. To get a good initial guess, we optimized repeatedly from different starting values using a simplex algorithm applied to a 10% sample of the data (to speed up computation)(Rowan (1990)). We then took each solution to the simplex algorithm and supplied it as an initial guess to the BFGS algorithm. With a good initial guess, the optimizer would converge quickly to the solution with approximately 1000 function evaluations. Many optimizations from different starting values were done to search for the global optimum. An initial optimization with the identity weight matrix is used to obtain solutions to compute the optimal weight matrix. The function is then minimized a second time with the optimal weight matrix.

Appendix C: Additional Tables (online)

Table C-1: Goodness of fit, STEM histories

History	Data	Simulation
10-11-12X10-11-12	50.3%	54.3%
10-10dX10-10d	8.77%	11.2%
10-11-12-12-12X10-11-12 or 10-11-12-12-12dX10-11-12	5.20%	1.00%
10-11-12-12X10-11-12	4.78%	2.93%
10-11-12dX10-11-12d	2.09%	1.21%
10-10-10dX10-10-10d	2.52%	5.82%
10-11-12-12dX10-11-12	3.07%	1.93%
10-11-11dX10-11-11d	2.12%	1.84%
10-10-11-12X10-10-11-12 or 10-10-11-12dX10-10-11-12	3.36%	5.53%
10-10-11-11dX10-10-11-11d	2.11%	0.60%
10-11-11-11dX10-11-11-11d	1.66%	0.20%
10-11dX10-11d	1.53%	1.61%
10-11-11-12X10-11-11-12	1.20%	1.68%
10-11-12-12X10-11-12-12	0.11%	0.16%
10-11-12-12-12dX10-11-12-12-12d or 10-11-12-12-12X10-11-12-12-12	0.25%	0.02%

The table shows the in-sample fit of the model for a subset of targeted moments: the proportion of STEM students with each history.

Table C-2: Goodness of fit, non-STEM histories

History	Data	Simulation
X10-11-12	75.8%	82.6%
X10-10-11-12	10.7%	8.33%
X10-11-12-12d	3.15%	1.84%
X10-11-12-12-12d	2.46%	0.20%
X10-11-12-12	2.08%	2.23%
X10-11-11-12	1.77%	2.48%
X10-10-11-12-12 or X10-10-11-12-12d	1.62%	0.50%
X10-10-11-11-12	0.66%	1.04%
X10-11-11-12-12 or X10-11-11-12-12d	0.54%	0.14%

The table shows the in-sample fit of the model for a subset of targeted moments: the proportion of non-STEM students with each history.

Table C-3: Retention impacts by socioeconomic status

	Math		*Portuguese*		
Low SES	Raw	S.D.	Raw	S.D.	Prop. with score
Retained in 10th/11th Grades	3.87	0.16	8.18	0.48	0.10
	(1.52)	(0.06)	(0.98)	(0.06)	(0.01)
Retained in 12th Grade	5.83	0.25	8.37	0.49	0.08
	(2.85)	(0.12)	(1.35)	(0.08)	(0.01)
Graduate in 4 Years	4.96	0.21	8.56	0.5	0.12
	(1.32)	(0.06)	(0.79)	(0.05)	(0.01)
High SES					
Retained in 10th/11th Grades	3.95	0.17	8.21	0.48	0.11
	(1.26)	(0.05)	(0.87)	(0.05)	(0.01)
Retained in 12th Grade	4.79	0.20	8.45	0.50	0.08
	(2.80)	(0.12)	(1.54)	(0.09)	(0.02)
Graduate in 4 Years	4.58	0.19	8.52	0.50	0.13
	(1.04)	(0.04)	(0.74)	(0.04)	(0.01)

The table presents estimates of the effect of grade retention on test scores for three treated subgroups disaggregated by socioeconomic status (SES). Low SES corresponds to an individual's parent qualifying for a public income subsidy. Treatment effects are computed by averaging the difference in test scores across the two counterfactuals, *for students whose scores are visible within four years of high school entry.* This does not include students who drop out before the 12th grade in either the status quo or counterfactual policy simulation. The standard deviations on the 12th grade math and Portuguese exams are 23.6 and 17.0 points, respectively.

Table C-4: Retention impacts by age relative to peers

	Math		*Portuguese*		
At Grade Level	Raw	S.D.	Raw	S.D.	Prop. with score
Retained in 10th/11th Grades	3.68	0.16	8.17	0.48	0.11
	(1.59)	(0.07)	(0.83)	(0.05)	(0.01)
Retained in 12th Grade	3.97	0.17	8.54	0.50	0.06
	(2.14)	(0.09)	(1.3)	(0.08)	(0.01)
Graduate in 4 Years	4.25	0.18	8.46	0.50	0.13
	(1.02)	(0.04)	(0.67)	(0.04)	(0.01)
One Year Older					
Retained in 10th/11th Grades	5.83	0.25	8.53	0.50	0.10
	(1.25)	(0.05)	(0.9)	(0.05)	(0.01)
Retained in 12th Grade	6.86	0.29	8.36	0.49	0.15
	(3.07)	(0.13)	(1.56)	(0.09)	(0.02)
Graduate in 4 Years	7.60	0.32	8.97	0.53	0.13
	(1.11)	(0.05)	(0.75)	(0.04)	(0.01)
Two or More Years Older					
Retained in 10th/11th Grades	7.34	0.31	7.70	0.45	0.07
	(2.23)	(0.09)	(1.03)	(0.06)	(0.01)
Retained in 12th Grade	8.87	0.38	7.48	0.44	0.17
	(1.74)	(0.07)	(1.40)	(0.08)	(0.01)
Graduate in 4 Years	9.88	0.42	8.49	0.50	0.07
	(1.58)	(0.07)	(0.93)	(0.05)	(0.00)

The table presents estimates of the effect of grade retention on test scores for three treated subgroups disaggregated by the age at which the student enters high school. Students who enter high school one or more year above grade level have typically been retained prior to entering high school. Treatment effects are computed by averaging the difference in test scores across the two counterfactuals, *for students whose scores are visible within four years of high school entry.* This does not include students who drop out before the 12th grade in either the status quo or counterfactual policy simulation. The standard deviations on the 12th grade math and Portuguese exams are 23.6 and 17.0 points, respectively.

Table C-5: Estimated parameters, log wage equation

	Dependent Variable:	
	Log wages	
	Coefficient	S.E.
Education	0.089	(0.000)
Experience	0.031	(0.000)
Experience^2	-0.0003	(0.000)
Constant	0.307	(0.002)
Observations	1,843,440	
RMSE	0.435	
R^2	0.313	

Appendix D: Identification of unobserved heterogeneity parameters (online)

The figures in this section demonstrate identification of the coefficients on the latent types, as well as the probability of each type. Figures D-1 through D-7 plot the SMM objective as a function of each coefficient for the latent types, holding every other parameter fixed. The center of each diagram is the estimated value that minimizes the objective function. The curvature of the objective function around the minimum shows that marginal perturbations in the type-specific parameters induce changes in the objective function. Note that the type probabilities are constrained to lie in the unit simplex, so a multinomial logit transformation was applied inside the function while optimizing, and the values along the X-axis in Figure D-7 corresponds to the values of the parameters before applying the transformation (they are not the estimated probabilities).

Figure D-1: Type Parameters, Math

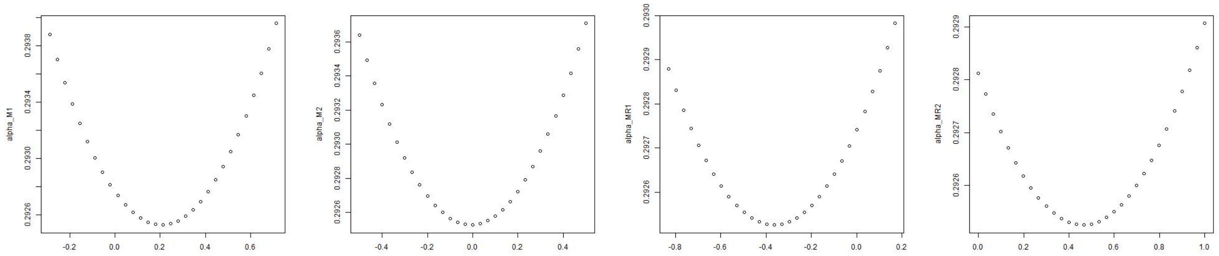


Figure D-2: Type Parameters, Portuguese

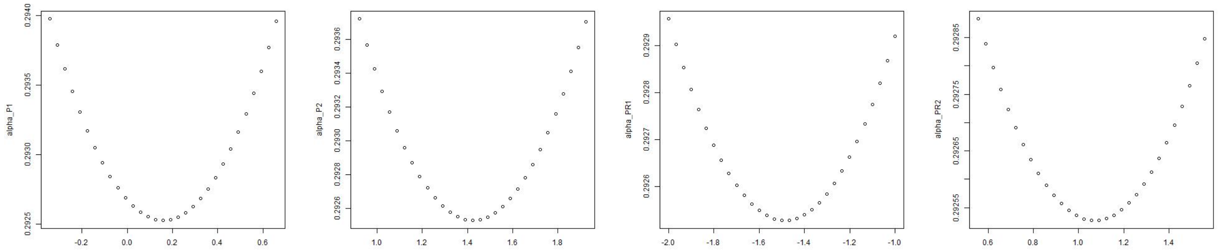


Figure D-3: Type Parameters, Dropout in 10th\11th Grades

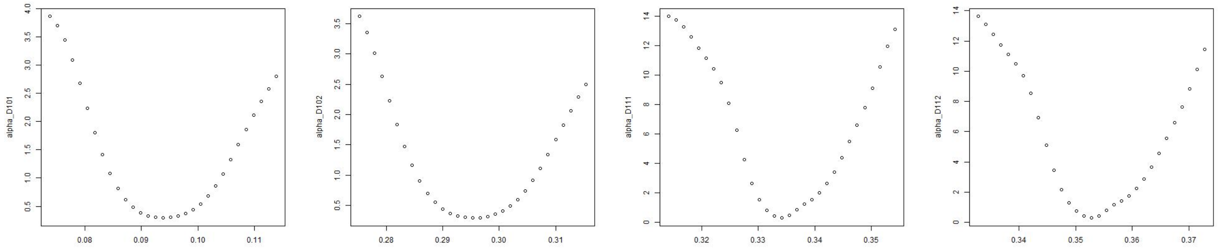


Figure D-4: Type Parameters, Dropout in 12th Grade

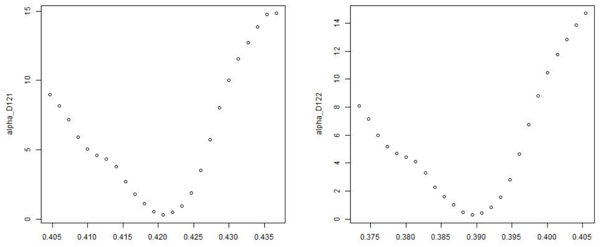


Figure D-5: Type Parameters, Retention in 10th\11th Grade

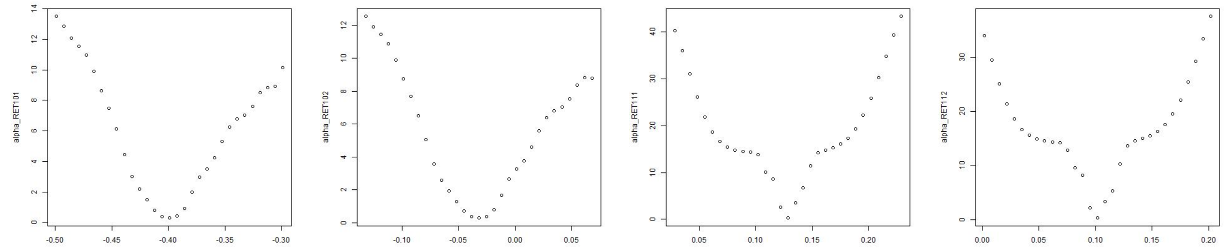


Figure D-6: Type Parameters, Retention in 12th Grade Math and Portuguese

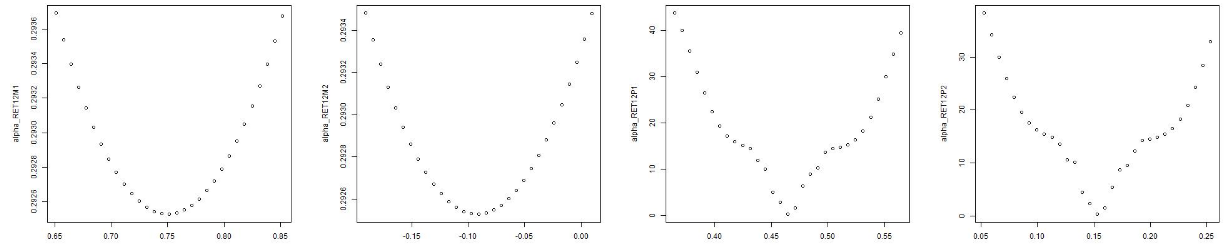
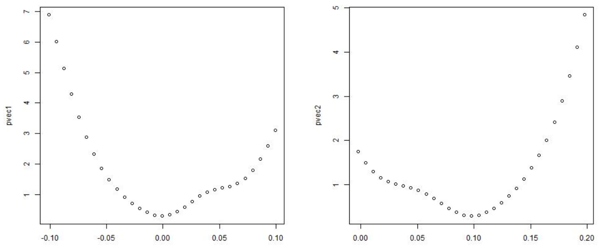


Figure D-7: Type Probabilities



Appendix E: Marginal Treatment Effects (online)

The marginal treatment effect (MTE) curve depicts how treatment effects vary with unobserved variables that determine selection into treatment. This section presents the results of MTE analysis of grade retention's effects on test scores and dropout. As in section 7.2, we use our estimated model to simulate a scenario in which all students are retained in grade g and a world in which no students are retained in any grade. We nonparametrically regress the difference in test scores across these two scenarios on the unobservabled variables influencing retention, which are a linear combination of the constant for the individual's unobserved type and their iid retention shock. We write these unobserved variables for grade $g = 10, 11$ as follows:

$$U^R_{i,g} = \sum_{m=1}^{M} \alpha^{m,g}_{RET} \mu^m_i + \nu^g_{i,t} \; .$$

In grade 12, we distinguish between unobservables affecting retention in math and those determining retention in Portuguese:

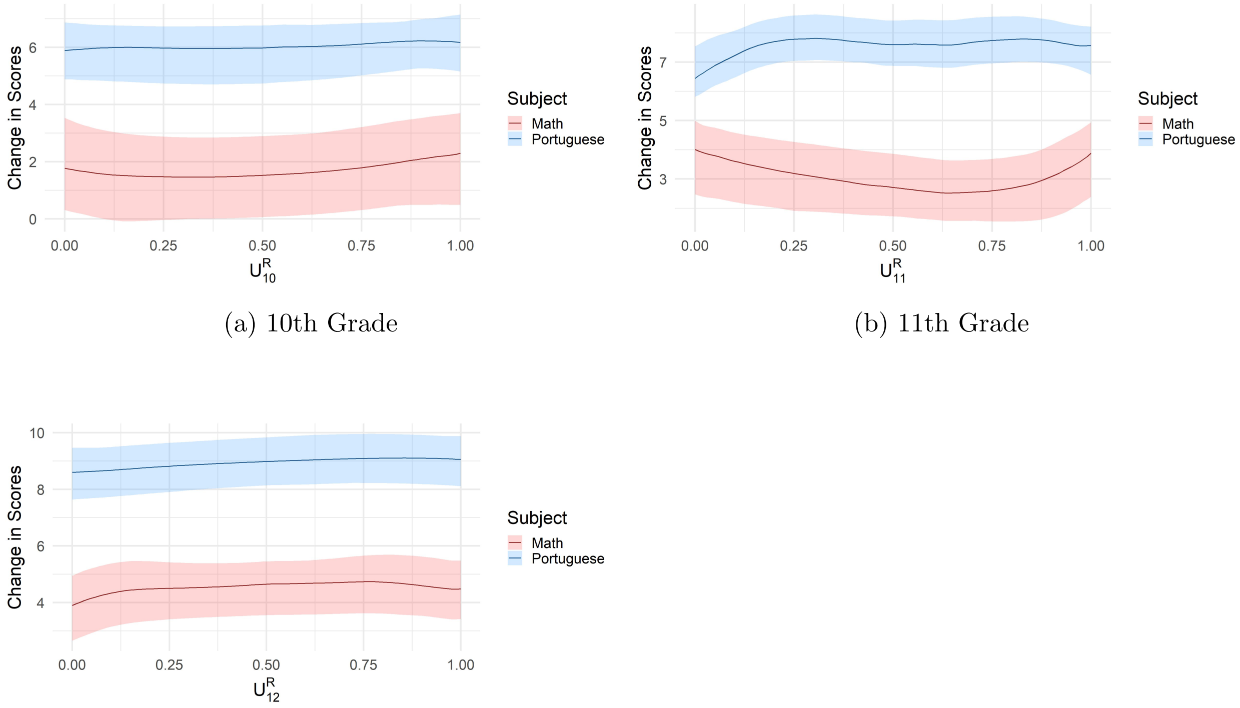
$$U^R_{i,12M} = \sum_{m=1}^{M} \alpha^{m,12M}_{RET} \mu^m_i + \nu^{12,M}_{i,t} \; ,$$
$$U^R_{i,12P} = \sum_{m=1}^{M} \alpha^{m,12P}_{RET} \mu^m_i + \nu^{12,P}_{i,t} \; .$$

In our analysis, a higher value of $U^R_{i,g}$ indicates that an individual is more likely to be retained in grade g, ceteris paribus. To facilitate comparisons between our analysis and the standard MTE setting, we nonparametrically regress the change in test scores on *quantiles* of $U^R_{i,g}$. Figures D-1 and D-2 displays the results of these nonparametric regressions.[36] Figure D-1 shows that the MTE curves for test scores in math and Portuguese are essentially flat in all grade levels. This is consistent with our estimation results that found that selection into retention was primarily on the basis on observed rather than unobserved factors. As the retention shocks, $\nu^g_{i,t}$, are independent of the value-added and dropout shocks, conditional on type, a nonconstant MTE function could only arise from the discrete multinomial types. However, Tables 5, 6, 7 showed that the type-specific

[36] We allow the bandwidth to differ for each plot. We select the bandwidth, following Fan and Gijbels (2018), to minimize the integrated mean square error across all evaluation points.

coefficients were typically small and statistically insignificant. The value-added MTE functions reflect this statistical insignificance, as do the dropout MTE functions, depicted in Figure D-2. They are flat, reflecting an absence of sorting on unobserved gains or losses. The primary sorting into retention is on the basis of observed characteristics, as demonstrated in section 7.2.

Figure D-1: Marginal treatment effects of retention on 12th grade test scores



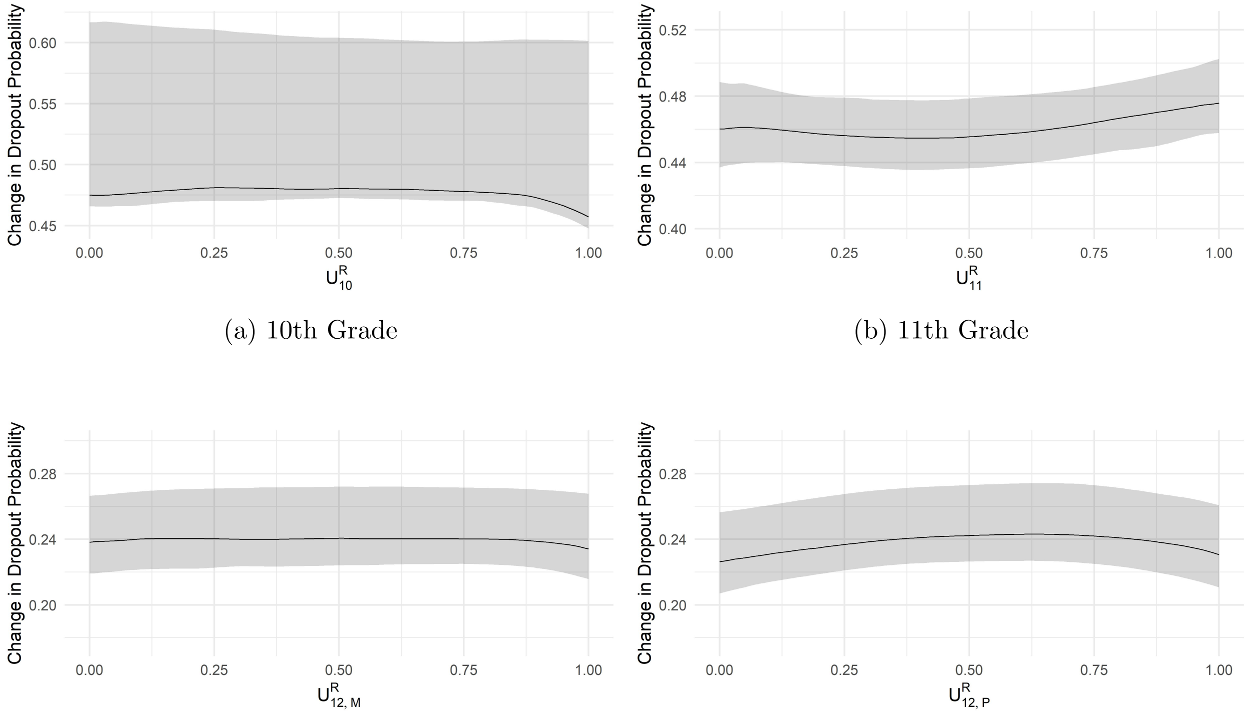
(a) 10th Grade

(b) 11th Grade

(c) 12th Grade

Note: Panels (a), (b), and (c) depict nonparametric regressions of the retention test score effect on quantiles of the unobservables determining retention in each grade and subject for the sample of students who take the 12th grade test. The regressions are estimated using local linear regression with an Epanechnikov kernel and a bandwidth designed to minimize the integrated mean square error. 95% confidence intervals are indicated by shaded regions. Confidence intervals are obtained by a parametric boostrap that samples 100 times from the asymptotic distribution of estimated parameters.

Figure D-2: Marginal treatment effects of retention on dropout



(a) 10th Grade

(b) 11th Grade

(c) 12th Grade, Math

(d) 12th Grade, Portuguese

Note: The figures depict nonparametric regressions of the dropout effect of retention on quantiles of the unobservables determining retention in each grade and subject. The regressions are estimated using local linear regression with an Epanechnikov kernel and a bandwidth designed to minimize the integrated mean square error. 95% confidence intervals are indicated by shaded regions. Confidence intervals are obtained by a parametric boostrap that samples 100 times from the asymptotic distribution of estimated parameters.