Quantitative Economics 7 (2016), 781–820 1759-7331/20160781

Grade retention and unobserved heterogeneity

Robert J. Gary-Bobo
CREST, ENSAE

Marion Goussé
Department of Economics, Université Laval

Jean-Marc Robin
Department of Economics, Sciences Po and University College London

We study the treatment effect of grade retention using a panel of French junior high-school students, taking unobserved heterogeneity and the endogeneity of grade repetitions into account. We specify a multistage model of human-capital accumulation with a finite number of types representing unobserved individual characteristics. Class-size and latent student-performance indices are assumed to follow finite mixtures of normal distributions. Grade retention may increase or decrease the student's knowledge capital in a type-dependent way. Our estimation results show that the average treatment effect on the treated (ATT) of grade retention on test scores is positive but small at the end of grade 9. Treatment effects are heterogeneous: we find that the ATT of grade retention is higher for the weakest students. We also show that class size is endogenous and tends to increase with unobserved student ability. The average treatment effect of grade retention is negative, again with the exception of the weakest group of students. Grade repetitions reduce the probability of access to grade 9 of all student types.

Keywords. Secondary education, grade retention, unobserved heterogeneity, finite mixtures of normal distributions, treatment effects, class-size effects.

JEL classification. C23, C36, C38, I2.

1. Introduction

Grade-retention practices are common in the schools of some countries but absent from others. Some educational systems have been designed to play the role of public certifica-

Robert J. Gary-Bobo: robert.gary-bobo@ensae.fr
Marion Goussé: marion.gousse@gmail.com
Jean-Marc Robin: jeanmarc.robin@sciencespo.fr

We thank Costas Meghir, Petra Todd, and three anonymous referees for useful comments. Robin gratefully acknowledges financial support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice, Grant RES-589-28-0001, and from the European Research Council (ERC), Grant ERC-2010-AdG-269693-WASP. Gary-Bobo's research is supported by Labex Ecodec (ANR-11-LABX-0047) and Investissements d'Avenir (ANR-11-IDEX-0003). Robin and Gary-Bobo gratefully acknowledge support from the Agence Nationale de la Recherche (ANR) white project "Econométrie des redoublements."

DOI: 10.3982/QE524

tion agencies. If this is the case, students are promoted to the next grade only if their test scores are sufficiently high, and the students who cannot pass are tracked or retained. France and Germany are good instances of such systems, in which grade retention is familiar. In contrast, social promotion, that is, the practice of passing students to the next grade, regardless of school performance, seems to prevail in more egalitarian societies or in countries promoting mass education. Scandinavian countries and the United Kingdom are good instances of the latter system. At the same time, grade retention is a form of second-best remedial education; in some countries it is the main if not the only form of remedial education, but it entails substantial costs. Grade repetitions consume resources, since they permanently increase the stock of enrolled students. There are opportunity costs, since grade repeaters could become productive sooner or have a longer productive life. There also exist substantial costs in the long run, since grade repeaters tend to obtain lower wages on the labor market, conditional on their highest credential.[1] Grade retention may also entail some benefits. The mere presence of grade repetitions acts as an incentive device and may increase study effort.[2] Finally, the distribution of skills in a given cohort of outgoing students may be improved if grade repeaters benefit from a longer period of schooling. Yet, many important aspects of a cost–benefit analysis are imperfectly known. As a consequence, in spite of its widespread use, it is hard to tell if grade retention dominates social promotion, or which of the two systems has the highest value as a social policy. As is well known, the question is hotly debated and international comparisons show trends in both directions. For instance, in recent years, France has relied less often on grade repetitions, while in the United States, grade retention has made a certain comeback, as an ingredient of school accountability policies.

The consequences of grade retention are not easy to estimate. This is essentially due to the endogenous character of the decision to hold a student back and to unobservable heterogeneity. Many studies in the past may have found a negative impact of grade retention on various outcomes because grade repeaters are a selected population with abilities below the average. In the sequel, we propose a way to evaluate the treatment effects of grade repetition in French junior high schools (grades 6–9), using a rich set of microdata, and taking the endogeneity of retention decisions and class size into account. We do not observe the students' wages and focus on educational outcomes.

In a preliminary study of the data, we find that the local average treatment effect (i.e., the LATE[3]) of grade retention on value-added, defined here as the difference between grade-9 and grade-6 scores, is significant and positive, using the quarter of birth as an instrument for retention. But the result does not seem to be very robust. We know that when treatment effects are heterogeneous, the linear instrumental variable (IV) estimator is a weighted average of marginal treatment effects (see the work of Heckman and Vytlacil (2005); see also Heckman (2010)). It follows that the IV estimates obtained with a particular instrument may not correctly identify the relevant effects. Indeed, in the following discussion, we show that the treatment effect of grade repetition varies with

[1] On this question, see Brodaty, Gary-Bobo, and Prieto (2012).

[2] On study effort, see De Fraja, Oliveira, and Zanchi (2010).

[3] On this concept, see Imbens and Angrist (1994).

unobserved characteristics of students, being positive for some individuals and negative for others.

Taking our inspiration from the work of Heckman and his co-authors, we propose a tractable model in which treatment effects are heterogeneous (see, e.g., Carneiro, Hansen, and Heckman (2003)). We assume the existence of a finite number of *latent student types* and that the effects of retention may vary from one type of individual to the next. Our approach is parametric: the observed outcomes and the latent variables, such as unobserved test scores, are modeled as finite mixtures of normal distributions. The model can then be used to compute counterfactuals and treatment effects.

We take dynamics into account, exploiting the data's panel structure. Our approach is similar in spirit to that of Cunha and Heckman (2007, 2008) and Cunha, Heckman, and Schennach (2010), but different (and somewhat simpler) in a number of technical details. The educational outcomes of the same individuals are observed recursively through time, either completely (quantitative test scores) or partially (qualitative promotion decisions). The successive observations are used to identify the model parameters and the latent student types. In particular, the coefficients of student types, that is, their impact on the different outcomes, are identified under a limited set of reasonable assumptions.

To be more precise, we specify a structural model of knowledge-capital accumulation in junior high school. The model explains grade retention, class size, promotion decisions, and test scores. It is estimated using panel data, on scores in grades 6 and 9, information on class sizes, and on student transitions (promotion to next grade, retention, and redirection toward vocational education). The panel provides a rich set of control variables describing family background and the environment of students. Repeated grades contribute to the accumulation (or destruction) of human capital (or skills) in a specific and type-dependent way. We present estimation results for a variant of our model with four unobserved student types or *groups*. Groups are clearly distinct and a clear hierarchy appears in terms of student ability. Groups are ranked in the same way if we use test scores in math or in French, or at the beginning of grade 6 or at the end of grade 9. The ranking of groups explains a similar ranking in the students' probabilities of grade retention (or promotion to the next grade). In a parallel fashion, the weaker the group, the smaller the class size, in every grade. This result shows the endogeneity of class size, which is used as a remediation instrument. Finally, to assess the impact of grade repetition on test scores at the end of grade 9, we compute the average treatment effect on the treated (ATT) and the average treatment effect (ATE) of the grade-repetition treatment. To this end, with the help of the model, we compute the counterfactual class size and test scores of grade repeaters (resp. nonrepeaters) that would be observed if they had not repeated a grade (resp. if they had repeated a grade), averaging over students and all possible types of each student, using their posterior probabilities of belonging to a group. We find that the ATE is negative, while the ATT is positive, but small and barely significant. The ATE and ATT are also computed within each of the four groups separately. This confirms that treatment effects are heterogeneous: grade retention is detrimental to able students but has some positive effects on the weakest students' final test scores. It is also shown that grade repetition has a negative impact on

the student's probabilities of access to grade 9. We conclude that grade retention should be replaced by some other form of remediation.

There is a substantial literature on grade retention, but many early contributions did not address endogeneity or selection problems in a convincing way (see, e.g., Holmes and Matthews (1984), Holmes (1989)). Few contributions have managed to propose a causal econometric evaluation. An early attempt, providing IV estimates on U.S. high-school data, is due to Eide and Showalter (2001). Also in the United States, Jacob and Lefgren (2004, 2009) use regression discontinuity methods to evaluate grade repetitions in the Chicago public-sector schools. Jacob and Lefgren (2004) find some positive short-term effects of grade retention on test scores for primary school children. Neal and Whitmore-Schanzenbach (2010) also propose an evaluation of the 1996 reforms that ended social promotion in Chicago public schools. Dong (2010) studies grade retention in kindergarten and finds positive effects. Closer to our approach, also using kindergarten data, Cooley-Fruehwirth, Navarro, and Takahashi (2011) estimated a multiperiod structural model with time-varying treatment effects. They find that the effect of grade retention depends on the timing of the treatment. Recently, Baert, Cockx, and Picchio (2013) used a structural dynamic choice model, estimated with Belgian data, and found that grade retention has a positive impact on the next evaluation, and persistent effects. On Latin American countries, see, for example, Gomes-Neto and Hanushek (1994). Manacorda (2012) applies a regression discontinuity approach to Uruguayan junior high-school data and finds negative effects on the dropout rate. In France, contributions on this topic (with a causal approach) are due to Mahjoub (2007), d'Haultfoeuille (2010), Brodaty, Gary-Bobo, and Prieto (2012, 2014), and Alet, Bonnal, and Favard (2013). Among these authors, d'Haultfoeuille (2010) applies a new nonparametric method for the estimation of treatment effects to French primary education data and also finds positive effects. Finally, Brodaty, Gary-Bobo, and Prieto (2012) find negative signaling effects of grade retention on wages. None of the quoted papers uses the methods and the data employed in the present article.

In the following discourse, Section 2 describes the data, Section 3 presents a preliminary analysis of grade retention using linear IV methods, and Section 4 presents our multistage skill accumulation model. The estimation strategy is exposed in Section 5. Sections 6 and 7 present the estimation results and the average treatment effects. Concluding remarks are given in Section 8.

2. Data

The data set used in this study is the 1995 secondary education panel of the French Ministry of Education (DEPP[4] Panel 1995), which follows 17,830 students in junior high-school (i.e., collège) from grade 6 to grade 9 (grade 6 is the equivalent of the French *classe de sixième*) during the years 1995–2001. The principals of a sample of junior high schools were asked to collect data on all pupils born on the 17th day of each month, with the exception of March, July, and October, and entering grade 6 in September 1995—about

[4]Département de l'Evaluation, de la Prospective et de la Performance.

1/40th of the whole cohort. A recruitment survey was conducted at the beginning of the first school year (1995–1996). Then a number of followup questionnaires were filled out by the principals in every subsequent year until 2001, and a questionnaire was filled out by the families in 1998 (with a response rate of 80%). Each student's junior high-school history was recorded without interruption, even when the student moved to another school. For each pupil and each year, we know the attended grade (6–9), the size of the class, and the promotion decision made by the teachers at the end of the year. In fact there are three possible decisions: promotion to next grade, grade retention, or redirection to vocational education (i.e., "steering"). These transition decisions are made during the last staff meeting (i.e., the *conseil de classe*), at the end of every school year, on the basis of test scores and other more or less objective assessments of the pupil's ability and potential in the next grade. Test scores in mathematics and French are available at the beginning of grade 6 and at the end of grade 9. Grade 9 test scores are missing for the individuals who dropped out of general education for apprenticeship or vocational training, and therefore never reached grade 9 in the general (nonvocational) middle schools. In addition, matching these data with another source from the Ministry of Education, the *Base Scolarité*, we obtain further information on school characteristics. In particular, total school enrollment and total grade enrollment (in each grade) for each year during the 1995–2001 period. These data will allow us to compute instruments for class size, based on local variations of enrollment. There are some missing data, but the quality of the panel is very good. For example, initial test scores are known for 95% of the sampled individuals. Discarding observations with obvious coding errors and missing data, and slightly more than 450 histories of pupils registered in special education programs (for mentally retarded children), we finally ended up with a sample of more than 13,000 individuals: 9403 of them are in grade 9 in 1999, 2594 are in grade 8, and 250 are in grade 7. The last subset contains the few individuals who repeated a grade twice. We chose to discard these observations to reduce the number of cases. The final sample has 13,136 students, which amounts to almost 75% of the individuals in the initial survey.

In the following discussion, grades are denoted by g, and $g \in \{1, 2, 3, 4\}$, where $g = 1$ corresponds to grade 6 and so on. The year is denoted t with $t \in \{1, 2, 3, 4, 5\}$, where $t = 1$ corresponds to year 1995 and so forth. Individuals are indexed by i. Let g_{it} denote the grade of individual i in year t. With this notation system, a student i who does not repeat any grade is such that $g_{it} = t$. A grade repeater is such that $g_{it} = t - 1$. Table 1 gives the observed distribution of grade histories (in junior high school). Each row corresponds to a different type of trajectory. The letter V stands for vocational education. For example, the sequence 11234 means that grade 6 was repeated and, therefore, that the student is observed in grade $g = 4$ in year $t = 5$. The sequence 123V indicates that the student was steered toward vocational education after grade 8. In total, about 30% of the pupils do not complete junior high school in 4 years: 18% are retained in one grade; 11% are redirected.

Individual histories are described by Table 2 and Figure 1. Table 2 presents two rows per year, except in year $t = 1$. During the first year, all students are in grade 6. Out of the 13,136 students initially enrolled in grade $g = 1$, 12,045 are promoted and 1091 are

TABLE 1. Individual grade histories.

Grade History	Count	
1234	9403	71.58%
12334	732	
12234	910	
11234	684	
Subtotal	2326	17.71%
1233V	33	
1223V	114	
1123V	154	
123V	147	
122V	146	
112V	246	
11V	7	
12V	560	
Subtotal	1407	10.71%
Total	13,136	

TABLE 2. Students promoted, retained, or redirected in each grade and year.

Year t	Grade	Initial Stock	Promoted (P)	Retained (R)	Redirected (V)
$t=1$	Grade 6	13,136	12,045	1091	0
$t=2$	Grade 6	1091	1084	0	7
	Grade 7	12,045	10,315	1170	560
$t=3$	Grade 7	2254	1862	0	392
	Grade 8	10,315	9403	765	147
$t=4$	Grade 8	2627	2326	0	301
	Grade 9	9403			

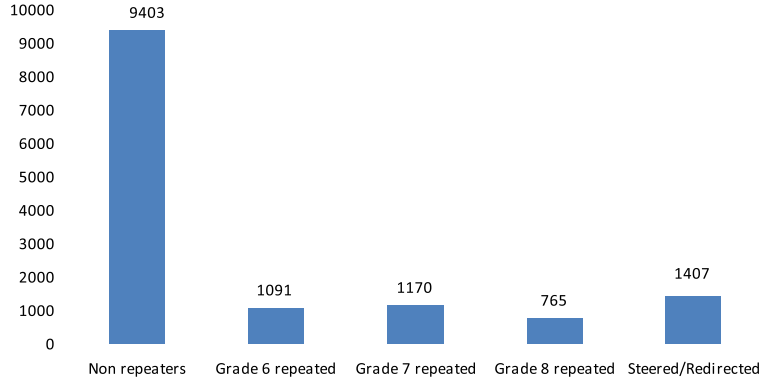

FIGURE 1. Number of repeaters in each grade.

retained. In year $t = 2$, we see that 1084 repeaters in grade $g = 1$ are promoted and only 7 students have been redirected. In year $t = 3$ there are $2254 = 1170 + 1084$ students in grade 7; 1170 students repeating grade 7 and 1084 students who were in grade 6 the year before. Figure 1 shows that the 9403 nonrepeaters constitute a majority of more than 70% of the students. Repeaters amount to less than 9% of the latter cohort each year.

3. Preliminary analysis: IV estimates

We start our study of the causal effect of grade retention on educational achievement, using the student's quarter of birth as an instrument for grade retention, in a linear model. The quarter or the month of birth has been used by various authors as an instrument (see, e.g., Angrist and Krueger (1991)). Recent work has shown that the month of birth can have long-lasting effects (see, e.g., Bedard and Dhuey (2006), Grenet (2010)). In his dissertation and a recent paper, Mahjoub (2007, 2009) used the quarter of birth as an instrument for grade retention. This approach yields a positive impact of grade retention on value-added scores. We follow the same approach here, as a preliminary step.

Value added (hereafter VA) is defined as the difference between standardized grade-9 and grade-6 scores, in mathematics and in French, respectively. This difference in test scores is higher for repeaters than for nonrepeaters. This is true in both French and mathematics. There exists a strong link between the age of a child, as measured by the month of birth or quarter of birth, and the probability of grade repetition (for details, see Appendix A). The probability of grade retention is clearly higher for children born later in the year. In principle, children must be 6 years old on September 1st of year t to be admitted in primary school, grade 1, year t. First-quarter students tend to be relatively older in their class, with an age difference that can reach 11 months, and relatively older children tend to perform better. At the same time, teachers are reluctant to retain older children in a grade, as retention may change a difference—being older—into a stigma—being too old.

It follows that the month, quarter, or season of birth is a candidate instrument for the grade-retention treatment, because it has good chances of being independent of the error term in an outcome equation with many controls. Note, in addition, as emphasized by Mahjoub (2007), that the value-added outcome being the difference of two test scores, possible specific and persistent effects of the birth quarter are "differenced out."

We now estimate the effect of grade retention on value-added by two-stage least squares (2SLS), using the quarter of birth as an instrument for grade retention. Some descriptive statistics on value-added, as well as further details on this IV approach, are relegated to Appendix A. Scores are standardized to have a mean of 50 and a standard deviation of 10 in grade 6 and in the whole sample (including all redirected pupils). Scores in grade 9 are standardized in the same way, using the subsample of individuals who reached grade 9. The first stage is a linear regression of the grade-retention dummy on birth quarter dummies and controls (the linear probability model). Results are displayed in Table 3. The fourth quarter being the reference in the regressions, we see that relatively older students have a significantly lower probability of being held back.

TABLE 3. Grade-retention probability.

Dependent Variable	Grade Retention
First quarter	−0.0513***
	(0.0110)
Second quarter	−0.0459***
	(0.00991)
Third quarter	−0.0133
	(0.0109)
R^2	0.054
F statistic for instruments	31.74

Note: Estimated by ordinary least squares (OLS). The dependent variable is the grade-retention dummy here. The control variables included in the regressions are gender, parental occupation, parental education, more than three children, indicator of grade repetition in primary school, and total school enrollment. Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

TABLE 4. The OLS and IV estimates of grade-retention effects.

	OLS		2SLS	
Dependent Variable	Math VA	French VA	Math VA	French VA
Grade repetition	1.757***	1.899***	21.94***	14.79***
	(0.200)	(0.196)	(5.391)	(4.510)
R^2	0.035	0.043		

Note: The table reports the estimated coefficient of the retention dummy in different regressions. VA, that is, value-added, the difference between test scores in grade 9 and grade 6, is the dependent variable. Gender is included as a control in all regressions in addition to parental occupation, parental education, more than three children in family indicator, indicator of grade repetition in primary school, and total school enrollment. Standard errors are given in parentheses. The asterisks ***, ** and * indicate significance at the 1, 5, and 10% levels, respectively.

Table 4 presents OLS and 2SLS estimates of the effect of grade retention on value-added scores using the same set of controls. Instrumenting grade retention by the quarter of birth has a dramatic impact: grade retention increases the score by about twice the standard deviation of value-added. These results confirm that the retention decision is endogenous.

Now, trying to estimate the impact of grade repetition in variants of this model, we found that the 2SLS results of Table 4 were not very robust. It is well known that IV estimates can be difficult to interpret when treatment effects vary with unobservable characteristics of individuals. To see this, we estimated several variants of a linear model with robust linear techniques. Table 5 shows the OLS and 2SLS estimates of a model in which standardized grade-9 scores are regressed on grade-6 entry test scores, the retention dummy, and controls, with the same first stage as in Table 4. In Appendix B (available in a supplementary file on the journal website, http://qeconomics.org/supp/524/supplement.pdf; also http://qeconomics.org/supp/524/code_and_data.zip), we present the three-stage least squares (3SLS) estimates obtained with an ex-

TABLE 5. The OLS and IV estimates of grade-retention effects.

	OLS		2SLS	
Dependent Variable	Math Grade 9	French Grade 9	Math Grade 9	French Grade 9
Grade repetition	−1.629***	−0.862***	19.93	−0.221
	(0.200)	(0.190)	(12.86)	(8.667)
Initial test score math	0.496***	0.237***	0.731***	0.244**
	(0.0109)	(0.0104)	(0.141)	(0.0951)
Initial test score French	0.108***	0.413***	0.314**	0.419***
	(0.0110)	(0.0105)	(0.124)	(0.0833)
R^2	0.418	0.472		

Note: The table reports the estimated coefficient of the retention dummy in different regressions. Gender is included as a control in all regressions in addition to parental occupation, parental education, more than three children in family indicator, indicator of grade repetition in primary school, and total school enrollment. There are $N = 11{,}694$ observations. Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

tended, simultaneous equations version of the model. This more elaborate version is used below as a point of comparison for our model with unobserved heterogeneity. Returning to Table 5, we immediately see that the sign of the OLS estimates of the grade-retention coefficient has changed. In addition, the corresponding IV estimates have lost their precision and significance.

There are several likely reasons for the nonrobustness of results. First, we do not know if the appropriate expression of value-added, say, in mathematics, is exactly $V_m = Y_{m1} - Y_{m0}$, that is, the difference between the final score Y_{m1} and the initial score Y_{m0}. The appropriate expression might well be $V_m = Y_{m1} - cY_{m0}$ with $c < 1$. Imposing $c = 1$, as in the model of Table 4, is too strong since Table 5 seems to indicate that $c \simeq 0.5$. It is also unclear that the chosen standardization is appropriate. This is why, in the model of Table 5, we treat initial scores as controls that may reduce the importance of endogeneity problems. But then, are error terms really independent from Y_{m0}? Is the IV strategy appropriate here and what are its shortcomings?

To see this, assume that the data are generated by a simple model in which the initial and final test scores, denoted Y_0 and Y_1, respectively, can be either the grades in math or in French, or an average of the two, for the sake of simplicity. Let θ be a random factor that represents unobserved "talent"; let R denote the grade-retention dummy and let Q be the semester-of-birth instrument (also a dummy), to simplify the exposition. We assume

$$
\begin{aligned}
Y_0 &= \theta + u,\\
R &= \alpha + \beta Q + \gamma\theta + v,\\
Y_1 &= a + (b + h\theta)R + c\theta + w,
\end{aligned}
$$

where $(a, b, c, h, \alpha, \beta, \gamma)$ are parameters, and (u, v, w, θ) are random variables with zero means and finite variances, assumed to be stochastically independent of each other. The first equation says that the entry test score is a measure of θ: it is just talent plus noise. The second equation is the auxiliary equation, that is, the first stage. The third equation

is the equation of interest, with a heterogeneous treatment effect of grade retention if $h \neq 0$.

Consider first the regression $Y_1 = a + bR + \epsilon$. Since $\epsilon = c\theta + hR\theta + w$, retention R is endogenous, and we can try the quarter of birth Q as an instrument. Assume that Q has the desirable properties $\mathbb{E}(\theta|Q) = 0$, $\mathbb{E}(u|Q) = \mathbb{E}(v|Q) = \mathbb{E}(w|Q) = 0$, and $\mathbb{E}(\theta^2|Q) = \sigma_\theta^2$ (i.e., talent has a constant variance). These assumptions are reasonable. If we now write the normal estimating equations for the IV estimates of (a, b) and take probability limits as the number of observations N goes to infinity, we find

$$\operatorname{Cov}(Y_1, Q) = b \operatorname{Cov}(R, Q) + \operatorname{Cov}(\epsilon, Q).$$

Given our assumptions, it is easy to check that $E(\epsilon) = hE(R\theta) = h\gamma\sigma_\theta^2$. We also have

$$\mathbb{E}(Q\epsilon) = \mathbb{E}\big[Q\mathbb{E}(\epsilon|Q)\big] = \mathbb{E}\big[Qh\mathbb{E}(R\theta|Q)\big],$$

but, again, it is easy to see that $\mathbb{E}(R\theta|Q) = \gamma\sigma_\theta^2$. We then find that $\mathbb{E}(Q\epsilon) = h\gamma\sigma_\theta^2\mathbb{E}(Q)$ and, therefore, $\operatorname{Cov}(Q, \epsilon) = \mathbb{E}(Q\epsilon) - \mathbb{E}(Q)\mathbb{E}(\epsilon) = 0$. From this we derive that

$$b = \frac{\operatorname{Cov}(Y_1, Q)}{\operatorname{Cov}(R, Q)}$$

and we conclude that $\hat{b}_{\mathrm{IV}}$, the IV estimator of b, is consistent. So the IV approach is justified, but we have learned nothing about the heterogeneity of treatment effects.

Next, if we now try to estimate the model

$$Y_1 = a + bR + cY_0 + \xi,$$

since $\xi = -cu + w + hR\theta$, the random disturbance ξ is correlated with R and Y_0. An IV approach is again needed. Note that if treatment effects were homogeneous, that is, if we had $h = 0$, the model could be estimated by OLS, since Y_0 would be an appropriate control. Assume now $h \neq 0$ and use Q as an instrument for R to estimate b and c. More precisely, we use (Q, Y_0) as a vector of instruments for (R, Y_0). Writing the normal equations and taking limits under standard assumptions about the instruments, we find the linear system

$$\operatorname{Cov}(Q, Y_1) = b \operatorname{Cov}(Q, R) + c \operatorname{Cov}(Q, Y_0) + \operatorname{Cov}(Q, \xi),$$
$$\operatorname{Cov}(Y_0, Y_1) = b \operatorname{Cov}(Y_0, R) + c \operatorname{Var}(Y_0) + \operatorname{Cov}(Y_0, \xi).$$

As before, it is easy to see that $\mathbb{E}(Q\xi) = \mathbb{E}(Q)\mathbb{E}(\xi) = \mathbb{E}(Q)h\gamma\sigma_\theta^2$. Hence, $\operatorname{Cov}(Q, \xi) = 0$. In addition, under our assumptions, we must have $\operatorname{Cov}(Q, Y_0) = \operatorname{Cov}(Q, u) + \operatorname{Cov}(Q, \theta) = 0$. It follows that $\hat{b}_{\mathrm{IV}}$ is consistent. But, on the other hand, $\operatorname{Cov}(Y_0, \xi) \neq 0$, implying that $\hat{c}_{\mathrm{IV}}$, the IV estimator of c, is biased.[5]

[5]To see this, using the fact that covariance is linear with respect to each of its arguments, we derive $\operatorname{Cov}(Y_0, \xi) = h \operatorname{Cov}(\theta, R\theta) - c\sigma_u^2$. Since $\mathbb{E}(\theta) = 0$, we easily find that

$$\operatorname{Cov}(\theta, R\theta) = \mathbb{E}\big(\theta^2 R\big) = \mathbb{E}\big[\theta^2(\alpha + \beta Q + \gamma\theta + v)\big] = \big(\alpha + \beta\mathbb{E}(Q)\big)\sigma_\theta^2 + \gamma\mathbb{E}\big(\theta^3\big).$$

Ideally, to solve this problem, we would need instruments for the entry test scores themselves. Yet, the IV strategy is a legitimate approach since it leads to consistent estimation of b. But this property holds only *if it is true that* $\mathrm{Cov}(Q, Y_0) = 0$. Given our data set, this seems to be false: if we look at Figure 6, in Appendix A, we see that Y_0 is a decreasing function of the quarter of birth. It follows that $\mathrm{Cov}(Q, Y_0) \neq 0$ and both IV estimators, $\hat{b}_{\mathrm{IV}}$ and $\hat{c}_{\mathrm{IV}}$, may be asymptotically biased. In practice, the arithmetic mean of Y_0 among individuals such that $Q = 1$ (i.e., students born during the second semester) is 50.41 in French and 50.39 in math, and the mean of Y_0 knowing $Q = 0$ (i.e., born during the first semester) is 51.8 in French and 51.62 in math, while the overall averages are, respectively, 51.20 and 51.09. The differences between these numbers are relatively small, so that $\mathrm{Cov}(Q, Y_0) \simeq -0.3$ in French (this is approximately one-half of the difference between the average of Y_0 knowing $Q = 1$ and the overall average). The correlation of Y_0 with Q is about -0.06, showing that this source of bias is small, in practice.

Finally, note in passing that if $c = 1$ (i.e., if value-added is the appropriate outcome), then an expression of the LATE estimator of b is the empirical counterpart of

$$b = \frac{\mathrm{Cov}(Q, Y_1 - Y_0)}{\mathrm{Cov}(Q, R)} = \frac{\mathbb{E}(Y_1 - Y_0 | Q = 1)}{\mathbb{P}(R = 1 | Q = 1) - \mathbb{P}(R = 1)}.$$

To sum up, the IV identification strategies based on the quarter of birth are legitimate methods and provide us with an estimation of the LATE, in principle, but (i) they do not allow for a study of heterogeneity in treatment effects, (ii) the quarter of birth poses problems as an instrument, in particular if it happens to be correlated with entry tests scores, leading to potential (but probably limited) biases, and (iii) in practice, the estimates of the treatment effect of grade retention obtained with the quarter of birth do not seem to be robust. Another distinct problem with the above IV estimates is that they are obtained with the subsample of the students who reached grade 9. This may obviously lead to a bias in the average treatment effect. The values that we find with the above IV strategy must be understood as conditional on the fact that students did not quit junior high school for vocational programs before grade 9. For these reasons, we propose to study the effects of grade retention with a different approach.

Following Cunha, Heckman, and Schennach (2010), we will model Y_1, Y_0, and R as explained by a common latent factor. Kotlarski's theorem (see, e.g., Kotlarski (1967)) says, in essence, that if $Y_0 = \theta + u$ and $Y_1 = \theta + v$, if we observe the distribution of (Y_1, Y_0), and if u, v, and θ are independent random variables, then the distribution of the latent factor θ is nonparametrically identified. In the absence of appropriate instruments, the latent factor's distribution can be identified with the help of several random measures, Y_0, Y_1, and R: the initial grades, the final grades, and the promotion or retention decisions. We obtain this in a relatively simple way at the cost of specifying structural equations. A consequence of this alternative approach is that the entry test scores Y_0 will be considered as endogenous, dependent variables, instead of possible control variables. Another important difference is that we no longer need the disappointing quarter-of-birth instrument (we may use it anyway, but this is not crucial). The key intuition here is that repeated noisy observations of a student's performance (i.e., the availability of several measures of the student's latent talent factor) are in a certain sense a substitute for the use of instruments in identification strategies.

4. A model of knowledge-capital accumulation

To uncover the mechanism of grade repetition and its impact on educational attainment, we construct a model of knowledge-capital accumulation with unobserved heterogeneity. We found a source of inspiration in a series of influential papers by Heckman and his co-authors, in which heterogeneity is captured by means of dynamic factor models. See, for example, Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010). Although close in spirit, the present approach relies on a somewhat simpler model. We use a multiperiod setting. We rely on the idea that in the educational process, inputs are imperfectly observed and outputs are imperfectly measured by means of test scores and teacher's decisions. Unobserved heterogeneity is modeled by means of a discrete set of unobserved individual types, generating finite mixtures of normal distributions.

The model is designed to match the following data features. We observe test scores in French and mathematics, but only at the beginning of grade 6 and at the end of grade 9. Promotion decisions (promotion to the next grade, grade retention, or redirection to vocational training) are observed in all years. In addition to these test scores and transitions, we also observe class size and total school enrollment. The students who do not drop off into vocational education at some point reach the terminal grade after 4 or 5 years, depending on retention, during the period 1995–2000. For children who never repeat a grade, we have observations in years $t = 1, 2, 3, 4$. For those who repeat a grade once and are not redirected to a vocational track, t can take all five values 1, 2, 3, 4, 5. Redirected children are the cause of attrition. Pupils are indexed by $i = 1, \ldots, N$. Let $g_{it} \in \{1, 2, 3, 4\}$ denote the grade of student i in year t, and let $S_{it} \in \{P, R, V\}$ denote the promotion decision (i.e., promotion, retention, and redirection) at the last staff meeting of year t. The term $g_{i,t+1}$ is missing if $S_{it} = V$. All students start in grade 6 in year 1 ($g_{i1} = 1$), so we set $S_{i0} = P$ for all i. There is no redirection of children toward vocational education in grade 6, so $S_{i1} \in \{P, R\}$.

4.1 *Initial conditions*

Initial scores in mathematics and French measure initial knowledge capital in mathematics and in French, denoted h_{m0} and h_{f0}, respectively. We assume that individuals have four possible unobservable types or, equivalently, belong to one of four possible *groups*. Let G_{ik} denote the dummy that is equal to 1 if i belongs to group k and equal to 0 otherwise. Let p_k denote the unconditional probability of belonging to group k and, of course, $p_1 + p_2 + p_3 + p_4 = 1$. Knowledge-capital levels at the beginning of grade 6, that is, h_{m0} and h_{f0}, have the form

$$h_{mi0} = c_{m01} + c_{m02}G_{i2} + c_{m03}G_{i3} + c_{m04}G_{i4}, \tag{1}$$

$$h_{fi0} = c_{f01} + c_{f02}G_{i2} + c_{f03}G_{i3} + c_{f04}G_{i4}. \tag{2}$$

In this formulation, group 1 is the reference group. It follows that c_{m01} and c_{f01} are the average initial levels of knowledge capital in mathematics, and French, respectively, for group 1 individuals. Subscript m (resp. f) indicates a coefficient related to the initial

mathematics capital (resp., the French language capital) equation. The average initial mathematics capital of group k is thus $c_{m01} + c_{m0k}$ for $k = 2, 3, 4, \ldots$.

Human capital is therefore discrete, but this should not be taken literally. We could add a random term with a continuous distribution, representing other unobserved inputs to the expressions of h_{mi0} and h_{fi0}, but the distribution of this term would not be identifiable, because it could not be distinguished from the teachers' "grading error," defined below. We suppose that the test scores in French, denoted y_f, and in math, denoted y_m, at the beginning of grade 6 are two different measures of the same knowledge capital, that is,

$$y_{mi} = h_{mi0} + \varepsilon_{mi0}, \tag{3}$$

$$y_{fi} = h_{fi0} + \varepsilon_{fi0}, \tag{4}$$

where ε_{m0} and ε_{f0} are random variables with a normal distribution and a zero mean, representing "grading" errors. The latter regression functions will identify the variance of ε_{m0} and ε_{f0}.

During the schooling of each student, we observe different variables that we regroup in different categories. There are time-invariant characteristics of the individual, such as family background observations, denoted X_0, time-varying characteristics of the individual, denoted X_t, $t = 1, \ldots, 5$, and time-varying characteristics of the school, used as instruments for class size, denoted Z_t. The variables used in regressions are listed in Table 6.

At this stage, we could also have added the list of controls X_0, including indicators of family-background characteristics, explaining the initial human capital h_{mi0} and h_{fi0}, in equations (1) and (2). By definition, the X_0 variables do not vary with time. To introduce a linear combination of the form $X_0 b_0$ in equations (1) and (2), we would have to assume that X_0 and the group indicators G_k are independent, which is a strong assumption. It follows from the adopted specification that the groups may capture some of the effects of family background. As a consequence, we will later use separate regressions to explain the impact of family-background variables and other controls on the probability of belonging to a given group. Another advantage of this formulation is to reduce the number

TABLE 6. Sets of variables.

Time-Invariant Characteristics X_0	Time-Varying Characteristics X_1, X_2, X_3, X_4, X_5	Time-Varying Instruments Z_1, Z_2, Z_3, Z_4, Z_5
Gender	Foreign language studied	Theoretical class size
Father's occupation	Special education zone	(i.e., Maimonides' rule)
Mother's education	Number of foreigners in school	
Number of siblings	Class size	
Grade retention in primary school	Total school enrollment	
Private sector in primary school	Size of the urban area	
	Private sector	

of parameters to be estimated. There is a small and finite number of groups k, while the number of possible family types may be very large. This is why the impact of family background would be modeled by means of a linear combination $X_0 b_0$, which can also be restrictive. Given this point of view, it may seem that there should be more than four groups. But such a parsimonious representation may, on the contrary, be appropriate, given the practical policy problem posed here, which is to classify students for the need of a pedagogical policy.[6] To assess the importance of these choices and their possible impact on our results, we present in Appendix C (in the supplementary file) a variant of our model in which some family background is added in the equations, and show that our main results are robust to these changes (see further comments below).

We treat class size as an endogenous variable and use an instrumental variable in the class-size equation (defined below), as in Angrist and Lavy (1999) and Hoxby (2000). The instrument for class size exploits discontinuities induced by the application of a class-opening threshold. Let N_{it} denote total grade enrollment in i's school in year t. The *theoretical class size* in year t, denoted Z_{it}, is the class size that would obtain if the headmaster's rule was to open a new class, as soon as total grade enrollment in grade g_{it} became greater than τq and to minimize class-size differences, where τ is the class-opening threshold and q is an integer. Given these definitions, the theoretical *number* of classes in grade g_{it}, denoted κ_{it}, is by definition

$$\kappa_{it} = \text{int}\left[\frac{N_{it}-1}{\tau}\right] + 1,$$

where int$[x]$ is the largest integer q such that $q \leq x$. The theoretical number of students per class in grade g_{it} is simply

$$Z_{it} = \frac{N_{it}}{\kappa_{it}}.$$

Piketty and Valdenaire (2006) and Gary-Bobo and Mahjoub (2013) show how this function of total grade enrollment fits the observed data in the French educational system. We set the threshold value $\tau = 25$ because it seems to provide the best fit with DEPP Panel 1995. We will see below that Z_{it} has a strong effect in class-size regressions.

4.2 *Knowledge-capital accumulation*

Knowledge, or human capital, accumulates according to the equation

$$h_{i1} = a_1 n_{i1} + b_1 X_{i1} + c_{11} + c_{12} G_{i2} + c_{13} G_{i3} + c_{14} G_{i4}, \tag{5}$$

where n_{i1} denotes class size in individual i's class, grade $g_{i1} = 1$. Again, in equation (5), group 1 is the reference, so that c_{11} is the impact of group 1 on h_{i1}, and the impact of group k is $c_{11} + c_{1k}$ for all $k > 1$.

Many studies have established that class size is an endogenous variable. In particular, available evidence for France shows that class size is positively correlated with student performance because smaller classes are typically used to redistribute resources in

[6] Eckstein and Wolpin (1999) used the same modeling strategy.

favor of weaker students or in favor of schools located in areas targeted for special help in education (see Piketty and Valdenaire (2006), Gary-Bobo and Mahjoub (2013)). We therefore model class size n_{i1} separately, as follows. Using group 1 as the reference, we have

$$n_{i1} = \alpha_{11} X_{i1} + \alpha_{12} Z_{i1} + \beta_{11} + \beta_{12} G_{i2} + \beta_{13} G_{i3} + \beta_{14} G_{i4} + \zeta_{i1}. \tag{6}$$

The random term ζ_{i1} is an independent, normally distributed error.

Since we do not have any quantitative measure of performance at the end of grades $g \in \{1, 2, 3\}$, repeated or not, we define a single, latent education score for those years. In grade 6 (i.e., if $g_{it} = 1$), we define the latent variable

$$y_{i1} = h_{i1} + \varepsilon_{i1}, \tag{7}$$

where ε_1 is an independent normal error with a zero mean.

An individual is promoted to grade 7 (i.e., $g_{i,2} = 2$) if his (her) human capital is high enough, and repeats a grade otherwise. The promotion decision is modeled as a simple Probit. Let C_{11} be a human-capital threshold above which students are promoted. We have

$$S_{i1} = \begin{cases} P & \text{if } y_{1i} \geq C_{11}, \\ R & \text{if } y_{1i} < C_{11}. \end{cases} \tag{8}$$

The distribution of ε_1 is assumed to be standard normal, as usual in such a case, to identify the coefficients of the latent index. Our specification of h_1 being given by (5) above, we see that the model will only identify the constant

$$\delta_{11} = C_{11} - c_{11}.$$

This is of course technically equivalent to normalizing C_{11}, but, in principle, C_{11} is the human-capital level above which students pass, while c_{11} is the specific mean level reached by group 1 students in the hypothetical situation $n_1 = X_1 = 0$. In essence, our model identifies differences between groups, not the absolute mean level of a group.

4.3 *From second to fifth year*

Similarly, still using group 1 as the reference, in the second and third years, the human capital has the following representation:

If $g_{it} = t$ (nonrepeaters), we have

$$h_{it} = a_t n_{it} + b_t X_{it} + c_{t1} + c_{t2} G_{i2} + c_{t3} G_{i3} + c_{t4} G_{i4}. \tag{9}$$

If $g_{it} < t$ (repeaters), we have

$$h_{it} = a_{tr} n_{it} + b_{tr} X_{it} + c_{t1r} + c_{t2r} G_{i2} + c_{t3r} G_{i3} + c_{t4r} G_{i4}. \tag{10}$$

The class-size equations are specified as follows:

If $g_{it} = t$ (nonrepeaters), we have

$$n_{it} = \alpha_{t1} X_{it} + \alpha_{t2} Z_{it} + \beta_{t1} + \beta_{t2} G_{i2} + \beta_{t3} G_{i3} + \beta_{t4} G_{i4} + \zeta_{it}, \tag{11}$$

where ζ_{it} is an independent normal random variable.

If $g_{it} < t$ (repeaters), we have

$$n_{it} = \alpha_{t1r} X_{ti} + \alpha_{t2r} Z_{it} + \beta_{t1r} + \beta_{t1r} G_{i2} + \beta_{t3r} G_{i3} + \beta_{t4r} G_{i4} + \zeta_{itr}, \tag{12}$$

where ζ_{itr} is an independent normal random variable. Therefore, the models for h_{it} (resp. n_{it}) have the same structure, but all the coefficients are free to vary with the student's status: repeater or nonrepeater.

At the end of the second and third years, if the student has not repeated a grade before, he or she can either pass to the next grade (P), repeat the year (R), or be redirected toward a vocational track (V). We model these three different transitions with an ordered Probit. Promotion or retention decisions are made by the teachers' staff meetings (i.e., the *conseils de classe*), at the end of every school year. In essence, these staff meetings base decisions on the student's grade-point average (hereafter GPA) at the end of the year, and decide whether to promote, to hold back, or to "steer" the student toward vocational education. Students with a GPA above a certain threshold are promoted; students with a low record are "steered"; students with a mediocre, below-the-average record repeat the grade if the teachers' committee thinks that they can benefit from the repetition. It seems reasonable to assume that the promotion decision is based on some average of the teachers' assessments of the student's cognitive capital plus an unobserved individual effect, reflecting other unobservable factors that the members of the teaching staff take into consideration. We have in mind that the student's unobservable GPA in year t is highly correlated with the latent capital h_{it}, or to fix ideas, that h_{it} is the GPA in year t plus some random factor. We then model the unobservable capital h_{it} as an educational output, which is the result of some educational inputs: class size, time-varying variables, and individual ability, as captured by the group indicator G_{ik}. Given this and given the clear hierarchy of the three possible decisions, it seems reasonable to use an ordered Probit structure.

Define first the latent variable

$$y_{it} = h_{it} + \varepsilon_{it},$$

where ε_t is an independent normal error. The decision S_{it} is then specified as

$$S_{it} = \begin{cases} V & \text{if } y_{it} < C_t, \\ R & \text{if } C_t \le y_{it} < D_t, \\ P & \text{if } y_{it} \ge D_t, \end{cases} \tag{13}$$

where C_t and D_t are the Probit cuts. We assume that ε_t has a standard normal distribution. As above, the model in fact identifies only the differences,

$$\delta_{t1} = C_t - c_{t1} \quad \text{and} \quad \delta_{t2} = D_t - c_{t1}.$$

In the sample, a student never repeats a grade twice. Thus, the model embodies the fact that if the student has already repeated a grade, he or she cannot repeat a second time. For repeaters, the possible decisions are promotion to the next grade or redirection. We model the two different transitions with a simple Probit. We first define the latent variable

$$y_{itr} = h_{it} + \varepsilon_{itr},$$

where ε_{tr} is an independent normal error. The decision S_{itr} is then specified as

$$S_{itr} = \begin{cases} P & \text{if } y_{itr} \geq C_{tr}, \\ V & \text{if } y_{itr} < C_{tr}, \end{cases} \tag{14}$$

where C_{tr} is a threshold, and we assume that ε_{tr} has a standard normal distribution. The model identifies only the difference, $\delta_{tr} = C_{tr} - c_{t1r}$.

It follows from these assumptions that the latent human capital h_{it} is affected by the promotion and retention decisions, because all the coefficients are free to vary in expressions (9) and (10), as well as in the auxiliary class-size equations (11) and (12), to describe a different productivity of inputs for students who repeated a grade.

The test scores in French, denoted y_{f4}, and in math, denoted y_{m4}, are two different measures of the final human capital. For nonrepeaters, with obvious notations for the random error terms, we have

$$y_{mi4} = h_{mi4} + \varepsilon_{mi4}, \tag{15}$$

$$y_{fi4} = h_{fi4} + \varepsilon_{fi4}, \tag{16}$$

where ε_{m4} and ε_{f4} are independent normal random variables. For repeaters, at the end of grade 9, test scores in French are observed in year $t = 5$ and denoted y_{f5}. Similarly, test scores in mathematics are denoted y_{m5}. We have two different measures of the repeaters' final human capital, with obvious notations for the independent random error terms:

$$y_{mi5} = h_{mi5} + \varepsilon_{mi5}, \tag{17}$$

$$y_{fi5} = h_{fi5} + \varepsilon_{fi5}. \tag{18}$$

The functions h_{mit} and h_{fit}, with $t = 4, 5$, have the same specification as h_{it} (as given by (9) above), with coefficients a_{mt}, b_{mt}, c_{mt} and a_{ft}, b_{ft}, c_{ft}, and so forth, that may be different for mathematics and French.

Our model is now fully specified. This model represents trajectories h_t that depend on the hidden type k, on time-varying covariates, and on the observable grade-repetition status. The value of h_{it} may depend on k in a different way at each period t through coefficients c_{tk}. There are no restrictions placed on the latter coefficients.

4.4 *Discussion*

The model presented above is flexible and quite general but has some limitations. First, we consider the issue of study-effort incentives. The very fact that grade repetitions have

a nonzero probability would act as a threat and the effort of students would on average be higher when the rate of grade repetition increases. This type of effect is very difficult to identify (see De Fraja, Oliveira, and Zanchi (2010)). In addition the effect could be weak. Indeed, incentives are provided by many other things: the parents, the labor market, and so forth. As noted by a referee, to estimate the effect, we would need data on schools where grade retention has been abolished. With our data, we did not find an important source of variability of the rate of grade repetition if we put aside the variability in the socioeconomic background of the students.

We ran regressions of the retention dummy, in each year t, on a long list of controls. A few variables have a significant impact, in addition to the obvious effect of the mother's education and the father's occupation. Students enrolled in schools with special subsidies (*zones d'éducation prioritaire* (ZEP) schools) are more (resp. less) likely to repeat a higher (resp. lower) grade; the size of the urban zone does not play a role; the rate of retention is not different in the private sector except in grade 9, where it is lower. We do take the predictable variations in grade-retention practices into account, since the variables that have an impact on the probability of retention in the linear probability model are used as controls is the promotion–retention Probits.[7]

Another interesting issue is school choice. Could it be that the model captures types that are matching types, reflecting a student-to-school matching, rather than latent student ability types? Latent types could indeed capture a typical matching effect, in addition to the student's unobservable ability, if we did not control for time-varying school and environmental characteristics. To a certain extent, we do control for the student's environment by means of X_t. Yet, it may be that our control variables miss an important aspect. This being said, our method is still valid if types also capture, to a certain extent, the fact that some categories of students are matched to certain kinds of schools, but the interpretation is more delicate, of course. If the proposed method is to be useful to policy makers, it should certainly be parsimonious in the sense that it would rely on a small number of relevant student groups. The study of "fixed effects" attached to school–student interactions would probably require many more groups and thus more parameters to estimate. This would go against the desire to summarize the most important effects with a parsimonious model, and such an attempt might fail because the data might not support the estimation of these parameters. But a study of the matching of students to schools could in principle become an interesting avenue of research—in the framework of the present paper, it is essentially out of reach, apart from controlling for observable school characteristics.

With the help of the model, we observe a given allocation of students to schools and we try to classify students rigorously. In practice, France is a country in which public schools are dominant, and private sector high schools are strictly regulated and highly subsidized by the government. In our sample, 18% of the students are enrolled in the private sector. In France, the private sector attracts only a slightly greater share of students from well-to-do families, but the social stratification is not extreme as in other

[7]Table S1, column (4), in Appendix B in the supplementary material, shows the estimated impact of family background on the probability of repeating a grade.

countries because tuition and fees are low. This is because free access is the legal counterpart of state subsidies for the bulk of French private schools. The educational system is therefore very homogeneous (or much less stratified) as compared to, say, the United States. For these reasons, it is very likely that our latent types are not reflecting school characteristics, since the private–public division would be the main source of such an effect, but the empirical basis is lacking. We could also consider different latent groups in French and mathematics, but that would go against parsimony, and we will see that groups explain the same ranking of scores in both disciplines.

Finally, we should pose the question of the external validity of our results. If samples with the same structure did exist in other countries, a variant of the model could be reestimated. But it would be asking too much to extrapolate on the basis of our results and to predict that grade retention has, say, a negative treatment effect of nearly the same magnitude in Germany. It seems reasonable to suggest, however, that qualitatively similar results would be obtained in Germany, like weak and/or heterogeneous treatment effects.

5. Identification and estimation method

The model can be viewed as a collection of finite normal mixture models. Under the normality assumption, these models are identified parametrically.[8] In addition, there are cross-equation restrictions since the latent groups k appearing in each equation are the same, with the same probability distribution. Without normality, the nonparametric identification of the distribution of groups k is a much more delicate question. A number of technical results can be proved; see Allman, Matias, and Rhodes (2009) and Kasahara and Shimotsu (2009). If we observe at least three conditionally independent random measures of an outcome, knowing the number of groups k, it is possible to identify a discrete mixture of probability distributions nonparametrically, provided that a number of technical conditions that bear on covariates hold true. We can certainly use the test scores in math and French in grade 6 as the first two measures. The final, grade 9 scores provide additional measures, but they are not independent conditional on the group. We may not be far from finding a few additional conditions under which a model such as ours would be identified without the normality assumption, but this difficult research question is beyond the scope of the present paper.

Intuitively, the proposed model performs an automatic classification of students based on a number of observations: test scores, promotion, and retention decisions. As the number of observations of the same student increases, an individual's posterior probability of belonging to a given group becomes closer to 0 or 1. To remove the risk of classification error completely we would need a large number of observations of the same student, related to the student's latent group (i.e., a long panel). The parametric normality assumption helps to identify a posterior distribution for the unknown group of each individual, knowing observed characteristics and observed outcomes, and thus helps to identify the individual's most likely group.

[8] See McLachlan and Peel (2000) and Geweke and Keane (1997).

The estimation method is a variation on the expectation-maximization (EM) algorithm. Let Y_i be the set of outcomes observed for individual i: $Y_i = (y_{mi0}, y_{fi0}, S_{i1}, \dots, S_{i4}, y_{mi4}, y_{fi4})$. Let $X = (X_1, X_2, X_3, X_4, X_5)$ and $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$. Then we denote θ the vector of all model parameters, namely, $\theta = (p_1, p_2, p_3, p_4, a_i, b_i, c_{ij}, \alpha_i, \dots)$. We replicate each individual i in the sample to create four different artificial observations of i. Student i's replicas differ by the unobserved type, or group k only, but the values of X_i, Y_i, and Z_i are the same for each replica. We arbitrarily choose initial values for the unconditional prior probabilities of the groups p_k, $k = 1, \dots, 4$, and for the posterior probabilities of belonging to a certain group knowing the observed characteristics of i, that is, $p_{ik} = \mathbb{P}(G_{ik} = 1|Y, X, Z)$. They will be updated after each iteration.

The estimation algorithm can be described as follows.

Step 1. We first run 20 *weighted* regressions and ordered Probits.

(a) Two regressions for the initial test scores in math and French.

(b) Two regressions of class size by grade: one for the repeaters and one for the nonrepeaters (except for the first year, because there are only nonrepeaters in year $t = 1$, and for year $t = 5$, because there are only repeaters). This amounts to eight regressions.

(c) One simple Probit to model the transition at the end of grade 6 in year $t = 1$. Two ordered Probits to model the decision at the end of grades 7 and 8 for nonrepeaters. Three simple Probits to model steering decisions relative to repeaters in grades 6, 7, and 8. There are four Probits and two ordered Probits in total.

(d) Two final test-score regressions in math and French, for repeaters and nonrepeaters (four regressions).

Step 2. We obtain an estimation of θ by means of our system of weighted regressions and weighted Probits.

Step 3. The residuals of regressions and the probabilities of passing to the next grade are collected to compute the individual contributions to likelihood, that is, by definition,

$$l_i(X, Z, Y, \theta) = \sum_{k=1}^{K} p_k l_i(Y, X, Z, \theta|G_{ik} = 1). \tag{19}$$

Step 4. Individual posterior probabilities p_{ik} of belonging to a group are then updated, using Bayes' rule and the likelihood as

$$p_{ik} = \mathbb{P}(G_{ik} = 1|Y, X, Z, \theta) = \frac{p_k l_i(Y, X, Z, \theta|G_{ik} = 1)}{\sum_{j=1}^{K} p_j l_i(Y, X, Z, \theta|G_{jk} = 1)}. \tag{20}$$

These individual probabilities are then averaged to update the prior probabilities p_k as

$$p_k = \mathbb{P}(G_k = 1) = \frac{1}{N} \sum_{i=1}^{N} p_{ik}. \tag{21}$$

Step 5. A new iteration begins until convergence of the estimated unconditional probabilities.

All standard deviations have been bootstrapped, using 50 drawings with replacement in the sample.

The estimation method used here has been advocated and justified by various authors (see, e.g., Arcidiacono and Jones (2003) and Bonhomme and Robin (2009)).

6. Estimation results

6.1 *Distribution of groups*

The results of the algorithm, using $K = 4$ groups, are given by Table 7. We chose to use only four groups because for $K > 4$ some groups become difficult to distinguish from each other. In Table 8, we compare the most likely groups of individuals, estimated with the full model, called classification 1, with the results of a limited submodel, based on grade-6 entry scores only, called classification 2. Both models have four unobserved types or groups. This has been done to try to assess the impact of initial test scores on the individual's posterior probabilities of belonging to a group. In other words, are students fully predetermined by their initial stock of knowledge? We observe that, according to classification 2, 75% of group 1 individuals are also most likely to become members of group 1, according to classification 1 (the full model). Observing the grade-6 scores in math and French only allows us to assign the student to the first group, to a large extent. But group 4 students are not predetermined by their entry test scores, since less than 2% of the students assigned to group 4 on the basis of the latter scores end up being members of group 4 in the full model. The corresponding percentages are 59% and 48%

Table 7. Estimated group probabilities.

	Group 1	Group 2	Group 3	Group 4
Probabilities	15.54%	31.16%	33.56%	19.74%
	(0.69)	(0.64)	(0.58)	(0.82)

Table 8. Comparison of two classifications.

	Classification 2				
Classification 1	Group 1	Group 2	Group 3	Group 4	Total
Group 1	74%	1%	0%	3%	2021
Group 2	24%	59%	2%	61%	4076
Group 3	0%	38%	48%	34%	4383
Group 4	2%	2%	50%	2%	2656
	100%	100%	100%	100%	
Total	2547	2883	4967	2739	13,136

for groups 2 and 3, respectively. We conclude that, with the exception of group 1, unobserved types are far from being perfectly predicted in year $t = 1$ (i.e., in grade 6). It seems that the weakest students are easily detected from the beginning, but the brightest students are not. We will come back to this point in the general discussion of estimation results below.

Table 9 presents the parameters obtained when we regress the individual posterior probabilities of belonging to a certain group k, defined as p_{ik} above, on the sociodemographic and family-background variables X_0. We find that the probabilities of belonging to the two extreme groups, group 1 and group 4, are significantly influenced by the social background. The results show, among other things, that when the mother is educated and the father is an executive, the probability of belonging to group 4 is significantly increased. Group 2 and group 3 are not so easy to distinguish on the basis of observed student characteristics. But the R^2 of these regressions—around 19% for group 1 and 14% for group 4—shows that the probabilities of belonging to a group are at best incompletely determined by observable family-background characteristics.

6.2 *Group effects on test scores*

We present here the estimated parameters of group effects and class size. Table 10 shows the estimated coefficients for the initial test scores (at the beginning of grade 6) and the final test scores (at the end of grade 9). Group 1 is the reference. We see how well the four groups are defined. Scores in French and math increase with group index k and the estimated coefficients yield the same ranking of ability groups in all columns, except the rightmost column of Table 10. More precisely, group 4 has everywhere the highest scores, with the exception of group 4 repeaters, in French, but the latter coefficient is estimated with less precision than the others. Intuitively, this is because group 4 students have a low probability of repeating a grade. Apart from this exception, group 4 is above group 3, which in turn dominates group 2, and group 1 is unambiguously the lowest ability group.

If we now focus on final scores, it is easy to see that group 1 gets higher scores on average when a grade was repeated (i.e., this is because the constant is higher). In contrast with group 1, individuals in groups 3 and 4 who did not repeat a grade obtain higher scores than the repeaters of these two groups. Take group 3 for instance. To obtain the final score in math of the average group 3 student who repeated a grade, we add the constant in the column (i.e., 43.31) to the differential impact of group 3 (i.e., 9.07). The total is 52.38. But if we compute the corresponding term for group 3 nonrepeaters, in math, we obtain, $15.80 + 41.88 = 57.68$. Grade repetition seems detrimental to group 3. The same is true with group 4. For the latter group, the corresponding additions yield 68.06 in the nonrepeaters' column and 59.36 in the repeaters' column. However, individuals in group 2 get approximately the same increase in their score, whether they repeat or not.

TABLE 9. Individual group probabilities and family background.

	Group 1	Group 2	Group 3	Group 4
Female	0.0493***	0.0388***	−0.00773	−0.0804***
	(0.00565)	(0.00766)	(0.00779)	(0.00638)
Mother education:	−0.0126	−0.0167	0.0234**	0.00594
Junior high school	(0.00873)	(0.0118)	(0.0120)	(0.00985)
Mother education:	−0.0521***	−0.0175	0.0531***	0.0165
Vocational certificate	(0.00937)	(0.0127)	(0.0129)	(0.0106)
Mother education:	−0.0901***	−0.103***	0.0550***	0.138***
High-school graduate	(0.0109)	(0.0147)	(0.0150)	(0.0123)
Mother education:	−0.0864***	−0.154***	0.0832***	0.157***
2 years of college	(0.0118)	(0.0160)	(0.0162)	(0.0133)
Mother education:	−0.103***	−0.174***	0.0240	0.253***
4 years of college and more	(0.0142)	(0.0192)	(0.0195)	(0.0160)
Father occupation:	−0.0514***	−0.0373	0.0100	0.0786***
Executive, professional	(0.0181)	(0.0245)	(0.0250)	(0.0204)
Father occupation:	−0.00141	0.0674***	−0.0314	−0.0346*
White collar	(0.0184)	(0.0249)	(0.0253)	(0.0207)
Father occupation:	0.0557***	0.0777***	−0.0696***	−0.0638***
Blue collar	(0.0169)	(0.0229)	(0.0233)	(0.0191)
More than three children in family	0.0845***	0.0004	−0.0432***	−0.0418***
	(0.00840)	(0.0114)	(0.0116)	(0.00949)
Retention in primary school	0.206***	0.0412***	−0.169***	−0.0781***
	(0.00731)	(0.00990)	(0.0101)	(0.00825)
Quarter of birth				
Q2	0.0000	0.0144	−0.0193*	0.0049
	(0.0077)	(0.0105)	(0.0107)	(0.0087)
Q3	0.0198**	0.0256**	−0.0331***	−0.0123
	(0.0085)	(0.0115)	(0.0117)	(0.0096)
Q4	0.0145*	0.0463***	−0.0260**	−0.0349***
	(0.0087)	(0.0117)	(0.0119)	(0.0098)
R^2	0.187	0.059	0.060	0.143

Note: Linear regressions of probabilities p_{ik} on controls X_0. Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively. There are 12,937 observations.

6.3 *Promotion decision model and effects of class size*

If we now look at the top rows in Table 10, we find that increasing class size has a negative impact in grade 9 for all students. The standard deviation of class size is around 3.[9] It follows that the estimated impact of a standard deviation of class size is around three-quarters of a normalized test-score point for nonrepeaters, or 7.5% of the standard deviation of test scores. The significant negative coefficient on class size appears because we control for unobserved heterogeneity and, therefore, for the endogeneity of this variable. Otherwise, the coefficient on class size would be positive (we return to this question below, when we discuss the class-size regressions). This being said, we do not find

[9]To be precise, the standard deviation of class size in year t, denoted σ_{nt}, has the values $\sigma_{n1} = 2.99$, $\sigma_{n2} = 2.90$, $\sigma_{n3} = 3.32$, and $\sigma_{n4} = 3.38$.

TABLE 10. Estimated impact of groups and class size on test scores.

	Score in Math			Score in French		
		Final			Final	
	Initial	Nonrepeaters	Repeaters	Initial	Nonrepeaters	Repeaters
Class size $t=4$		−0.25***			−0.25***	
		(0.03)			(0.04)	
Class size $t=5$			−0.19***			−0.25***
			(0.07)			(0.05)
Group 2	10.44***	8.14***	5.32***	10.82***	9.10***	5.80***
	(0.27)	(0.57)	(0.67)	(0.23)	(0.65)	(0.69)
Group 3	19.17***	15.80***	9.07***	19.16***	16.65***	10.13***
	(0.22)	(0.62)	(0.91)	(0.30)	(0.61)	(0.73)
Group 4	25.42***	26.18***	16.05***	25.60***	27.50***	9.22**
	(0.25)	(0.62)	(5.24)	(0.28)	(0.68)	(5.18)
Constant	35.34***	41.88***	43.31***	35.20***	40.87***	44.29***
	(0.24)	(0.92)	(1.71)	(0.26)	(0.94)	(1.20)
R^2	0.68	0.60	0.18	0.68	0.63	0.21

Note: Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

TABLE 11. Estimated impact of groups and class size on promotion decisions.

Dependent Variable ↓	Class Size	Group 2	Group 3	Group 4	Cut 1 δ_{t1}	Cut 2 δ_{t2}	Cut R δ_{tr}
S_1	−0.025***	0.67***	2.24***	2.45***	−1.13***		
	(0.007)	(0.04)	(0.12)	(0.72)	(0.16)		
S_2 repeaters	0.010**	4.29***	4.22***	3.17***			−1.80*
	(0.04)	(0.42)	(0.61)	(1.3)			(1.12)
S_2	−0.004	0.63***	1.62***	2.72***	−0.85***	−0.08	
	(0.006)	(0.044)	(0.057)	(0.64)	(0.14)	(0.14)	
S_3 repeaters	−0.016*	0.38***	0.92***	4.43***			−0.93***
	(0.012)	(0.017)	(0.24)	(1.37)			(0.29)
S_3	0.045***	0.33***	0.92***	1.67***	−0.64***	0.34**	
	(0.006)	(0.05)	(0.06)	(0.17)	(0.18)	(0.16)	
S_4 repeaters	0.035***	0.33***	0.65***	0.55			−0.002
	(0.01)	(0.07)	(0.12)	(1.91)			(0.24)

Note: The promotion decisions S_t are modeled with the help of an ordered Probit. They take the value 0 for redirection, 1 for retention, and 2 for pass. Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

a very strong class-size effect on final scores (a quarter of a point or 1/40th of the standard deviation of test scores for a one student reduction in class size). Table 11 shows the main parameters of the promotion decision model. Dependent variables determine rows, while the coefficients of a given explanatory variable in equations are displayed in the same column. A higher group label means a higher average knowledge capital. As

a consequence, the greater the group label, the greater the probability of passing to the next grade, for nonrepeaters as well as for repeaters, in each grade. The estimated coefficients reflect this ranking of groups very clearly, again, with the exception of the impact of group 4 in the Probit concerning grade-8 repeaters (i.e., S_4 repeaters). The latter coefficient is not estimated with precision because group 4 students have a small probability of repeating a grade. Apart from this exception, all other coefficients are estimated with good precision. The first column of Table 11 shows that increasing class size decreases the probability of promotion to grade 7, but has a nonsignificant (or even a positive impact) on pass rates in later grades.

6.4 *Endogeneity of class size*

Table 12 finally gives the coefficients of group dummies and of instruments in class-size equations. Each row in the table corresponds to a dependent variable. One of the class-size instruments is theoretical class size (i.e., Maimonides' rule), that is, the class size that would be experienced by the student if a class-opening threshold of 25 was applied, given total grade enrollment. The coefficient of this variable is significant and positive, as expected. We also find that class size increases with the ability (i.e., the group) of students. The only exceptions are the coefficients on group 4 dummies, that cannot be estimated with precision among grade repeaters. These results prove that class size is strongly endogenous and that it is used as a remediation instrument by school principals.

TABLE 12. Estimates of class-size equation parameters.

Dependent Variable ↓	Maimonides' Rule	Constant	Group 2	Group 3	Group 4	R^2
Class size $t=1$	0.32***	16.09***	1.12***	1.75***	1.78***	0.20
	(0.02)	(0.36)	(0.18)	(0.15)	(0.16)	
Class size $t=2$	0.49***	14.81***	0.68***	−5.75***	3.15*	0.25
(repeaters)	(0.05)	(1.09)	(0.27)	(1.51)	(2.35)	
Class size $t=2$	0.37***	15.17***	1.07***	1.85***	1.96***	0.21
	(0.02)	(0.17)	(0.30)	(0.14)	(0.16)	
Class size $t=3$	0.36***	16.06***	0.53***	1.13***	−1.96*	0.18
(repeaters)	(0.05)	(0.91)	(0.17)	(0.32)	(1.27)	
Class size $t=3$	0.35***	13.66***	1.87***	2.90***	3.10***	0.24
	(0.05)	(0.40)	(0.22)	(0.20)	(0.26)	
Class size $t=4$	0.33***	15.61***	0.95***	2.00***	1.85	0.19
(repeaters)	(0.05)	(0.90)	(0.22)	(0.29)	(2.16)	
Class size $t=4$	0.35***	14.05***	1.62***	2.62***	2.94***	0.26
	(0.02)	(0.46)	(0.34)	(0.26)	(0.29)	
Class size $t=5$	0.26***	16.34***	0.91***	2.67***	0.32	0.22
(repeaters)	(0.04)	(0.73)	(0.32)	(0.31)	(2.92)	

Note: Standard errors are given in parentheses. The asterisks ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

Our estimates are robust if the group dummies are exogenous variables in each year. To check this, we regressed the posterior probabilities of belonging to a group over a set of permanent individual characteristics X_0 and the time-varying characteristics X_1, X_2, X_3, X_4, X_5. The results of these latter regressions are not presented here, but they show that if the coefficients on X_0 are strongly significant, in contrast, time-varying characteristics are not significant. Thus, our model seems well specified (and we found a confirmation of well known results). A better social background (that is, more educated and more qualified parents) significantly increases the initial capital and, therefore, the probability of belonging to high-ability groups.

6.5 *Robustness*

Our first goal is to compare the results of the approach developed above with a standard approach, to make sure that our mode of treatment of unobserved heterogeneity is not yielding unreasonable results. To this end, using the same sample, we estimated a relatively easy to handle system of five simultaneous linear equations by means of three-stage least squares (3SLS). The system explains final test scores in math and French, grade retention, and class size in grades 6 and 9. As is well known, the consistency of 3SLS estimates does not depend on a normality assumption. The instrumental variables used are, as above, the semester of birth (to instrument grade retention), the theoretical class size, defined above, and total grade enrollment (to instrument class size). We add a long list of variables including controls for family background. The model and results are presented in Appendix B, available in the supplementary material. It is reassuring to find that this standard approach yields good results, and confirms a number of things that we learned with our more sophisticated model—if we put aside the treatment effects of grade retention themselves! We find reasonable effects of family-background variables (parental occupation, education, etc.) with the expected signs; we find significant and negative class-size effects; lower initial test scores are associated with a significantly higher probability of grade retention and a significantly lower class size, showing that class size is used as a remedial education tool in France. But the coefficients of the grade-retention dummy in the outcome equations are not particularly credible, as explained in Section 3. It is interesting to note that the impact of family-background variables is in essence the same in the 3SLS-estimated system and in the regression of probabilities of belonging to a latent group over the same controls, as shown by Table 9.

We then try a more ambitious robustness check, namely, to reestimate the whole model with unobserved heterogeneity, while adding controls in all equations. We added only a limited number of controls that we expect to be important: dummies indicating (i) if the mother has a college degree, (ii) if the father is an executive or professional, (iii) student gender, and (iv) if the student repeated a grade in primary school. In addition, we introduce quarter-of-birth indicators in all the Probit equations describing promotion to the next grade or retention. The results of the EM algorithm for this variant are presented and discussed in Appendix C, available in the supplementary material. In a nutshell, we find that if the weakest and the strongest groups (groups 1

and 4) are still well identified, groups 2 and 3 become difficult to distinguish. In Appendix C, we compare the groups in the two variants of the model. But it is reassuring to find that the main conclusions do not change. Computing the ATTs and ATEs with the variant, we reach, in essence, the same conclusions as in Section 7 below: ATT is small and barely significant; ATE is unambiguously negative; the treatment effect by groups shows that grade retention would have positive effects only in the weakest group.

7. The treatment effects of grade retention

We now turn to the key question of the present paper: the treatment effects of grade repetition. The model will be used to compute counterfactuals.

7.1 *Effect of grade retention on grade-9 scores*

Each individual i has a posterior conditional probability p_{ik} of belonging to each of the four groups $k = 1, \ldots, 4$. For each individual and each of his (her) possible types, we compute a counterfactual class size and a counterfactual final test score. Each individual has four counterfactual final scores and four counterfactual final class sizes. Using the posterior probabilities, we can then compute expected counterfactual grades.

For each group and for each student who has not repeated a grade, we perform the following operations:

(i) We compute the class size he or she would have experienced in grade 9 if he or she had repeated a grade. To do this, we assume that the student does not move to a different school and that his or her class environment has the same characteristics (same size of the urban area, same sector (private or public), same classification as priority education zone, ...). However, we use the information that we have on total school enrollment and total grade enrollment in the same school 1 year later.

(ii) We compute the score predicted in grade 9 if the student had repeated a grade (this counterfactual is denoted Y_r^c).

For each grade repeater and each group, we perform the following operations:

1. We compute the class size predicted in grade 9 if the student had not repeated a grade.

2. We compute the student's predicted score in grade 9 if he or she had not repeated a grade (this counterfactual is denoted Y^c).

Let N_r denote the number of individuals who repeated a grade and let N_p denote the number of individuals who did not repeat a grade. Of course, we have $N = N_p + N_r$. Let y_{ri} be the observed final grade of i if i is a repeater. Let y_i be the observed final grade of i if i never repeated a grade. We can now compute the following treatment effects.

The average treatment effect (i.e., ATE) is defined as

$$ATE = \frac{1}{N}\left(\sum_{i\in\mathbf{N}_p}\sum_{k=1}^{4}\big(\mathbb{E}\big(Y_{ri}^c|G_{ik}=1\big)-y_i\big)p_{ik} + \sum_{i\in\mathbf{N}_r}\sum_{k=1}^{4}\big(y_{ri}-\mathbb{E}\big(Y_i^c|G_{ik}=1\big)\big)p_{ik}\right), \tag{22}$$

where $p_{ik}=\mathbb{P}(G_{ki}=1|X,Z,Y)$ is i's posterior probability of belonging to group k. In the above expression, $\mathbb{E}(Y_{ri}^c|G_{ik}=1)$ and $\mathbb{E}(Y_i^c|G_{ik}=1)$ are the predictions of i's final grades in the counterfactual situations of grade repetition and not repeating, respectively, using the estimated regression functions and conditional on belonging to group k.

The average treatment effect on the treated (i.e., ATT) is then

$$ATT = \frac{1}{N_r}\sum_{k=1}^{4}\sum_{i\in\mathbf{N}_r}\big(y_{ri}-\mathbb{E}\big(Y_i^c|G_{ik}=1\big)\big)p_{ik}. \tag{23}$$

We also compute an ATE by group. For group k, the average treatment effect ATE_k is defined as

$$ATE_k = \frac{1}{Np_k}\left(\sum_{i\in\mathbf{N}_p}\big(\mathbb{E}\big(Y_{ri}^c|G_{ik}=1\big)-y_i\big)p_{ik}+\sum_{i\in\mathbf{N}_r}\big(y_{ri}-\mathbb{E}\big(Y_i^c|G_{ik}=1\big)\big)p_{ik}\right), \tag{24}$$

where $p_k=(1/N)\sum_i p_{ik}$. The ATT within group k, denoted ATT_k, can be defined in a similar way:

$$ATT_k = \frac{1}{\sum_{i\in\mathbf{N}_r}p_{ik}}\sum_{i\in\mathbf{N}_r}\big(y_{ri}-\mathbb{E}\big(Y_i^c|G_{ik}=1\big)\big)p_{ik}. \tag{25}$$

7.2 *Effect of grade retention on the probability of access to grade 9*

Individual i's estimated probability of access to grade 9, knowing group k, is denoted P_{9ik} and can be decomposed in the manner

$$\begin{aligned}
P_{9ik} =\ & \Pr(S_{i1}=P|k)\Pr(S_{i2}=P|k)\Pr(S_{i3}=P|k) \quad \text{(does not repeat)}\\
&+\Pr(S_{i1}=P|k)\Pr(S_{i2}=P|k)\Pr(S_{i3}=R|k)\Pr(S_{i4r}=P|k)\\
&\text{(repeats grade 8)}\\
&+\Pr(S_{i1}=P|k)\Pr(S_{i2}=R|k)\Pr(S_{i3r}=P|k)\Pr(S_{i4r}=P|k)\\
&\text{(repeats grade 7)}\\
&+\Pr(S_{i1}=R|k)\Pr(S_{i2r}=P|k)\Pr(S_{i3r}=P|k)\Pr(S_{i4r}=P|k)\\
&\text{(repeats grade 6)},
\end{aligned}$$

where, to simplify notation, we denote $\Pr(S_{it} = X|k) = \Pr(S_{it} = X|G_{ik} = 1)$ for all $X = P, R, V$. If the government decides to abolish grade retention (but keeps the possibility of steering students toward the vocational track), then the only way to reach grade 9 is to pass the three grades directly. Let P^c_{9ik} be the counterfactual probability of accessing grade 9 when grade retention is abolished. Given that no student is redirected to the vocational track at the end of grade 6, this probability can be expressed as

$$P^c_{9ik} = \Pr(S_{i2} = P|k)\Pr(S_{i3} = P|k).$$

To find the average treatment effect of grade retention, we need to compute the individual probabilities P_{9ik} and P^c_{9ik} for all the students in the sample, including those who have actually been redirected. This requires the computation of many counterfactuals. For those who repeated grade 6 and then passed or were redirected, we need counterfactual class sizes and counterfactual school-environment characteristics for years 2 and 3 that they would have experienced had they not repeated a grade. For those who repeated grade 5 or have been redirected at the end of grade 5, we need their counterfactual class size and counterfactual characteristics for year 3, as if they had not repeated this grade. Finally, for those who were never held back, we need the counterfactual class size and characteristics that they would have experienced had they repeated a grade. Table 13 summarizes the counterfactual probabilities and the counterfactual class size we computed for each different grade history. Then we can compute the following treatment effects. The average treatment effect is

$$ATE = \frac{1}{N}\sum_{k=1}^{4}\left(\sum_{i\in\mathbf{N}_r}(P_{9ik} - P^c_{9ik})p_{ik} + \sum_{i\in\mathbf{N}_p}(P_{9ik} - P^c_{9ik})p_{ik}\right).$$

TABLE 13. Counterfactuals required to compute the probabilities of accessing grade 9.

	Grade 7		Grade 6R		Grade 8		Grade 7R		Grade 8R	
History	$\Pr(S_2)$	n_2	$\Pr(S_{2r})$	n_{2r}	$\Pr(S_3)$	n_3	$\Pr(S_{3r})$	n_{3r}	$\Pr(S_{4r})$	n_{4r}
1234			C	C			C	C	C	C
12334			C	C			C	C		
12234			C	C	C	C				
11234	C	C			C	C				
1233V			C	C			C	C		
1223V			C	C	C	C				
1123V	C	C			C	C				
123V			C	C			C	C	C	C
122V			C	C	C	C			C	C
112V	C	C			C	C			C	C
12V			C	C	C	C	C	C	C	C
11V	C	C			C	C	C	C	C	C

Note: The letter C indicates that a counterfactual value has been computed. The letter R indicates that a grade-repeater model is used. The term $\Pr(S_t)$ means the probability distribution of decision $S_t \in \{P, V, R\}$. n_t denotes class size in year t. The subscript r indicates the specific model for grade repeaters, $S_{tr} \in \{P, V\}$.

The average treatment effect on the treated is then

$$ATT = \frac{1}{N_r}\sum_{k=1}^{4}\sum_{i\in \mathbf{N}_r}\left(P_{9ik} - P^c_{9ik}\right)p_{ik}.$$

7.3 *Results and discussion*

Table 14 displays the results of the various computations. The last row in this table shows the overall results. If we consider the final tests scores in math and French (at the end of grade 9), the ATT is positive, but small. Given that the mean value of the scores is 50 with a standard deviation of 10, the effects are smaller than a tenth of a standard deviation and barely significant. The ATE is clearly negative in math and in French. As we will see, this is mainly due to the fact that the most able students would suffer from grade repetitions. If we now look at the values of ATE_k and ATT_k, the treatment effects within group k, it is easy to see that only group 1 students benefit for grade repetitions. The effect of grade repetitions is not significantly different from zero for group 2 students. In contrast, in the case of group 3 and group 4, both the ATE and the ATT are negative in math and in French. This shows that grade repetition hurts the students belonging to top groups.[10] We conclude that grade repetitions have some usefulness for the weakest students, with an effect on the order of a quarter of a standard deviation on the final grades.

We now discuss the effect on the probability of access to grade 9. The treatment effects of grade repetition on final scores rely essentially on the regression equations that

TABLE 14. Average treatment effects of grade retention.

	Mathematics		French		Probability of Access to Grade 9	
	ATE	ATT	ATE	ATT	ATE	ATT
Group 1	2.43	2.45	3.09	3.20	−0.11	−0.11
	(0.76)	(0.76)	(0.81)	(0.80)	(0.014)	(0.014)
Group 2	0.12	0.36	0.18	0.47	−0.12	−0.12
	(0.42)	(0.42)	(0.41)	(0.42)	(0.012)	(0.012)
Group 3	−3.79	−2.92	−3.66	−2.77	−0.09	−0.10
	(0.76)	(0.75)	(0.60)	(0.59)	(0.022)	(0.023)
Group 4	−6.68	−14.08	−6.86	−14.22	−0.06	−0.06
	(4.61)	(4.53)	(4.54)	(4.52)	(0.07)	(0.07)
All	−2.56	0.27	−3.73	0.71	−0.09	−0.11
	(0.85)	(0.31)	(0.94)	(0.33)	(0.017)	(0.008)

Note: Standard deviations are given in parentheses.

[10]Note that ATT_k and ATE_k should be equal for each k if group k was the only variable used to predict counterfactual scores. But other control variables are used to predict these scores, such as class size and family background characteristics. This determines the differences between ATT_k and ATE_k in Table 14. However, the differences are neither large nor significant.

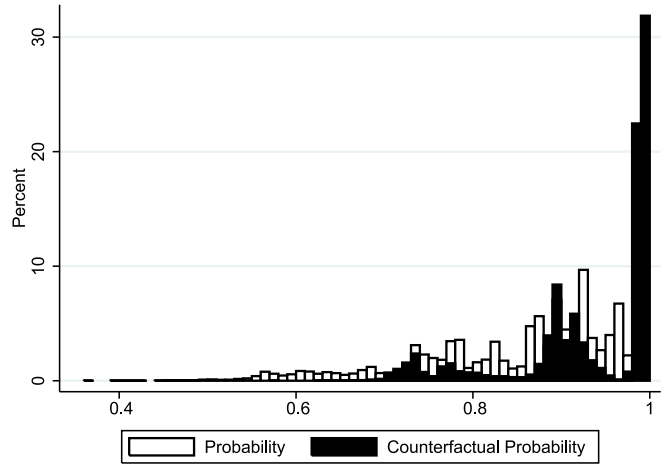

FIGURE 2. Histogram of individual probabilities of access to grade 9.

determine the final test scores, and the latter equations are estimated with the subset of individuals who reached grade 9. The fact that this population is selected is taken into account by the posterior individual probabilities p_{ik}. But it is reassuring to derive results for an outcome that depends on the entire structure of the model. This is the case of access to grade 9, because the probabilities P_{9ik}, defined above, depend on all the decision and class-size equations.

It is striking to see that in Table 14, the ATTs and ATEs of grade retention are all negative, even if we consider within-group treatment effects. This means that introducing grade retention, if grade retention does not already exists, will be detrimental to students, on average, and detrimental to students of each group, taken separately. The effects are particularly strong for groups 1 and 2. To see this, we computed the distribution of the individual probabilities P_{9ik} and individual counterfactual probabilities P^c_{9ik} in the student population. The histograms of these distributions are displayed on Figure 2.

On Figure 2 it is easy to see that the counterfactual probabilities have a mass near 1, meaning that the abolition of grade repetitions would help many students to reach grade 9. Yet, there are clearly subgroups of individuals that keep a low probability of access: these individuals bear a high risk of being tracked in vocational programs. We will understand the effect of grade repetition on access to grade 9 more fully if we compute the histograms of P_{9ik} and P^c_{9ik} separately for each group. This is done in the following figures. Figure 3 gives the distributions of P_{9ik}, while Figure 4 displays the distributions of the counterfactual P^c_{9ik}.

Comparing the histograms, it immediately comes to mind that when grade repetitions are abolished, access to grade 9 becomes certain for group 3 and group 4 students. The effect of abolition is less obvious for the weakest groups, 1 and 2, but in fact, these probabilities increase and become more favorable. To sum up, these effects explain why the treatment effects of grade repetitions on access to grade 9 are unambiguously negative. We see also that these effects are very strong, since a drop of 11 or 12 points of prob-

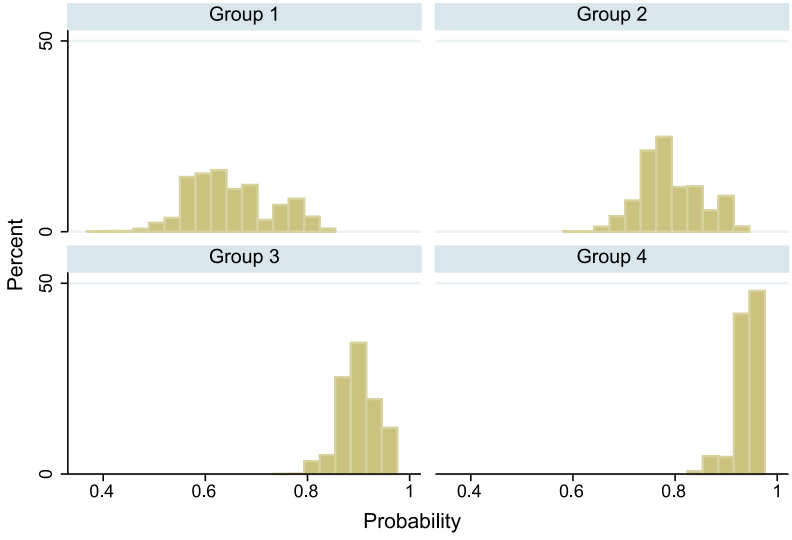

FIGURE 3. Histograms of probabilities of access to grade 9, by group.

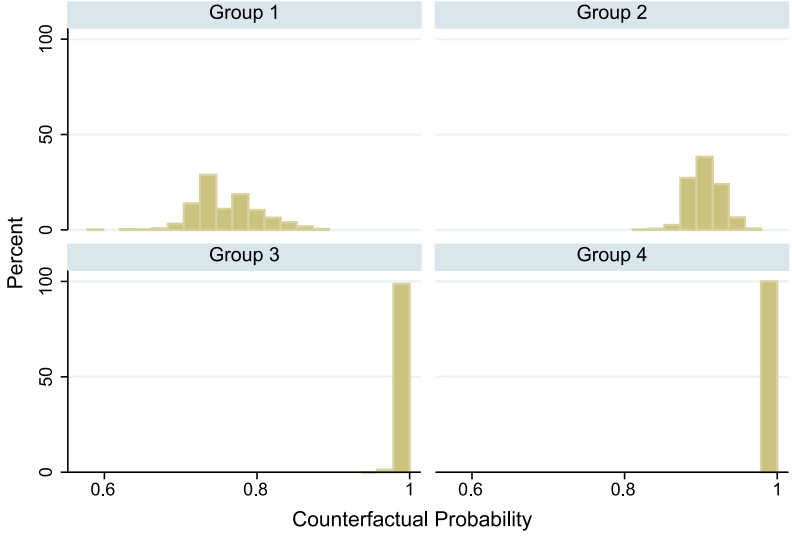

FIGURE 4. Histograms of counterfactual access probabilities, by group.

ability, very roughly, amounts to 50% of the best chances of access to grade 9 among group 1 and group 2 students.

The treatment effects are positive only for the weakest students, and these effects are weak when they are positive. Given these results, and the results of Table 14 in general, it seems that we can only recommend the abolition of grade retention. The results of Table 8 suggests a path for reform. Coming back to this table, we see that the weakest (i.e., the group 1) students are more easily detected in grade 6 than other types. In cases of grade retention, forcing weaker students to follow the same teaching twice is only a

rough second best. It would be more efficient to track these students from the start of junior high school, with additional remediation resources. One could imagine a slow track and a fast track, with, say, a year of difference in duration to reach the certification exams at the end of grade 9, and with flexible possibilities of track changes in both directions. To avoid the stigma of tracking, the slow track should probably be the norm, and students who seem promising would be steered toward the fast track. A system of that sort would lead to a more efficient use of resources than grade repetitions. It would clearly give weak students better chances of reaching the end of grade 9 with the required stock of knowledge and skills.

8. Conclusion

Grade retention is difficult to evaluate because grade repeaters have been selected on the basis of many characteristics that the econometrician does not observe. The difficult problem is to find a reasonable model to compute what would be the counterfactual performance of a student who has repeated a grade, if instead of being held back, he or she had been promoted to the next grade. To this end, we have assumed that the distribution of student test scores can be represented by a finite mixture of normal distributions, conditional on observed covariates, during each year of the observation period. The class size experienced by a student is also assumed to be distributed as a mixture of normals. All such mixtures are relying on the same finite number of latent student classes, called *groups*. In a flexible formulation, we show that class size, probabilities of grade retention, and test scores all depend on the unobserved group in a nontrivial and consistent way. We estimated a model with four groups and found that the four groups are unambiguously ranked. The higher the group index, the greater the student's ability and the larger his/her class size. This proves that class size is endogenous, smaller classes being used by school principals to redistribute resources toward weaker students. With the help of our model, we computed counterfactual test scores to evaluate the average treatment effect and the average treatment effect *on the treated* of grade retention. We found that the ATE is negative, while the ATT is generally positive, but small. We computed treatment effects in each student group separately, and found that the ATE is positive for less able students and negative for more able students. Finally we computed the ATT and ATE of grade retention on the probability of access to grade 9, and found that this effect is significant and negative. Grade retention is a form of remedial education and seems to help the weakest students insofar as it tends to increase their test scores at the end of grade 9. But these effects are weak. It follows that grade retention could probably be replaced by a form of tracking or by different forms of remediation. Other studies have shown that grade retention is a stigma and that repeated years are interpreted as a negative signal by employers (on this point, see Brodaty, Gary-Bobo, and Prieto (2012)). The long-run effects of grade retention seem to be detrimental. We can only conclude that grade retention is unlikely to be an efficient public policy, because its impact on student performance—when positive—is weak.

Appendix A: Details on quarter of birth as an instrument for grade retention

Table 15 gives summary statistics relative to the main variables of our sample. Table 16 displays descriptive statistics on value-added. Scores in grade 6, ranging between 0 and 20, as is usual in French schools, are standardized to have a mean of 50 and a standard deviation of 10 in the whole sample in grade 6 (including all redirected pupils). Scores

Table 15. Summary statistics.

Variable	Mean	Std. Dev.
Female	0.501	0.5
Age at grade-6 entry	11.158	0.492
Retention in primary school	0.204	0.403
Education of mother		
No education	0.176	0.381
Junior high-school certificate (i.e., BEPC)	0.312	0.463
Secondary vocational certificate (i.e., CAP, BEP)	0.222	0.415
High-school degree (i.e., Baccalauréat)	0.126	0.332
Associate's degree	0.101	0.302
Bachelor and more	0.064	0.244
Father's occupation		
Farmer	0.032	0.176
Self-employed, owner of a business	0.099	0.298
Executive, professional, higher education	0.163	0.369
Intermediate profession, technician, middle manager	0.18	0.384
White collar employee	0.109	0.312
Blue collar worker	0.366	0.482
Unemployed	0.051	0.22
Class size 1	25.567	2.993
Total grade enrollment	150.95	57.573
Total school enrollment	564.557	214.467
Rural area	0.097	0.296
Less than 5000 inhabitants	0.107	0.309
5000–9999 inhabitants	0.099	0.299
10,000–19,999 inhabitants	0.077	0.267
20,000–49,999 inhabitants	0.104	0.306
50,000–99,999 inhabitants	0.082	0.274
100,000–199,999 inhabitants	0.081	0.272
200,000–1,999,999 inhabitants	0.212	0.409
Paris	0.141	0.348
Special subsidies (ZEP school)	0.105	0.307
Number of observations	13,036	

Note: The BEPC (brevet d'études du premier cycle) is the French junior high-school diploma; CAP (Certificat d'aptitude professionnelle) and BEP (brevet d'études professionelles) are vocational degrees. The high-school degree is the baccalauréat. Class size 1 is the number of students per class in year $t = 1$; statistics for years 2–5 are very similar. ZEP schools (zones d'éducation prioritaire) receive special subsidies and more teachers per pupil.

TABLE 16. Descriptive statistics for value-added.

	Math		French	
Standardized Score	Balanced Sample[a]	Repeaters	Balanced Sample[a]	Repeaters
Grade 6	51.10	43.25	51.21	43.38
	(9.55)	(8.48)	(9.47)	(8.44)
Grade 9	50	43.37	50	43.46
	(10)	(8.23)	(10)	(7.87)
VA = grade 9 − grade 6	−1.10	0.11	−1.21	0.08
	(8.55)	(9.63)	(8.39)	(9.18)

Note: [a]Sample of all pupils for whom a test score is available both in grade 6 and in grade 9.

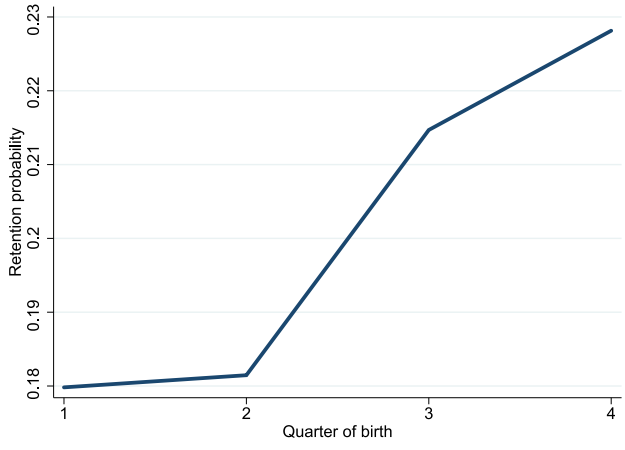

FIGURE 5. Probability of grade retention by quarter of birth.

in grade 9 are standardized in the same way in the sample of individuals who reached grade 9. Table 16 shows that value-added, the sign of which is irrelevant because scores are measures of performance relative to each grade, is nevertheless higher for repeaters than for nonrepeaters. This is true both in French and mathematics.

There exists a strong link between the age of a child, as measured by the month of birth or quarter of birth, and the probability of grade repetition. A look at Figure 5 shows the frequency of grade retention by quarter of birth. The probability of grade retention is clearly higher for children born later in the year. In principle, children must be 6 years old on September 1st of year t to be admitted in primary school, grade 1, year t. In practice, many 5-year-old children born between October and December are admitted, but the 5-year-old children born in the first quarter typically have to wait until the next year. It follows that first-quarter students tend to be relatively older in their class, with an age difference that can reach 11 months. Older children being more mature, they tend to perform better, and at the same time, teachers are reluctant to retain them because they are older, everything else being equal.

Figure 6 shows that initial (grade-6 entry) scores decrease with quarter of birth. The decreasing trend also exists for final scores but is less pronounced. Figure 7 shows that

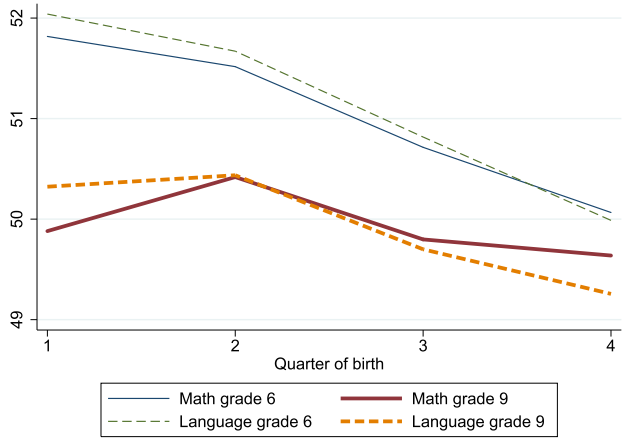

FIGURE 6. Scores by quarter of birth.

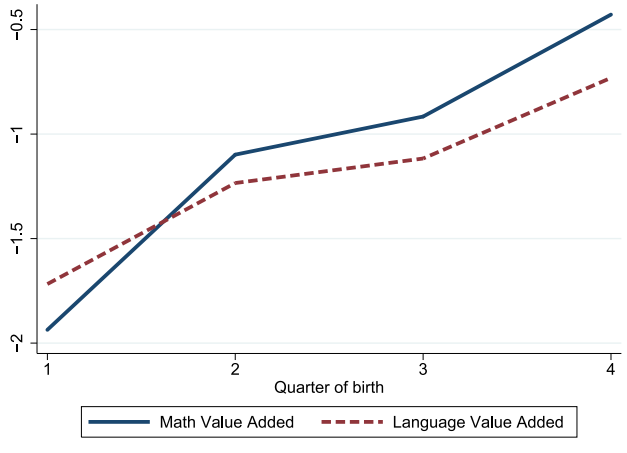

FIGURE 7. Value-added by quarter of birth.

value-added scores tend to be higher for relatively younger students, who seem to be catching up during their junior high-school years. In a first attempt to check if this is attributable to grade retention, we plot value-added by quarter of birth separately for repeaters and nonrepeaters. Figure 8 clearly shows that value-added age profiles are steeper for repeaters than for nonrepeaters. To understand the kind of effect captured by Figure 8, suppose that the underlying model has the following structure, as discussed above. Let $V = Y_1 - Y_0$ denote value-added, where Y_1 is the final test score and Y_0 is the entry test score. Assume that we have the two equations

$$V = a + bR + c\theta + \eta, \tag{26}$$

$$R = \alpha + \beta Q + \gamma\theta + v, \tag{27}$$

where the retention dummy is R, the semester of birth is Q, the variables η, v, and θ are normal, independent random shocks with variances σ_η^2, σ_v^2, and σ_θ^2, and a, b, c, α, β,

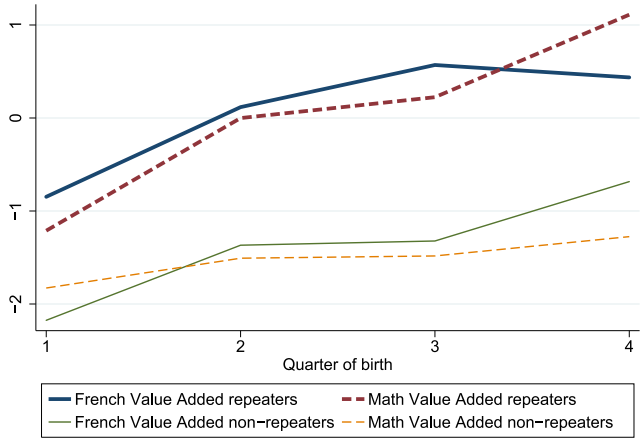

FIGURE 8. Value-added by quarter of birth for repeaters and nonrepeaters.

and γ are parameters.[11] The expectation of value-added, conditional on (R, Q) can be expressed as

$$\begin{aligned}\mathbb{E}(V|R,Q) &= a + bR + \mathbb{E}(c\theta + \eta|R,Q)\\ &= a + bR + \mathbb{E}(c\theta + \eta|\gamma\theta + v)\\ &= a + bR + \delta(\gamma\theta + v)\\ &= a + bR + \delta(R - \alpha - \beta Q),\end{aligned}$$

where

$$\delta = \frac{\mathrm{Cov}(c\theta + \eta, \gamma\theta + v)}{\mathrm{Var}(\gamma\theta + v)} = \frac{\gamma c\sigma_\theta^2}{\gamma^2\sigma_\theta^2 + \sigma_v^2}.$$

Now clearly, we have

$$\mathbb{E}(V|R,Q) = a + (b + \delta)R - \beta\delta Q - \alpha\delta.$$

It follows that, on Figure 8, the gap between repeaters ($R = 1$) and nonrepeaters ($R = 0$), knowing Q, is $b + \delta$. The LATE of grade repetitions is the IV estimator of b, that is, b_{IV}, given by Table 4, and the fact that the OLS estimator of b (also given by Table 4) is smaller than b_{IV} implies that $\delta < 0$ and hence $c\gamma < 0$, $\delta\beta < 0$, a very natural result.

REFERENCES

Alet, E., L. Bonnal, and P. Favard (2013), "Repetition: Medicine for a short-run remission." *Annals of Economics and Statistics*, 111, 227–250. [784]

[11] We use semester instead of quarter of birth for the sake of simplicity. We also dropped the controls from the equations to simplify the discussion.

Allman, E. S., C. Matias, and J. A. Rhodes (2009), "Identifiability of latent structure models with many observed variables." *The Annals of Statistics*, 37, 3099–3132. [799]

Angrist, J. D. and A. B. Krueger (1991), "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 106 (4), 979–1014. [787]

Angrist, J. D. and V. Lavy (1999), "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics*, 114, 533–575. [794]

Arcidiacono, P. and J. B. Jones (2003), "Finite mixture distributions, sequential likelihood, and the EM algorithm." *Econometrica*, 71 (3), 933–946. [801]

Baert, S., B. Cockx, and M. Picchio (2013), "On track mobility, grade retention and secondary school completion." Unpublished manuscript, Ghent University, Belgium. [784]

Bedard, K. and E. Dhuey (2006), "The persistence of early childhood maturity. International evidence of long-run age effects." *Quarterly Journal of Economics*, 121 (4), 1437–1472. [787]

Bonhomme, S. and J.-M. Robin (2009), "Assessing the equalizing force of mobility using short panels: France, 1990–2000." *Review of Economic Studies*, 76 (1), 63–92. [801]

Brodaty, T., R. J. Gary-Bobo, and A. Prieto (2012), "Does speed signal ability? The impact of grade retention on wages." Unpublished manuscript, CREST-ENSAE, France. [782, 784, 813]

Brodaty, T., R. J. Gary-Bobo, and A. Prieto (2014), "Do risk aversion and wages explain educational choices?" *Journal of Public Economics*, 117, 125–148. [784]

Carneiro, P., T. K. Hansen, and J. J. Heckman (2003), "Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice." *International Economic Review*, 44 (2), 361–422. [783]

Cooley-Fruehwirth, J., S. Navarro, and Y. Takahashi (2011), "How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatments effects." *Journal of Labor Economics* (forthcoming). [784]

Cunha, F. and J. J. Heckman (2007), "The Technology of Skill Formation." Technical report, National Bureau of Economic Research, Cambridge, Massachusetts. [783]

Cunha, F. and J. J. Heckman (2008), "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of Human Resources*, 43 (4), 738–782. [783, 792]

Cunha, F., J. J. Heckman, and S. M. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78 (3), 883–931. [783, 791, 792]

De Fraja, G., T. Oliveira, and L. Zanchi (2010), "Must try harder: Evaluating the role of effort in educational attainment." *Review of Economics and Statistics*, 92 (2), 577–597. [782, 798]

d'Haultfoeuille, X. (2010), "A new instrumental method for dealing with endogenous selection." *Journal of Econometrics*, 154 (1), 1–15. [784]

Dong, Y. (2010), "Kept back to get ahead? Kindergarten retention and academic performance." *European Economic Review*, 54 (2), 219–236. [784]

Eckstein, Z. and K. I. Wolpin (1999), "Why youths drop out of high school: The impact of preferences, opportunities, and abilities." *Econometrica*, 67 (6), 1295–1339. [794]

Eide, E. R. and M. H. Showalter (2001), "The effect of grade retention on educational and labor market outcomes." *Economics of Education Review*, 20 (6), 563–576. [784]

Gary-Bobo, R. J. and M.-B. Mahjoub (2013), "Estimation of class-size effects, using Maimonides' rule and other instruments: The case of French junior high schools." *Annals of Economics and Statistics*, 111, 193–225. [794, 795]

Geweke, J. and M. P. Keane (1997), "Mixture of Normals Probit Models." Research Department Staff Report 237, Federal Reserve Bank of Minneapolis. [799]

Gomes-Neto, J. B. and E. A. Hanushek (1994), "Causes and consequences of grade repetition: Evidence from Brazil." *Economic Development and Cultural Change*, 43 (1), 117–148. [784]

Grenet, J. (2010), "Academic performance, educational trajectories, and the persistence of date-of-birth effects: Evidence from France." Unpublished manuscript, Center for Economic Performance, London School of Economics, London, UK. [787]

Heckman, J. J. (2010), "Building bridges between structural and program evaluation approaches to evaluating policy." *Journal of Economic Literature*, 48 (2), 356–398. [782]

Heckman, J. J. and E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73 (3), 669–738. [782]

Holmes, T. C. (1989), "Grade level retention effects: A meta-analysis of research studies." In *Flunking Grades: Research and Policies on Retention*, Falmer Press, Bristol. [784]

Holmes, T. C. and K. M. Matthews (1984), "The effect on nonpromotion on elementary and junior high-school pupils: A meta-analysis." *Review of Educational Research*, 54 (2), 225–236. [784]

Hoxby, C. M. (2000), "The effects of class size on student achievement: New evidence from population variation." *Quarterly Journal of Economics*, 115, 1239–1285. [794]

Imbens, G. W. and J. D. Angrist (1994), "Identification and estimation of local average treatment effects." *Econometrica*, 62 (2), 467–475. [782]

Jacob, B. A. and L. Lefgren (2004), "Remedial education and student achievement: A regression-discontinuity analysis." *Review of Economics and Statistics*, 86 (1), 226–244. [784]

Jacob, B. A. and L. Lefgren (2009), "The effect of grade retention on high school completion." *American Economic Journal: Applied Economics*, 1 (3), 33–58. [784]

Kasahara, H. and K. Shimotsu (2009), "Nonparametric identification of finite mixture models of dynamic discrete choices." *Econometrica*, 77 (1), 135–175. [799]

Kotlarski, I. I. (1967), "On characterizing the gamma and normal distributions." *Pacific Journal of Mathematics*, 20, 69–76. [791]

Mahjoub, M.-B. (2007), "The treatment effect of grade repetitions." Unpublished manuscript, Paris School of Economics, Paris, France. [784, 787]

Mahjoub, M.-B. (2009), *Essais en micro-économie de l'éducation*. Ph.D. thesis, P, R. Gary-Bobo, supervisor, University of Paris I Panthéon-Sorbonne. [787]

Manacorda, M. (2012), "The cost of grade retention." *Review of Economics and Statistics*, 94 (2), 596–606. [784]

McLachlan, G. and D. Peel (2000), *Finite Mixture Models*. John Wiley and Sons, New York. [799]

Neal, D. and D. Whitmore-Schanzenbach (2010), "Left behind by design: Proficiency counts and test-based accountability." *Review of Economics and Statistics*, 92 (2), 263–283. [784]

Piketty, T. and M. Valdenaire (2006), "L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français. Estimations à partir du panel primaire 1997 et du panel secondaire 1995." Les Dossiers, Ministère de l'Education Nationale, no. 173. [794, 795]

Co-editor Petra E. Todd handled this manuscript.

Submitted December, 2014. Final version accepted January, 2016.