DON'T HOLD BACK? THE EFFECT OF GRADE RETENTION ON STUDENT ACHIEVEMENT

Ron Diris
Department of Economics
Maastricht University
6200 MD Maastricht, The Netherlands
r.diris@maastrichtuniversity.nl

Abstract

This study analyzes the effect of age-based retention on school achievement at different stages of education. I estimate an instrumental variable model, using the predicted probability of retention given month of birth as an instrument, while simultaneously accounting for the effect of month of birth on maturity at the time of testing. The analysis further assesses heterogeneity in retention effects by achievement, by background characteristics, and by type of skill. Using international data from multiple waves of the PISA international assessment test, I find that grade retention in primary school harms student achievement across the distribution, while delayed school entry can produce positive results for those at the lower end. The identified local average treatment effect indicates that letting students retain in primary school because of a low relative age is harmful for their future school achievement.

doi:10.1162/EDFP_a_00203

1. INTRODUCTION

Grade retention is a common approach in dealing with students who lag behind their classroom peers. There is an ongoing debate on the effectiveness of holding children back a year in response to learning difficulties. Students could benefit from developing more fundamental skills before they progress to the next grade.[1] On the other hand, literature shows that remediation of early achievement gaps is difficult (Cunha et al., 2006), and retention could negatively affect self-esteem and motivation (Jimerson, Anderson, and Whipple 2002). Empirical studies have traditionally found that retention harms student achievement, especially in the long run, and is also associated with higher school dropout.[2] Studies that thoroughly address the inherent endogeneity of a student's retention status have appeared only recently, however. Retention is also unlikely to affect different types of students equally. Evidence on heterogeneity in the effects of retention across student characteristics is still limited. Additionally, given that the learning process is strongly dynamic, the implications of retention at different stages of education can be very different as well.

The aim of this study is to estimate the effects of delayed school entry and grade retention in primary school and analyze how these effects differ by student characteristics and type of skill. An instrumental variable (IV) approach is used to estimate the impact of retention. The IV model exploits the strong correlation between retention status and month of birth. The analysis is based on multi-country data from the 2003, 2009, and 2012 PISA cohorts.[3] I identify a substantial negative local average treatment effect (LATE) of retention in primary school (Early Grade Retention) on school achievement, and a negligible effect of delayed school entry (Late Start). The LATE that is identified by the IV model indicates that letting students retain a grade in primary school because of a low relative age in class is harmful for future levels of school achievement. An alternative IV model that relies on stronger assumptions suggests that the impact on achievement of retention in secondary school is of roughly similar size as that of retention in primary school. The main IV model is combined with a quantile regression approach to estimate effects by percentile of the achievement distribution. Early Grade Retention (GRE) produces negative impacts across the distribution while Late Start (LS) can have positive effects for those near the bottom of the distribution. There is also some evidence that retention especially harms boys and those with highly educated parents. Further analysis shows that the effect of GRE on achievement is present across questions of all difficulty levels, including the very easiest test questions.

Studies on the impact of retention have been traditionally conducted in educational research but the topic has received increasing attention from economists in recent years. These recent papers have put more emphasis on addressing the endogeneity of retention treatment. They predominantly use a regression discontinuity design (RDD), which exploits the fact that retention is often based on specified achievement thresholds. Jacob and Lefgren (2004) and Roderick and Nagaoka (2005) apply this approach to public schools in Chicago, Greene and Winters (2007, 2012) and Schwerdt and West (2013) to public schools in Florida, and Manacorda (2012) to students in Uruguay. These

1. A belief often voiced by teachers, as documented by Tomchin and Impara (1992).
2. See, for example, overviews by Holmes and Matthews (1984) and Grissom and Shepard (1989).
3. PISA is an international assessment test. See section 3 for more details.

studies identify treatment effects for retention that are generally less harmful than those estimated by previous studies relying on ordinary least squares (OLS) or propensity score matching approaches. A common pattern is that early grade retention (grade 3) leads to short-term gains in achievement that decrease over time and are either small or statistically insignificant after several years. Jacob and Lefgren (2004) and Roderick and Nagaoka (2005) both find that later grade retention (grade 6) harms student achievement. Manacorda (2012) finds a consistently negative effect for retention in grades 7–9 for students in Uruguay. RDD approaches have also noted that retention increases the probability of high school dropout (Jacob and Lefgren 2009) and has negative (short-run) effects on behavior (Özek 2015). Eide and Showalter (2001) alternatively use an IV approach to address the endogeneity problem in estimating retention effects, using the number of days between birth date and school cutoff date as an instrument. They identify positive but statistically insignificant estimates on dropout and earnings, using long-run data from the United States. The study does not account for maturity differences between students with different birth dates.

The current study uses an IV approach that identifies the fraction of retained students within a certain month of birth and country as an instrument for retention. Using an instrument based on month of birth requires simultaneously correcting for the effect of month of birth on maturity at the time of the test. These two effects can be separated because the effect of month of birth on retention is not strictly proportional to the effect of relative age on test scores, as retention decisions are based on more than achievement. As such, there is still variation in retention prevalence across birth months after including a control function for maturity.

Additionally, this study analyzes heterogeneity in the impact of retention. Marginal treatment effects identified by RDD studies are very valuable because they tell us whether retention is executed too often—but the effects of retention can be very different for different types of students. This study designs an approach that estimates effects by percentile of the achievement distribution. It furthermore assesses to what extent the effect of retention differs by timing (in kindergarten versus in primary school), by type of skill (by school subject and difficulty level of the questions), and by background characteristics (gender, parental background, and country-level indicators). Additionally, previous studies on the effects of retention are country-specific, whereas this study uses a multi-country analysis. This reduces the odds that the estimates are particular to a specific school system. Moreover, the dataset used here is predominantly European, whereas the majority of the empirical literature is focused on either the United States or South America.

This paper proceeds as follows. Section 2 specifies a theoretical framework on the relation between retention and skill formation. Section 3 describes the PISA data that are used in the estimation. Section 4 specifies the methodology of the study. Results are discussed in section 5, both for the main IV model and additional heterogeneity analyses. Robustness analyses are presented in section 6. Section 7 concludes.

2. THEORY

This section presents a theoretical depiction of the relationship between retention and a student's learning process. From an economic point of view, the choice between

retention or promotion is a choice between two different investment paths. Cunha and Heckman's (2007) Technology of Skill Formation specifies that the development of skills over the lifecycle is a function of the skill set already present in period $t(\theta_t)$, parental characteristics (h), and investments made in period $t(I_t)$:

$$\theta_{t+1} = f_t(\theta_t, h, I_t).$$

The development of skills over the lifetime depends on the impact of investments, and on the self-productivity of skills that are already accumulated—skills beget skills. Schooling represents one important type of investment. Let us characterize I_t^g as the total investment of a school into a student in grade g at time t, which can be seen as the curriculum for that grade. At point t, the choice between promotion (p) and retention (r) is a choice between investment I_{t+1}^{g+1} and (repeated) investment I_{t+1}^{g} for the following year. Let us assume (for now) that retention simply postpones all curricular investments by one year. The effect of retention after one year ($\theta_{t+1}^r - \theta_{t+1}^p$) depends on whether the student learns more from the promoted curriculum or from the repeated curriculum. Where this effect is a priori unclear, it should hold that the repeat student enters grade $g+1$ with a higher θ, at point $t+1$, than his promoted counterfactual did, at point t ($\theta_{t+1}^{r,g+1} > \theta_t^{p,g+1}$), assuming that the repeated curriculum I_{t+1}^g has at least some positive effect. Moreover, as skills are self-productive, the larger set of incumbent skills at the beginning of grade $g+1$ also implies that student r learns more in grade $g+1$ than counterfactual student p did:

$$\theta_{t+2}^{r,g+1} - \theta_{t+1}^{r,g} > \theta_{t+1}^{p,g+1} - \theta_t^{p,g}.$$

Two preliminary conclusions follow. First, the negative impact of retention can never exceed a grade equivalent (i.e., what is learned in one year of schooling). Second, if the effect of retention is negative after one year, this negative effect will gradually decrease over time. Alternatively, when the effect of retention is positive after one year it will gradually accelerate over time.

These conclusions only consider the self-productivity aspect of skill formation and do not consider that retention can impact investments or inputs other than the yearly curriculum, however. Retention could affect the attention given to the child by teachers, parents, and peers. It could also affect the odds of attending a specific school track or participation in specific programs such as summer schools or extra tutoring classes. Additionally, retention can have effects on the student's willingness to exert effort in learning. A priori, the directions of these effects are not clear. Student self-esteem could be improved by facing a less-demanding curriculum, but it could also be reduced through the potential shock and negative stigma associated with retention. The impact on (other) investments largely depends on whether the approach taken to differences in achievement (by parents, teachers, schools, or other relevant decision makers) is more focused on remedying the achievement gaps of relatively less able students or rather focused on maximizing the talents of more able students. If such investment effects are present, neither of the previous two conclusions necessarily holds anymore—negative retention effects can accelerate over time if retention lowers investment and effort in the future, and the effect of retention can also potentially exceed a grade equivalent. A common finding among RDD studies is that early retention effects are initially low or even positive but worsen over time. This goes against the self-productivity argument

and therefore implies the impact of retention on other investments in skill formation is important (at least for the marginal student).

Retention effects are likely to be heterogeneous. The model indicates that retention is more favorable the higher the impact of I_{t+1}^{g} and the lower the impact of I_{t+1}^{g+1}. The impact of the more demanding I_{t+1}^{g+1} likely increases with ability, whereas the relationship between ability and I_{t+1}^{g} could be more complex. High-ability students are faster learners but they are also likely to be subject to strong diminishing returns from receiving a repeated curriculum. As such, large benefits for students of very high ability are unlikely, but the relationship between the effect of retention and ability is not necessarily monotonic.[4] Additionally, students of different ability might differ in the way that retention affects future inputs (e.g., students from more affluent families might receive more support in response to retention).

With respect to heterogeneity by the *age* at which retention occurs, several relevant aspects are implied by the model. First, self-productivity implies that earlier retention has more potential to be beneficial because there is a longer period in which the student may benefit from entering a specific grade with a higher skill level than when promoted, and also because self-productivity tends to be stronger at younger ages (Cunha and Heckman 2008). On the other hand, earlier retention also implies that effects of retention on other investments operate and accumulate over a longer period of time. Additionally, the timing of retention can also determine *how* investments are affected. For example, it is often argued that negative psychological effects of retention are lower for early retention, because it does not remove children from long-established peers and because the social stigma from being held back a year in kindergarten is generally assumed to be smaller.[5]

A final caveat to the framework presented here is that the potential self-productivity gains of retention assume that the repeated curriculum has at least some positive impact at $t + 1$. Nevertheless, skills also depreciate and retention can cause a shock to the learning process through stigma and placement with new classroom peers, which could imply that $\theta_{t+1}^{r,g} > \theta_{t}^{r,g}$. Given that current empirical evidence generally shows that (early) retention effects are positive in the very short run but weaken or reverse over time,[6] this appears unlikely to hold in reality.

3. DATA

The data in this study are captured from PISA. PISA is an international test of school achievement of 15-year-old students. Studies were conducted in the years 2000, 2003, 2006, 2009, and 2012. PISA tests students in reading, mathematics, and science. Standardized achievement test scores in each of these topics serve as dependent variables in the analysis. Scores are based on the mean of the five plausible values that are reported in the data. These scores are standardized to have a mean of 0 and a standard deviation of 1.

4. Cooley-Fruehwirth, Navarro, and Takahashi (2016) identify that high-ability repeat students benefit especially from retention, but these "treatment on the treated" effects should not be generalized to all high-ability students, as repeat students of high ability are likely to be very specific (and rare) cases.
5. See, for example, Shepard and Smith (1986).
6. See, for example, Jacob and Lefgren (2004), Roderick and Nagaoka (2005), and Greene and Winters (2007, 2012).

The data contain students who are born within one year of each other in a particular country, which is generally a specific calendar year. The majority of students are in grade 10. A first restriction on the set of countries included in the analysis of this study is that they are all Organization for Economic Co-operation and Development (OECD) countries to ensure a relatively homogenous group of countries with well-developed educational systems. The analysis further requires that we compare students who would be in the same grade in absence of retention. This implies that there is a specific and known school cutoff date from which point on students are allowed to enroll in formal education. As such, countries without a clear cutoff date are excluded, as well as countries with region-specific cutoff dates for which it was not possible to reliably identify the cutoff date to which each student was exposed (e.g., Australia and the United States). Additionally, I exclude countries where the frequency of retention is very low (<1%). Finally, some countries are dropped because of data irregularities. A separate online appendix that can be accessed on *Education Finance and Policy*'s Web site (www.mitpressjournals.org/doi/suppl/10.1162/EDFP_a_00203) contains a more detailed discussion on why certain countries are excluded from the final sample.

Several countries in PISA include students from a specific calendar year, although the cutoff date lies during the year (generally in either September or October). Given the setup of the analysis requires that students are in the same grade in absence of retention, students born on the right side of the cutoff date (i.e., born after September or October) are excluded for these countries. This applies to six country cases—Austria, Estonia, Luxembourg, the Netherlands, Slovakia, and (part of) Canada. The countries included in the analysis are: Austria, Belgium, Canada, Denmark, Estonia, Finland, France, Italy, Luxembourg, the Netherlands, New Zealand, Poland, Portugal, Slovakia, Spain, and Sweden.

There are three main reasons for choosing the PISA data for the analysis. First of all, they consist of students who are born within one year of each other, which ensures repeat students and nonrepeat students can be compared at the same age level. Additionally, PISA is aimed at testing general knowledge. Tests therefore are not biased toward the curriculum of a specific grade. Finally, PISA contains information on the stage at which students are retained. This allows one to identify delayed school entry (labeled as Late Start), retention in primary school (labeled as Early Grade Retention) and retention in secondary school (labeled as Late Grade Retention). The latter two groups can be identified through the question in PISA that asks students whether they were ever retained in either primary (ISCED 1) or secondary school (ISCED 2 and ISCED 3), and the former by looking at those who are one grade behind but report that they never repeated any formal grade (or are two grades behind and report only one occurrence of retention). Repeat students with missing information on the retention questions and students for which the number of reported retention occurrences exceeds the number of grades behind the modal grade are excluded from the sample.[7] Table 1 shows

7. This concerns 5.8 percent of all repeat students. An exercise in which all these observations are jointly labeled as either Late Start, Early Grade Retention, or Late Grade Retention produces virtually identical results as those found for the main estimation.

Table 1. Degree of Retention by Country

	LS	GRE	GRL
Austria	0.125	0.028	0.080
Belgium	0.042	0.161	0.190
Canada A	0.030	0.022	0.0050
Canada B	0.023	0.066	0.103
Denmark	0.124	0.030	0.0060
Estonia	0.029	0.027	0.022
Finland	0.024	0.027	0.0048
France	0.021	0.159	0.203
Italy	0.027	0.010	0.143
Luxembourg	0.075	0.213	0.242
Netherlands	0.063	0.215	0.083
New Zealand	0.041	0.013	0.0064
Poland	0.0047	0.017	0.027
Portugal	0.063	0.169	0.162
Slovakia	0.137	0.025	0.017
Spain	0.019	0.097	0.274
Sweden	0.0053	0.021	0.0042

Notes: The table lists all countries included in the sample and the corresponding shares of Late Starts (LS; delayed school entry), Early Grade Retention (GRE; retention during ISCED 1), and Late Grade Retention (GRL; retention during ISCED 2 and ISCED 3) at age 15 years. Canada A refers to Canadian provinces with a 1 January cutoff date, Canada B refers to Canadian provinces with a 1 October cutoff date.

mean retention rates at different stages of education for all countries included in the sample.[8] Only students in the 2003, 2009, and 2012 PISA cohorts report information on when they were retained, hence only those three cohorts are included in the analysis. The total number of observations for all cohorts and countries jointly equals 344,551.

4. METHODOLOGY

Counterfactual

Because one cannot observe the same person in both a retained and nonretained state, one needs to compare retained students with a group of nonretained students. This comparison can be conducted at an age (i.e., birth cohort) level or at a grade level. Among previous studies, both approaches are common. Choosing a grade-level comparison favors retained students compared with age-level comparisons, as retained students are older and have had more schooling in the former case. Estimating the causal effect of retention requires identifying the difference in achievement when retained, compared with the achievement when that same student would not have been retained.

8. The remainder of the study will use "retention" as a general term encompassing grade repetition at any of the three identified stages, including Late Start.

This comparison should take place at the same age, because the counterfactual should be of the same age as the actual observation.[9]

Same-age comparisons automatically imply that promoted students are in a higher grade than retained students. As such, promoted students might have been exposed to additional curricular topics. If such topics appear in the test, this can provide a disadvantage for retained students that can be seen as unfair. This issue can be circumvented if the data contain vertically-scaled test scores that are aligned to the content of each grade, as for example used by Schwerdt and West (2013), but such tests are not available here. Moreover, this disadvantage is essentially a genuine part of the effect of grade retention. Retained students are always "behind" on the investment path. Whenever they catch up, their nonretained peers have learned something new again, built up work experience, and so forth. Curricular issues are further assessed in section 6.

IV Model

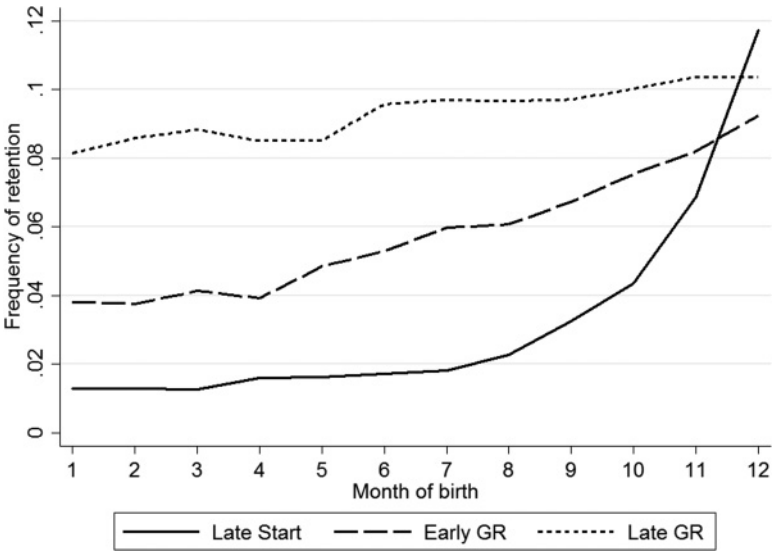
The aim of this study is to estimate the impact of Retention (R_i) on student achievement (A_i). The traditional approach in estimating retention effects is to use OLS analysis and control for a wide range of observed individual characteristics (X_i'):

$$A_i = \beta_0 + \beta_1 R_i + \sigma X_i' + \varepsilon_i. \tag{1}$$

However, R_i likely is strongly correlated with other determinants of achievement, and it is unlikely that all these determinants are observed by the researcher. This assumed correlation between R_i and the error term ϵ_i will bias the estimates of β_1. I address this issue by using an IV model that exploits variation in month of birth. Figure 1 shows the pattern of retention by month of birth, separately for delayed school entry, retention in primary school, and retention in secondary school. Month of birth is recoded so that month 1 represents the first month after the cutoff date for school entry. For all stages, there is a clear positive correlation between retention and month of birth; students born just before the cutoff date are more likely to retain than students born earlier in the year. The figure also shows that the strength and functional form of the pattern differs across the three stages.

There are two primary concerns with respect to the validity of an instrument based on month of birth. One is that an individual's birth month can be related to (family) background characteristics, the other is that month of birth is correlated with maturity at the time of the test. The first concern is elaborately addressed in section 6. The second concern is a more pressing matter. Previous studies that also rely on date of birth to instrument retention have argued that maturity effects fade out at later ages (see, e.g., Eide and Showalter 2001; Garcia-Perez, Hidalgo-Hidalgo, and Robles-Zurita 2011) but there currently exists extensive evidence that maturity effects persist into later grades (see, e.g., Bedard and Dhuey 2006; Borghans and Diris 2014). Analysis in this study confirms this. A cross-sectional regression reveals a statistically significant and negative relation between month of birth and student achievement at age 15 years, controlled for retention status. As the OLS estimate of the effect of retention is likely to be negatively

9. Grade comparisons are still informative but from an economic point of view should always be weighed against the cost of retention in terms of extra educational expenditures and the opportunity costs of losing a year in time.



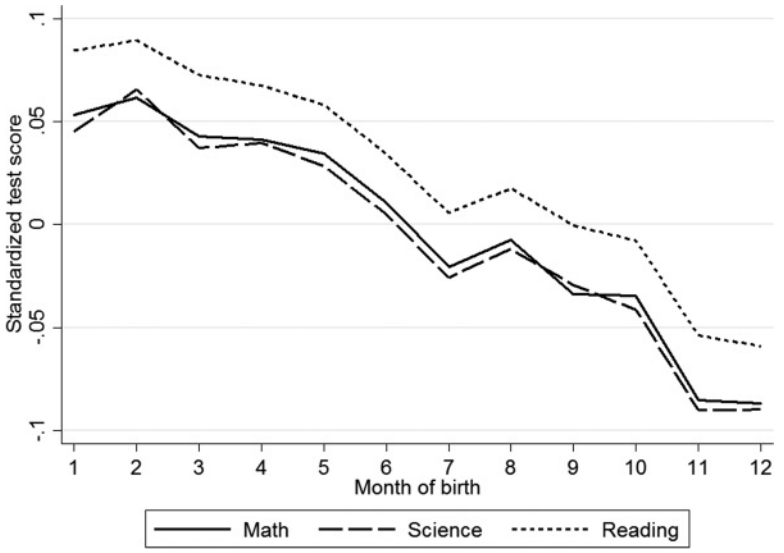
Notes: The figure shows the average level of delayed school entry (LS), retention during ISCED 1 (GRE), and retention during ISCED 2–3 (GRL) over the month of birth of the corresponding students. Month of birth is counted from the official school cutoff date on, 1 being the first month after the cutoff date and 12 being the month just before the cutoff date for the next year. Countries that do not have observations for each month of birth (Austria, Canada region B, Estonia, Finland, Luxembourg, the Netherlands, Portugal, and Slovakia) are excluded for the construction of this figure.

Figure 1. Frequency of Retention by Month of Birth.

biased, this estimate of the maturity effect is likely underestimated (scores of repeat students are conditioned too strongly and this is more prominent in later birth months). Hence, the fact that the maturity estimate is statistically significant while it is likely biased toward zero is strong evidence that maturity effects are present at age 15 years.

It is still feasible to exploit differences in month of birth as an instrument for retention, even when maturity effects are controlled for. The underlying reason is that differences in the prevalence to retain across birth months are not necessarily proportional to achievement differences across birth months. Tomchin and Impara (1992) analyze how much weight teachers put on different types of information in making decisions on retention. Their results show that achievement and ability are the most important factors but maturity and relative age in class are not far behind. Home environment, gender, and physical size also play a role. The influence of maturity and relative age is especially high in early grades. Hence, aside from the fact that relatively younger students are less likely to obtain passing grades because of lower achievement, relative age is also taken into account in making retention decisions beyond this achievement effect. Moreover, the nonlinear pattern for Late Start in figure 1 indicates that such relative age considerations are given especially high emphasis for those born in the last few months before the cutoff date.

The analysis in Tomchin and Impara (1992) essentially reveals that the decision-making process for letting a student repeat or progress depends on three types of information: achievement, relative age, and background factors. The change in the functional form and slope of the patterns in figure 1 is a result of a shifting in the weights in this decision process. A higher weight on relative age leads to disproportionate



Notes: The figure shows the average achievement test scores for mathematics, science, and reading by month of birth. Month of birth is counted from the official school cutoff date on, 1 being the first month after the cutoff date and 12 being the month just before the cutoff date for the next year. Countries that do not have observations for each month of birth (Austria, Canada region B, Estonia, Finland, Luxembourg, the Netherlands, Portugal, and Slovakia) are excluded for the construction of this figure.

Figure 2. School Achievement by Month of Birth.

increases in retention probability for later months; a higher weight on achievement leads to more linear, but still upward-sloping, patterns; a higher weight on background factors leads to a flatter pattern (as these are presumably exogenous to the birth month). The pattern for Early Grade Retention does not exhibit the same degree of nonlinearity as the pattern for Late Start, suggesting that (disproportionate) relative age considerations play a smaller role, but it does show a positive and slightly increasing slope after a flat pattern across the first four months. For Late Grade Retention, the slope is low and linear, which indicates that achievement and background factors largely dominate the decision process.[10]

To exploit the "disproportionality" of the effect of month of birth on retention prevalence, I create an instrument that measures the share of retained students for every cohort-country-month cell, while including a "smooth" control function for relative age in class. To avoid correlation between individual retention status and the mean level of retention, as well as a potential bias from a "bad draw" of students in a particular month and cohort, I take the mean level of retention for the country-month cell averaged over the other two PISA cohorts as the instrument value for each cohort-country-month cell. The instrument is separately calculated for each type of retention so that the effects of retention can be estimated at multiple stages of the educational process.

It is important for the validity of the estimates that the control for maturity has the proper functional form. Figure 2 shows the relation between month of birth and

10. An alternative explanation is that the less steep pattern for Late Grade Retention occurs because there are more low-ability students already filtered out by earlier retention. However, the pattern is similar when we restrict the sample to countries that have little retention before secondary school.

achievement is fairly linear. Indeed, a linear control for month of birth provides a strong fit but previous research has shown that maturity effects are more severe for earlier school starting ages (Bedard and Dhuey 2006; Borghans and Diris 2014). Borghans and Diris (2014) show that the best fit is provided by jointly including a linear and a quadratic term of the school starting age in months that is predicted by the student's month of birth and the official school cutoff date (labeled *PS*). This specification captures that maturity effects are slightly stronger for students with lower (predicted) starting ages. The instrumentation is too weak for conventional thresholds of first stage power to estimate the effect of Late Grade Retention, as there is essentially no relation between month of birth and Late Grade Retention, conditional on maturity. The instrumentation for Late Start and Early Grade Retention is very strong.[11] I conduct an analysis in section 5 on the impact of Late Grade Retention that relies on somewhat more restrictive assumptions. The main estimation model for estimating the effects of Late Start (LS) and Early Grade Retention (GRE) becomes:

$$\begin{aligned} LS_i &= \alpha_o + \alpha_1 \rho_i^{LS} + f_k\left(PS_i\right) + \theta X_i' + \kappa_{co} + \varepsilon_i \\ GRE_i &= \beta_o + \beta_1 \rho_i^{GRE} + f_k\left(PS_i\right) + \zeta X_i' + \delta_{co} + \psi_i \\ A_i &= \gamma_o + \gamma_1 \widehat{LS}_i + \gamma_2 \widehat{GRE}_i + f_k\left(PS_i\right) + \sigma X_i' + \omega_{co} + \eta_i. \end{aligned} \tag{2}$$

The terms ρ_i^{LS} and ρ_i^{GRE} refer to the previously defined instruments for LS and GRE, respectively. A_i can refer to achievement scores in math, science, or reading, and also to an overall score that takes the average score of the three domains tested in PISA. $f_k\left(PS_i\right)$ indicates the polynomial control function for maturity, which contains a linear and a quadratic term in the main approach. The vectors κ_{co}, δ_{co}, and ω_{co} capture fixed effects for every combination of country (c) and cohort (o). A vector of controls (X_i') is also included, comprising gender, family structure, ethnicity, language spoken at home, parental education level, working status of mother and father, and a composite measure of home possessions. ε_i, ψ_i, and η_i represent classical error terms. The model does not control for Late Grade Retention, as any effect of Late Start and Early Grade Retention on the probability of later retention is a genuine part of the total treatment effect and therefore should not be controlled for.[12] To ensure that the sample is representative of the total population, inverse sampling probabilities as reported by PISA are used as weights for each observation. Additionally, observations are weighted so that the joint weight of each country-cohort cell is identical. As such, the results are not driven by one or a few large countries and it is ensured that each educational system has the same weight.

The model controls for the impact of relative age on test scores but still assumes a homogeneous maturity effect for a given starting age. This implies that the estimation exploits not only the disproportionality in the effect of month of birth on retention, but also differences in the strength of this relationship between countries. This can be problematic when such differences are driven by variation in country-specific maturity effects. Nevertheless, descriptive evidence indicates that countries with a stronger

11. Kleibergen-Paap test statistics for first stage power are reported in the last row of table 2. The test statistic drops to 0.83 when Late Grade Retention is also included.
12. Sensitivity of the estimates to including GRL as an additional control is assessed in section 6.

relationship between month of birth and retention are not more likely to also have a stronger relationship between month of birth and achievement.[13] It is found that countries in which the relationship between month of birth and retention is weaker are also countries in which retention is based—to a relatively strong extent—on sociodemographic factors (such as gender, ethnicity, and parental education).[14] The assumption of homogeneous maturity effects is more extensively addressed in section 6.

5. RESULTS

This section reports results on the estimated effect of Late Start and Early Grade Retention on student achievement. Results are presented for both an OLS approach (model 1), the main IV model (model 2), and additional heterogeneity analysis. An alternative model that also instruments the effect of secondary school retention is presented later (see subsection "Retention in Secondary School"). All reported coefficients refer to the estimated effect of retention on achievement test scores that are standardized with a mean of 0 and a standard deviation equal to 1.

Overall Results

The OLS estimates are portrayed in the top panel of table 2. The OLS model without control variables provides estimates of −0.658 for LS and −1.28 for GRE, for the overall score. This reduces to −0.490 and −1.05, respectively, when controls are added. The bottom panel of table 2 shows results for the IV model. Instrumental variable estimates for LS are all low. Estimates for GRE are negative and statistically significant across domain scores. Science achievement is most strongly affected and math achievement least strongly, and this difference is statistically significant. The estimate for the overall score equals −0.486 (in the model with controls). Table 2 also shows that the differences between the estimates with controls and the estimates without controls are very small. This lends support to the validity of the instrument, as it is not strongly correlated with observable student characteristics. This issue is further explored in section 6.

The magnitude of the effect of GRE on achievement is sizable. The OECD calculates, for each PISA edition, the score equivalent of a grade level (i.e., the gain in achievement from being one grade higher). This estimated effect is around 40 points for the OECD on average, but equals 47.4 points if we only include the countries used in the sample of this study.[15] The estimated effect of GRE on the unstandardized overall score equals −41.8 points. Hence, the estimates indicate that GRE leads to a negative effect on school achievement that requires around 0.88 of a full year of education to compensate. This also suggests that the estimated effect of retention in a grade-level comparison would be low but positive (although given the current level of precision

13. I test the correlation between both parameters for all forty-seven country-cohort observations. The crude correlation is negative and statistically significant, but this estimate also incorporates the treatment effect of retention. When the effect of maturity on score is estimated only for the first five months after the cutoff (where the retention prevalence remains relatively flat) or when scores of repeat students are corrected with the estimates obtained in the main analysis, the correlation is low and statistically insignificant. See figure A1 in the online appendix.
14. There is a very strong negative correlation (–0.71) between the share of the variance in retention explained by X' and the strength of the relationship between month of birth and retention for each country-cohort combination. See figure A2 in the online appendix.
15. See OECD (2004, 2009, 2013) for grade equivalent calculations for PISA 2003, 2009, and 2012, respectively.

Table 2. Effect of Late Start (LS) and Early Grade Retention (GRE) on School Achievement

	LS	GRE	LS	GRE
Panel A: OLS — model 1				
Math score	−0.629***	−1.25***	−0.495***	−1.04***
	(0.016)	(0.012)	(0.014)	(0.011)
Science score	−0.612***	−1.19***	−0.450***	−0.965***
	(0.016)	(0.013)	(0.014)	(0.011)
Reading score	−0.648***	−1.24***	−0.462***	−0.994***
	(0.016)	(0.013)	(0.014)	(0.012)
Overall score	−0.658***	−1.28***	−0.490***	−1.05***
	(0.016)	(0.013)	(0.014)	(0.011)
Panel B: IV — model 2				
Math score	−0.0093	−0.471***	0.000047	−0.400***
	(0.054)	(0.141)	(0.049)	(0.134)
Science score	0.012	−0.616***	0.020	−0.542***
	(0.057)	(0.146)	(0.052)	(0.139)
Reading score	0.048	−0.537***	0.056	−0.453***
	(0.054)	(0.140)	(0.049)	(0.132)
Overall score	0.018	−0.566***	0.027	−0.486***
	(0.055)	(0.143)	(0.050)	(0.136)
KP-stat	175.61		174.95	
Controls	No	No	Yes	Yes

Notes: The table reports the estimated effects of Late Start (delayed school entry) and Early Grade Retention (retention during ISCED 1) on achievement test scores, using either an OLS approach (model 1) or an IV approach (model 2), each both with and without the vector of controls X'. Test scores are standardized with a mean of 0 and a standard deviation of 1. Controls included are gender, family structure, ethnicity, language spoken at home, parental education, working status of mother and father, and a composite measure of home possessions. Standard errors are in parentheses and are robust and corrected for clustering at the school level. KP-stat = Kleibergen-Paap statistic.

***Significant at the 1% level.

it would not be statistically distinguishable from zero). As discussed in section 2, the effect of retention will always be smaller than a full grade equivalent if retention has no impact on other investments. The fact that the estimates are close to a grade equivalent indicates that either the gain from redoing a grade and the corresponding future self-productivity gains are very low or future inputs are indeed negatively affected.

As indicated by Imbens and Angrist (1994), an IV model estimates a LATE for individuals for whom the treatment indicator responds to changes in the instrument. In this case, these are students for whom the retention status is responsive to changes in their month of birth. These individuals are being retained because decision makers put (disproportionate) emphasis on relative age in retention decisions. As such, these are not necessarily students that lack the innate ability to understand the curriculum for the modal grade. We know from previous research that achievement gaps by month of birth are relatively large in early grades, but that relatively younger students in class (partially) catch up in later grades.[16] This reducing relative age gap indicates that achievement growth in early grades is higher for relatively young students, when promoted, than for their older peers. The results from the main analysis suggest that retention

16. Crawford, Dearden, and Meghir (2010) find this for England, where retention is absent, whereas other studies such as Bedard and Dhuey (2006) estimate an IV model for those who comply to their predicted grade. Hence, these results strictly represent relative age effects and are not directly affected by retention.

can distort that process of catching up and therefore could be especially harmful for these type of students. As such, the estimated LATE is likely to be more negative than the average treatment effect of retention. This also implies that the bias in the OLS estimates is even stronger than the difference between the OLS and IV estimates in table 2 would suggest, which highlights the severe endogeneity bias of OLS estimates of retention effects and underlines that controlling for observed characteristics only partially conditions for this. Additionally, a comparison with the results of recent RDD studies suggests the estimated LATE is substantially more negative than the effect for the student at the achievement margin for promotion.

The finding that delayed school entry is less harmful than retention in primary school is in line with other literature. For example, Jacob and Lefgren (2004) and Roderick and Nagaoka (2005) identify more favorable or less harmful treatment effects for earlier retention compared with later retention. It also corresponds to the theoretical framework, which specifies that the potential benefits through self-productivity effects of skills are larger the earlier a student repeats a grade. One has to consider, however, that each LATE is estimated at a different margin. As such, the difference in the estimates could also reflect that those who delay school entry (because of a low relative age) and those who repeat a grade in primary school (because of a low relative age) are different individuals. An analysis of background characteristics of each type of student reveals that both groups are disadvantaged compared with nonrepeat students, but GRE students are especially so. This pattern is not consistent across all indicators, however. Late Start has an especially strong link to having foreign-born parents, whereas GRE has an especially strong link to low parental education. This difference in the composition of each group could also contribute to the estimated difference in treatment effects.

Heterogeneity by Quantile

The results from section 5 indicate that the LATE of retention is zero at best, but this does not preclude that certain groups benefit from treatment. This section presents results from an analysis on the heterogeneity of retention effects across the achievement distribution. The quantile regression approach is based on the original work of Koenker and Bassett Jr. (1978) and the extensions provided by Abadie, Angrist, and Imbens (2002) and Chernozhukov and Hansen (2005).

The approach recodes the outcome variable to the rank in the achievement distribution. The rank is based on the percentage of people in the sample with a similar or higher test score for the particular test domain (within the country-cohort cell). Everyone below the percentile for which the effect is estimated receives the value of the lowest threshold, and everyone above that percentile receives the value of the higher threshold. For example, when estimating the effect for the third decile, those with a rank below 0.20 receive a rank of 0.20, those above 0.30 receive a rank of 0.30, and those in between 0.20 and 0.30 receive their precise rank. An example is given below, for the case in which 10 deciles are defined:

$$\begin{aligned} Score1_i &= \min(0.1, score_i) \\ Score2_i &= \min(0.1, \min(0.2, score_i)) \ldots \\ Score10_i &= \max(0.9, score_i). \end{aligned} \tag{3}$$

The method distributes the total treatment effect over all separate quantiles. All quantile effects add up to the total effect. If one wants to individually interpret the quantile estimates, they have to be rescaled. If retention would be equally divided over the distribution, it would simply be a matter of multiplying by 10 for deciles and 100 for percentiles. This is not the case, and I therefore instead divide by the number of retained students located in a certain quantile as a fraction of the total number of retained students in the whole sample.[17]

The results for this analysis are displayed in table 3, for the case where the ranking is defined by percentile. The analysis identifies that some students can benefit from delayed school entry. This especially holds with respect to reading scores, which are positively affected up until the 15th percentile, but also math and science scores of students very low in the distribution are improved by delayed school entry. The effect sizes for LS are modest. Estimates for GRE are negative and substantial. They are consistently statistically significant throughout the distribution. The effect of GRE on reading achievement is relatively lower for those low in the ranking and higher for those high in the ranking, compared with the estimates for math.

The quantile estimates are also graphically portrayed for the "overall score" in figure 3. The graph plots coefficients and 95 percent confidence intervals. Early Grade Retention is harmful from the first percentile on, whereas the effect of LS is positive and statistically significant (at the 5 percent level) up to the 9th percentile. Confidence intervals widen for higher percentiles, as relatively few repeat students are located there. This holds especially for GRE, as the instrumentation for LS is stronger and because LS is still relatively prevalent above the lowest deciles (see figure A3 in the online appendix).

It should be emphasized that the ranking in this quantile regression approach is based on an outcome variable that is also affected by the treatment, which is different from a heterogeneity analysis based on an ex ante measure of ability. A quantile regression estimate takes up those changes in the ranking that occur within the quantile and those that cross the quantile boundary. It is not straightforward to precisely induce from this to which achievement rank net of treatment an estimate pertains. If one, for example, takes the 10th percentile and if one knows that the treatment effect ranges between 0 and −0.10, this implies that the estimate for the 10th percentile takes up a mixture of the treatment effects of those who would be between the 10th and 20th percentile net of treatment.[18] Nonetheless, the results indicate that there is substantial heterogeneity in the effects of retention across the achievement distribution, that positive effects of delayed school entry are present for individuals low in the achievement distribution, and that the effect of GRE becomes more negative when we move upward in the distribution.

The results of this heterogeneity analysis contrast with findings by Cooley-Fruehwirth, Navarro, and Takahashi (2016), who find an opposite dynamic and even identify positive effects from retention for high-ability students in the United States.

17. For example, if 50 percent of all repeat students are located in the lowest decile, one divides the initial coefficient for the first decile by 0.5. If only 2 percent of all repeat students are in the 7th decile, one divides by 0.02 for that decile, and so on.

18. This issue mainly pertains to the estimates for GRE. Given the low magnitude of the LS estimates, the overlap between the ranking used here and a hypothetical ranking net of treatment should be large.

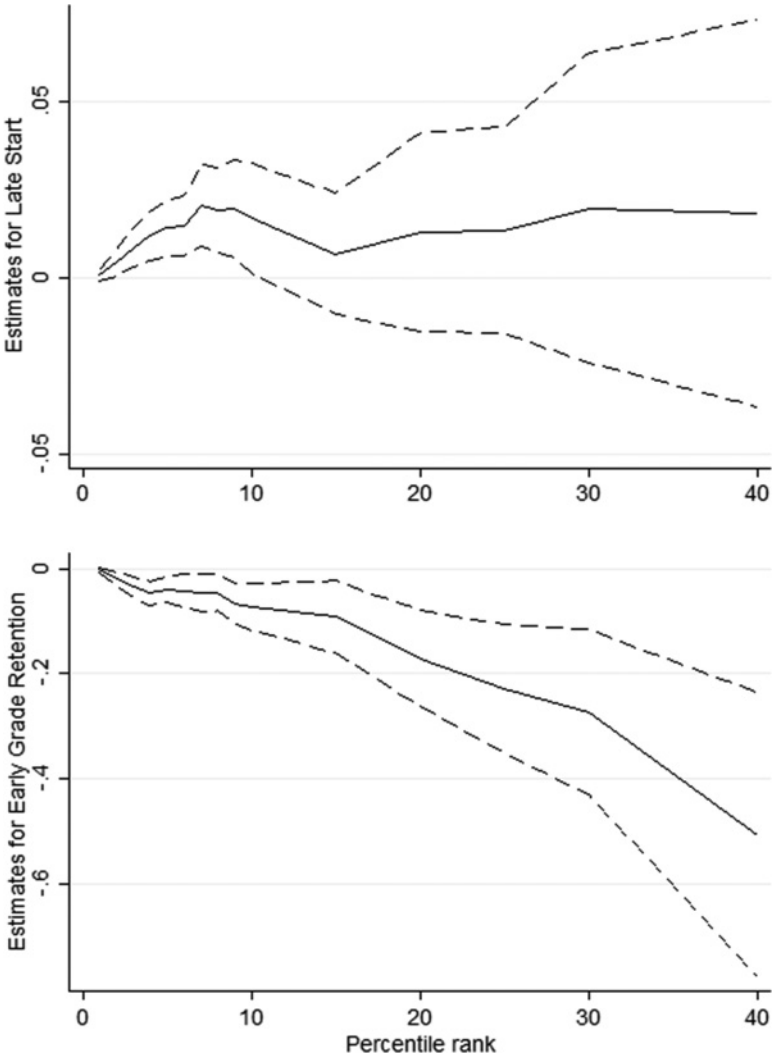
Table 3. Effect of Late Start (LS) and Early Grade Retention (GRE) by Percentile

	Math	Science	Reading		Math	Science	Reading
LS							
(1)	0.0017	0.00092	0.0015	(9)	0.0077	0.0095	0.018***
	(0.00068)	(0.00092)	(0.00086)		(0.0075)	(0.0082)	(0.0068)
(2)	0.0032	0.0041**	0.0051**	(10)	0.0072	0.013	0.020***
	(0.0018)	(0.0021)	(0.0020)		(0.0081)	(0.0080)	(0.0071)
(3)	0.0074***	0.0088***	0.010***	(15)	0.0076	0.016	0.017*
	(0.0027)	(0.0030)	(0.0030)		(0.011)	(0.011)	(0.0096)
(4)	0.0082***	0.010***	0.014***	(20)	0.0092	0.016	0.012
	(0.0032)	(0.0037)	(0.0039)		(0.014)	(0.015)	(0.012)
(5)	0.0077*	0.013***	0.018***	(25)	0.011	0.027	0.018
	(0.0040)	(0.0043)	(0.0044)		(0.017)	(0.018)	(0.018)
(6)	0.0075	0.012**	0.019***	(30)	−0.012	0.020	0.0074
	(0.0047)	(0.0056)	(0.0054)		(0.023)	(0.023)	(0.021)
(7)	0.0067	0.0089*	0.019***	(40)	0.011	0.011	0.011
	(0.0057)	(0.0054)	(0.0058)		(0.033)	(0.027)	(0.026)
(8)	0.0070	0.0093	0.020***	(All)	−0.0086	0.0056	0.0074
	(0.0058)	(0.0065)	(0.0063)		(0.015)	(0.015)	(0.015)
GRE							
(1)	−0.0036	−0.0044	−0.0038	(9)	−0.067***	−0.086***	−0.055***
	(0.0027)	(0.032)	(0.0031)		(0.021)	(0.024)	(0.020)
(2)	−0.013**	−0.021***	−0.022***	(10)	−0.083***	−0.084***	−0.063***
	(0.0060)	(0.0071)	(0.0074)		(0.024)	(0.026)	(0.023)
(3)	−0.025***	−0.039***	−0.025***	(15)	−0.138***	−0.107***	−0.085**
	(0.0085)	(0.010)	(0.0097)		(0.136)	(0.039)	(0.034)
(4)	−0.038***	−0.047***	−0.026**	(20)	−0.180***	−0.196***	−0.125***
	(0.011)	(0.013)	(0.012)		(0.054)	(0.049)	(0.042)
(5)	−0.037***	−0.053***	−0.032**	(25)	−0.187***	−0.240***	−0.168***
	(0.012)	(0.014)	(0.015)		(0.056)	(0.068)	(0.056)
(6)	−0.059***	−0.059***	−0.042**	(30)	−0.295***	−0.369***	−0.245**
	(0.016)	(0.018)	(0.017)		(0.088)	(0.103)	(0.072)
(7)	−0.051***	−0.070***	−0.040**	(40)	−0.420***	−0.481***	−0.410***
	(0.016)	(0.019)	(0.017)		(0.151)	(0.143)	(0.137)
(8)	−0.061***	−0.071***	−0.045**	(All)	−0.110**	−0.152***	−0.114***
	(0.019)	(0.019)	(0.019)		(0.039)	(0.040)	(0.037)

Notes: The table reports the effect of LS and GRE on the ranking of the student in the particular test subject, by percentile, using model 2 and coding the outcome variable as defined in equation 3. Original estimates are rescaled by dividing through the share of all Late Start or Early Grade Retention that is situated in that specific percentile. Ranking is defined on a country level and ranges from 0 to 1. Standard errors are in parentheses and are robust and corrected for clustering at the school level.

*Significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

The analysis in this study can accurately estimate treatment effects of, at most, students just below the median of the distribution. Still, the pattern of results across the lowest deciles is different when comparing this study with that of Cooley-Fruehwirth, Navarro, and Takahashi (2016). This difference can exist because of institutional differences in the nature of retention between the largely European dataset used in this study and the United States. For example, it might be that U.S. students of relatively higher ability benefit from additional investments and resources tied to retention, such as summer schools and increased guidance from schools. Another reason for the discrepancy can lie in the particular LATE that is identified by model 2. As discussed before, this estimated effect can be very different from the effect of retention on the average repeat student, which is what Cooley-Fruehwirth, Navarro, and Takahashi (2016) estimate.



Notes: The figure shows the effect of Late Start (LS) and Early Grade Retention (GRE) on achievement across the test score distribution. Effects are estimated by percentile, using Model 2 and coding the outcome variable as defined in equation 3. Coefficients indicate the effect of LS and GRE on the percentile ranking, which is defined on a country level and based on the mean of the student's score on math, science, and reading. The solid line provides the estimates, the dotted lines form 95% confidence intervals. Specification 3 recodes rankings from 0 to 1. Hence, a coefficient of −0.10 refers to a drop in the ranking by 10 percentiles.

Figure 3. Effect of Late Start and Early Grade Retention by Percentile.

Heterogeneity by Demographic and Country Characteristics

Additionally, interaction effects between retention and background characteristics are estimated. These interactions are instrumented with an interaction between the background variable and each of the two main instrumental variables. The most noteworthy results are reported in table 4. Overall, the imprecision of the interaction terms is high and few statistically significant effects are identified. All point estimates indicate that retention harms boys more than girls, but only the estimate for GRE with respect to math is statistically significant. Figure A4 in the online appendix shows percentile estimates

Table 4. Effect of Late Start (LS) and Early Grade Retention (GRE) by Demographic Characteristics

	Math	Science	Reading	Overall
LS	0.028	0.067	0.117	0.074
	(0.085)	(0.087)	(0.081)	(0.084)
LS*Male	−0.047	−0.077	−0.098	−0.078
	(0.098)	(0.100)	(0.094)	(0.098)
GRE	−0.216	−0.455***	−0.358**	−0.359**
	(0.167)	(0.170)	(0.161)	(0.167)
GRE*Male	−0.354**	−0.168	−0.160	−0.237
	(0.176)	(0.180)	(0.171)	(0.176)
LS	−0.0052	−0.0058	0.038	0.0095
	(0.056)	(0.057)	(0.054)	(0.056)
LS*LowEF	0.126	0.213	0.251	0.206
	(0.146)	(0.160)	(0.157)	(0.152)
GRE	−0.390***	−0.469***	−0.417***	−0.445***
	(0.151)	(0.157)	(0.149)	(0.153)
GRE*LowEF	−0.072	−0.280	−0.020	−0.130
	(0.211)	(0.224)	(0.216)	(0.215)
LS	0.0042	−0.00015	0.050	0.019
	(0.055)	(0.056)	(0.052)	(0.054)
LS*LowEM	−0.023	0.120	0.099	0.069
	(0.160)	(0.168)	(0.161)	(0.162)
GRE	−0.421***	−0.504***	−0.464***	−0.484***
	(0.146)	(0.151)	(0.144)	(0.148)
GRE*LowEM	0.092	−0.087	0.087	0.031
	(0.210)	(0.215)	(0.205)	(0.208)

Notes: The table reports the IV estimates of LS and GRE on achievement test scores, including interactions with gender and parental education. Interactions are instrumented using an interaction between the main instrument and the background characteristic. Test scores are standardized with a mean of 0 and a standard deviation of 1. LowEF (LowEM) is coded as 1 if the father (mother) of the student is low educated (ISCED 0–2), and 0 otherwise. Standard errors are in parentheses and are robust and corrected for clustering at the school level.

**Significant at the 5% level; ***significant at the 1% level.

separately for boys and girls. The figure shows that girls benefit more from delayed school entry low in the distribution, but boys still experience positive treatment effects well above the lowest decile. The differences are statistically significant for the lowest four percentiles (in favor of girls), and between the 15th and the 20th percentile (in favor of boys). With respect to GRE, the point estimates are generally more negative for boys, but none of the differences is statistically significant.

There is also suggestive evidence that LS has positive effects for those with low-educated fathers. The interaction is close to a 10 percent statistical significance level for reading. Moreover, percentile estimates show that, with respect to overall scores, the interaction terms are statistically significant between the third and 20th percentile. Interactions between GRE and low parental education provide statistically insignificant results throughout, and the same result is obtained for other indicators of disadvantage.

Heterogeneity is also assessed with respect to country characteristics, but a lack of statistical power prevents drawing strong conclusions. Point estimates indicate that treatment effects are more negative for countries with a high frequency of LS or GRE. This could occur because the local average student likely is located higher in the

Table 5. Effect of Late Start (LS) and Early Grade Retention (GRE) by Difficulty Level

	Not Reached Is Wrong		Not Reached Is Missing	
	LS	GRE	LS	GRE
Overall	0.018	−0.391***	0.027	−0.471***
	(0.052)	(0.134)	(0.055)	(0.166)
Easiest 25%	0.072	−0.374***	0.077	−0.503***
	(0.054)	(0.139)	(0.056)	(0.172)
25%–50%	0.026	−0.388***	0.042	−0.449***
	(0.053)	(0.136)	(0.057)	(0.171)
50%–75%	0.016	−0.222	0.0081	−0.370**
	(0.052)	(0.137)	(0.055)	(0.170)
Hardest 25%	−0.022	−0.454***	−0.020	−0.441**
	(0.053)	(0.145)	(0.056)	(0.179)

Notes: The table reports the effect of LS and GRE on achievement test scores (taking all test subjects together), for groups of questions that differ by level of difficulty. Difficulty is defined by the share of students with a correct answer. One point is given for each correct question answered and the sum of correct answers is divided by the number of total questions. A distinction is made between treating a question for which an answer is not reached by the student as either a wrong answer or a missing question. Scores are standardized with mean 0 and standard deviation 1. Standard errors are between parentheses and are robust and corrected for clustering at the school level.

**Significant at the 5% level; ***significant at the 1% level.

distribution in those countries. None of the interactions with country characteristics is statistically significant, however.

Effects by Difficulty Level

One main justification for retention is that it allows students to improve fundamental skills through a repeated curriculum before they move on to more advanced material. One might consequently expect that retention benefits basic knowledge at the expense of more advanced knowledge that has not yet been (extensively) discussed in the curriculum. By estimating treatment effects by difficulty level of the questions, I assess whether this pattern is apparent in practice as well.

Difficulty level is based on the fraction of students who answer a question correctly. Scores are calculated by giving one point for every correct answer. Results in table 5 are portrayed separately for when questions that are not reached by the student are treated as incorrect answers or as a missing question. The estimate for LS with respect to easy test questions is positive but statistically insignificant. The point estimates gradually become less favorable for more difficult questions but none of the estimates is statistically significantly different from 0. The estimates for the 25 percent easiest and the 25 percent hardest questions, however, are statistically significantly different from each other. GRE leads to substantial negative effects for both easy and difficult test questions. A statistically significant difference is identified when comparing the 25 percent hardest questions with those in the quarter below (50 to 75 percent), but none of the other estimates is statistically distinguishable.[19] Hence, primary school retention

19. This statistically significant difference is only identified for the method in which not reached questions are treated as incorrect.

is (also) ineffective at enhancing those fundamental skills that it aims to improve. In light of the theoretical framework presented in section 2, this either suggests that the self-productivity gains from improving more fundamental skills by repeating a grade are limited, or that these are negated by adverse impacts on other inputs.

Retention in Secondary School

The IV approach used in model 2 does not allow for estimation of the effect of secondary school retention (GRL) because of a lack of first-stage power. As an alternative, one can use the share of retention in the same cohort as an instrument, rather than the share of retention in the other two PISA cohorts. Nevertheless, given the relatively minor changes in the prevalence of GRL across birth months, simply having a bad draw of students in one particular cohort-country-month cell can lead to a severe bias. To mitigate this, such volatility can be smoothed out by predicting the probability of retention given month of birth assuming a strictly linear relation. To increase statistical power, this instrument is constructed by region-cohort cell rather than by country-cohort cell.[20] This alternative instrument provides sufficient first stage power to estimate the effect of GRL, although the measures are relatively imprecise. The results from this model are provided in table A1 in the online appendix. The estimate for GRL in relation to overall scores equals −0.543 and is statistically significant at the 5 percent level. Science scores are more severely affected by GRL than math or reading scores. These results should only be seen as indicative, as the model strictly relies on variation in the strength of the (linear) relation between month of birth and GRL across country-cohort cells, and any of the sensitivity tests that we apply to the main model is not feasible for this approach because of limited first stage power.

Additionally, model 2 is estimated using GRL as an outcome, to assess whether early retention affects the chance of retention at a later stage. A negative and statistically significant estimate is obtained for LS (–0.043) and a positive but statistically insignificant estimate for GRE (0.065).

6. ROBUSTNESS

Maturity Effects

Model 2 exploits variation in retention rates across month of birth, while simultaneously controlling for the effect of maturity at the time of test. The estimates for retention can be affected by how one controls for maturity. This subsection addresses how sensitive the results are to different assumptions and specifications with respect to the effect of maturity on achievement.

Homogeneity Assumption

The results presented in the main analysis rely on the assumption of a homogeneous maturity effect across countries (conditional on the starting age). It is known from previous research that maturity effects can differ across countries (Bedard and Dhuey 2006). This can lead to biased estimates. If the average maturity effect underestimates the maturity effect for some country C_A, the scores of repeat students in C_A (conditional on

20. PISA distinguishes regions for Belgium, Canada, Finland, Italy, Portugal, and Spain. There are a total of 61 regions in the sample.

their month of birth) are biased downward, given that repeat students are born more often in later months. This leads to a negative bias in the retention estimate. For a country C_B with a relatively low maturity effect, the bias will be in the opposite direction. The total bias depends on the number of repeat students in C_A relative to C_B. If countries with a high frequency of retention tend to have stronger relative age effects, the retention estimates will be negatively biased, and vice versa. Essentially, one wants countries with high retention levels to have more weight in the estimation of the maturity effect. Simply increasing the weights of countries with high retention frequencies would not solve the issue, however, as the weights would increase the difference in relative retention frequencies between the countries by the same degree.

I develop an iteration exercise in which the effect of maturity (again measured through the use of a linear and quadratic term of *PS*) on achievement is separately estimated from the effect of retention, in a way that ensures the negative and positive biases from assuming a homogeneous maturity effect cancel out. In a first step, the effect of maturity on achievement is estimated, while weighting observations by the total number of retained students in each country-cohort cell, without controlling for retention. A new score that corrects for this maturity effect is constructed, and model 2 is executed on the corrected score, absent the maturity control and without weighting.[21] Because not correcting for retention in step 1 leads to an overestimated maturity effect (assuming a negative effect of retention), the retention effect in step 2 is positively biased. The obtained estimate is still negative, however. Scores for retained students are then corrected with this (underestimated) coefficient, and a reweighted maturity effect is estimated. This estimate of the maturity effect is still overestimated, but less so than before, which in turn leads to a retention estimate that is also still positively biased but less so than before. If one repeats this exercise, the estimate converges to a point at which both the maturity effect and the retention estimates are stable. The obtained estimates are −0.042 (standard error of 0.045) for LS and −0.567 for GRE (standard error of 0.086). Hence, the assumption of homogeneous maturity effects (given the school starting age) does not strongly impact the estimates. The underlying reason for this minimal impact is that there is no correlation between a country's retention rate and the strength of its maturity effect, as also indicated by figure A1 in the online appendix.

The iteration exercise depends on the assumption that the retention effect is nonpositive, and also requires that there is no severe heterogeneity in treatment effects across countries. To assess sensitivity with respect to the latter assumption, the same iteration exercise is conducted while dropping specific countries from the sample. Table 1 shows that GRE is especially frequent in the Netherlands, Portugal, France, and Belgium. Dropping any of these countries for this exercise results in highly similar coefficients.[22]

Sensitivity to Specification

I further assess how sensitive the results are to alternative control functions for maturity, and also test sensitivity to alternatively defined instruments. Results are portrayed

21. The weights used in the main analysis still apply but are not multiplied with the retention frequency of the country-cohort cell, as in the previous step of the iteration exercise.
22. These estimates range from −0.473 to −0.598. The corresponding range for LS is −0.045 to −0.053.

Table 6. Sensitivity Analysis

	LS	GRE
Panel A: Maturity		
Main model	0.027	−0.486***
	(0.050)	(0.136)
Month	0.013	−0.593***
	(0.051)	(0.125)
Month dummies	0.011	−0.578***
	(0.058)	(0.132)
PISA-specific	0.016	−0.485***
	(0.051)	(0.135)
Region-specific	0.021	−0.648***
	(0.056)	(0.260)
Country-specific	−0.037	—
	(0.110)	
Panel B: Additional		
GRL control	−0.0017	−0.443***
	(0.050)	(0.132)
LS only	0.037	—
	(0.051)	
GRE only	—	−0.489***
		(0.135)
Mean month	0.022	−0.552***
	(0.048)	(0.098)
Mean quarter	0.092	−0.326***
	(0.062)	(0.127)
No country weights	0.068	−0.446***
	(0.062)	(0.166)

Notes: The table reports estimates of the effect of retention on school achievement under alternative specifications. In panel A, different control functions for the effect of maturity on achievement are used: a linear birth month effect, dummy variables for all birth months, interactions between the predicted starting age (*PS*) and PISA cohort, interactions between *PS* and geographical region (Scandinavia, southern Europe, eastern Europe, western Europe and Anglo-Saxon countries), and interactions between *PS* and every country dummy. Main model refers to model 2. Panel B includes an additional control for secondary school retention (GRL), estimates both treatment effects separately, uses alternative instruments or uses different weights. Mean month uses the mean share of retention for every country-month of birth-cohort cell. Mean quarter uses the mean share of retention for every country-quarter of birth-cohort cell. No country weights uses the inverse sampling probability as weights, without additionally weighting countries. Overall scores are used as an outcome in all instances. Standard errors are in parentheses and are robust and corrected for clustering at the school level.

*** Significant at the 1% level.

in table 6. The table shows that alternative control functions, such as a strictly linear function or dummies for each birth month, induce very little variation in either estimate.[23] The table further shows that results are similar when the control function is

23. A range of alternatives is tested (including higher-order polynomials); the additionally included terms have very low estimates and therefore lead to near-identical retention estimates. Marked changes only occur for control functions that clearly fit the data poorly (such as including only a higher-order polynomial).

allowed to vary across cohorts or by geographical region. Allowing for country-specific maturity effects is only feasible (in terms of first stage power) for estimation of LS and leads to a statistically insignificant estimate of −0.037.

Panel B of table 6 further assesses sensitivity to the exact specification of model 2. Including GRL as an additional control has only a minor effect on the estimated effects.[24] Estimating LS and GRE in separate models also leads to very similar results as in the main specification. The alternative instruments used in panel B are still based on date of birth but differ slightly in their exact construction. The first entry measures the prevalence of retention in the same cohort-country-month cell, rather than the average of the country-month cells in the other two PISA cohorts. This induces a negative bias in the GRE estimate, because this instrument is influenced by one's own retention status and because results can be affected by a bad draw of students in a particular cohort-country-month cell—although this bias is shown to be small. A similar result is obtained when the instrument is calculated by quarter of birth rather than by month of birth. Finally, table 6 shows that results are similar when observations are not weighted on a country level (inverse sampling probability weights still apply). I also conduct a "jackknife" exercise that leaves out each country in turn, to assess the sensitivity of the estimates with respect to the sample composition. The coefficients produced by this exercise range from −0.0047 to 0.085 for LS and −0.367 to −0.624 for GRE. When specific PISA cohorts are dropped, the obtained range of coefficients is −0.0033 to 0.070 for LS and −0.400 to −0.620 for GRE.

Effect of Retention on Self-esteem and Parental Support

As specified in section 2, the long-run effects of retention can potentially be affected by how retention affects future investments or inputs in the learning process, other than through a direct effect on the specified curriculum. In this subsection, potential mechanisms in these areas are explored.[25]

Using PISA data from parental questionnaires, we find that LS has a negative effect on the reported educational expenses that parents make with respect to the participating child and on the frequency of helping students with homework. One should interpret these results with caution, as the differences can also be related to being in a lower grade. They could be seen as suggestive evidence that parents devote fewer resources and less attention to children who have delayed school starts, which could mitigate potential positive effects of delayed school entry from other sources. These effects could not be estimated for GRE because these parental data are only available for a limited part of the sample, which strongly decreases first-stage power.

Additionally, it has been argued that retention impacts students' self-esteem and attitudes toward school.[26] PISA data are used to construct several such measures, using factor analysis. The majority of coefficients in table A2 (see the online appendix) are

24. One potential concern with excluding GRL in the main estimation is that this could bias the estimation of the maturity effect and thereby also of the other retention estimates. The results from table 6 show this issue has no strong impact on the estimates, and also indicate that the importance of GRL as a channel for the effects of earlier retention is limited.
25. The results of this analysis are reported in table A2 in the online appendix.
26. A meta-analysis by Holmes and Matthews (1984) finds a small but statistically significant negative effect from retention on self-esteem and attitude toward school, averaged across nine different studies. Alexander, Entwisle, and Dauber (2002) find a positive effect of retention on self-esteem.

low and statistically insignificant. LS has a positive effect on students' attitudes and work ethic toward mathematics.[27] The former measures students' enjoyment, interest, and perceived value of mathematics as a subject and the latter measures how hard they work on homework and exams for math, as well as how well they pay attention in class. Such effects might stem from increases in the relative ranking in class that is associated with delayed school entry. In relation to the theoretical model, this suggests that delayed school entry can possibly provide benefits through higher future effort levels. The estimates for GRE are all statistically insignificant, although the (positive) estimate with respect to math attitudes is close to a 10 percent statistical significance threshold.

Taken together, these estimates are somewhat in favor of LS relative to GRE, although they are not consistently positive. A typical problem in assessing these effects is that many potentially relevant inputs are not being observed or are measured with substantial error. As such, it remains inconclusive to what extent these effects can explain retention effects, which provides a potential avenue for future research.

Exogeneity of Month of Birth

Individual-level Characteristics

Recent research challenges the exogeneity of month of birth, arguing that it is correlated with indicators of socioeconomic status.[28] I analyze the correlation between month of birth (relative to the cutoff) and the demographic indicators that are part of the control vector X'.

The exercise shows there is no statistically significant relationship between month of birth and ethnicity, language spoken at home, gender, living in a single-parent family, or whether the father works full-time. There are statistically significant but small correlations between month of birth and maternal education, whether the mother works full-time, and home possessions. Several of the cohort-specific and virtually all of the country-specific correlations are not statistically significant, and the sign of the correlation strongly differs across countries. The size of these effects is modest in all cases; the correlation with mother's education is relatively largest and suggests that moving from the lowest to the highest category is associated with a birth date that lies 0.6 days earlier. Hence, there is some evidence of a relationship between social status and month of birth, but it is very weak—only statistically significant in very large samples and inconsistent by year, country, and indicator. Additionally, the main model controls for a (homogeneous) maturity effect, which likely captures at least part of this relationship. Placebo tests are conducted in which control variables are plugged in as outcome variables in model 2. All placebo tests provide statistically insignificant estimates for both LS and GRE.

These findings are already implicitly reflected by the small impact of adding control variables on the estimates for retention. Buckles and Hungerman (2013) show that including controls can severely affect estimates when month of birth is used as an instrument for years of schooling. The contrast with the results of this study is mainly

27. Math was the focal point for both PISA 2003 and PISA 2012, and therefore an extensive set of questions is directed toward math attitudes in both cohorts.
28. See, for example, Buckles and Hungerman (2013).

because month of birth is a markedly weaker instrument in those settings and therefore even a minor association with other determinants of the outcome can have strong effects. The results of this study are also not sensitive to dropping certain birth months or birth quarters from the sample (provided it is not the complete last quarter of birth, on which the instrumentation strongly relies).

School-level Characteristics

Additionally, a student's birth month could affect the quality of the school she is selected into. This subsection analyzes the relation between month of birth and school quality indicators. The nature of the analysis is different than for demographic characteristics. The variables in the vector X' are very unlikely to be influenced by the retention status of the child, but school-level characteristics can act both as a possible identification threat and as a potential mechanism. For example, school quality can differ because retention directly affects achievement and thereby student sorting across schools, but it can also differ because a low relative age can affect student sorting net of retention effects. The latter example can lead to a bias in the results, if the effect is not accurately captured by the maturity control. On the other hand, the former example reflects a genuine part of the effect of retention.

When plugging in school-level characteristics as outcome variables in model 2, a positive and statistically significant relationship is identified with respect to the share of certified teachers in the school (LS), public schools (LS) and student–teacher ratios (LS and GRE). Statistically insignificant estimates are identified for school enrollment, urbanization level of the school, and share of teachers with a tertiary education degree. Additionally, the model estimates a positive and (marginally) statistically significant relationship between LS and the chance of being in an academic track. With the current data, it is not possible to determine whether these relations are directly due to relative age or indirectly due to retention. Additionally, some of these differences could occur because of grade effects, in particular with respect to student–teacher ratios. Nonetheless, sensitivity to a potential bias from school sorting is assessed by including school quality indicators as additional controls in model 2. When adding all school-level characteristics to the main estimation (allowing for country-specific track effects), the estimates for LS and GRE change from 0.027 and −0.486 to 0.017 and −0.453, respectively. Hence, even if one believes that these differences in school quality are completely net of retention and grade effects, the size of such a bias appears very marginal at best.[29]

As stated above, some of these differences could be part of the overall effect of retention. In particular, the identified effect on the chance of being in an academic track could potentially be a mechanism that can explain part of the difference between the estimated effect for LS versus GRE. The effect size is rather limited (0.015), however, hence it is unlikely to explain a large share of the difference. The positive sign of the estimate is likely because decisions with respect to tracks are generally based on achievement relative to the grade of the student rather than relative to the age of the student. As argued before, the results imply that LS has a favorable effect in same-grade comparisons.

29. Additional analysis using TIMSS data for the same set of countries indicates that the relationship between month of birth and *primary* school characteristics is weak as well.

Measurement Error

Retention variables are based on self-reported information of students, which can suffer from measurement error. The main approach labels all those who are a grade behind and do not report repeating a grade as delayed school entrants. This approach might mistakenly label those who repeated during formal education but failed to report this as delayed entrants. Alternatively, some students might mistakenly interpret retention during kindergarten as retention during ISCED 1. The main estimation results can be biased if there is a specific type of repeat student who is especially likely to misreport. A comparison between retention rates in the PISA data and retention rates from administrative data shows highly similar average rates for LS and GRE (see table A3 in the online appendix). Still, some country-level differences exist and it is also possible that different types of misreporting cancel out on average. I use alternative data from PISA on self-reported school starting ages to assess sensitivity toward measurement error.

The school starting age question in PISA asks students at what age they started ISCED 1. The answers are clearly subject to measurement error as well, as it appears that students have a tendency to round up their age if their birthday lies closely after the school starting date. Nonetheless, the variable can be used in a sensitivity analysis, assuming that at least a substantial share of (potentially) mislabelled delayed school entrants in the main analysis report correct starting ages. In a first exercise, LS-students that report "on-time" starting ages are classified as GRE-students instead.[30] In a second exercise, those who report retention during ISCED 1 but report "late" starting ages are classified as LS-students instead of GRE-students. This sensitivity exercise is rather strong, as it redistributes roughly 15 percent of all repeat students. This analysis provides coefficients of 0.096 and −0.471 for LS and GRE in the first exercise and −0.037 and 0.366 for LS and GRE, respectively, in the second exercise. In a final exercise, only those repeat students for whom the starting age deduced from the retention data matches the reported starting age are included in the analysis. This produces estimates of 0.060 and −0.467 for LS and GRE, respectively. Hence, although using self-reported retention data is a limitation and attenuation bias cannot be ruled out completely, it appears very unlikely that this has a strong impact on the main results.

Curriculum Bias

To further assess to what extent the results can be influenced by what is recently discussed in the curriculum, retention effects are estimated with respect to the performance of students on a problem-solving test that was taken for PISA 2003 and PISA 2012. The test asks students how to design the quickest or cheapest way between two subway stations, which gates have to be opened to let water flow from A to B, and so forth. Any strong influence of the exact curriculum in the current grade is very unlikely. The estimated effect of retention on achievement in this test is very similar to the estimates identified for the other test subjects (see table A4 in the online appendix). Hence, similar retention effects are identified for a test for which the influence of

30. For students for whom it cannot be identified whether their birth date is before or after the school starting date (e.g., the birth month is August and the country has a flexible school starting date somewhere within the month of August) LS and GRE are defined as in the main analysis.

specific curricular knowledge should be very minimal. It should still be acknowledged that promoted students are exposed to the curriculum of an additional grade. In absence of vertically scaled scores, it is not possible to completely narrow down how this affects the results. However, as stated before, one can also perceive this as an essential part of the treatment effect of retention.

7. CONCLUSION

In this paper, I present evidence on the effect of age-based retention in kindergarten and in primary school on school achievement at age fifteen, using international data from PISA. The IV model exploits the fact that differences in retention prevalence across birth months are not necessarily proportional to differences in achievement across birth months, as a young relative age in class is given weight in retention decisions over and above its relation to achievement. The analysis identifies a low and statistically insignificant effect of delayed school entry and a negative effect of primary school retention. The negative estimates for primary school retention are severe; they are close to what PISA reports as a grade equivalent difference in test scores. Heterogeneity analysis indicates that delayed school entry can have positive effects for the very lowest percentiles of the achievement distribution (especially for girls), while the impact of primary school retention is negative throughout. Retention in primary school harms performance on questions across all difficulty levels.

The identified effects are considerably more negative than the findings of recent studies that use regression discontinuity designs (if one compares the results for primary school retention with the third grade retention results from those studies). It is difficult to assess to what extent this difference is driven by a difference in institutional setting, as the studied policies in Chicago and Florida also involved, for example, summer schools, instructional support, and better-quality teachers. Nonetheless, a crucial distinction is that those studies estimate effects for the marginal student at the achievement threshold for promotion, whereas this study estimates a LATE for students that are being retained because they were born late in the year. As such, the results indicate that putting emphasis on relative age in class for retention decisions in primary school is harmful for the achievement of students who are retained as a result. The students for whom the effect is estimated are not necessarily lacking in innate ability. The findings can also be seen in relation to literature on month of birth effects. Studies have found that achievement differences by relative age are high early in primary school but reduce in later years. This indicates that relatively young students learn relatively fast in the years after school entry. Letting them repeat a grade might hinder that process of catching up. In relation to the theoretical framework of this study, this is suggestive evidence that relatively younger students especially benefit from the promoted curriculum I_{t+1}^{g+1} rather than the repeated curriculum I_{t+1}^{g}. Additionally, the negative findings could suggest that the effect of retention worsens as students grow older. Previous research generally identifies that retention effects are initially positive, but rapidly decline or become negative over time. The age at which students are tested in this study is high compared with other studies. As such, the high effect sizes can also be partly the result of primary school retention effects worsening through secondary education.

The large difference between the estimates for delayed school entry and the estimates for retention in primary school in this study is in line with previous literature that identifies more harmful effects of grade retention in grade 6 versus grade 3 (Jacob and Lefgren 2004) and estimates severe negative effects of retention in junior high (Manacorda 2012). It also fits with the theoretical framework and research on early childhood investments, as the potential benefits of the self-productivity of skills are higher for interventions that take place earlier in life (Cunha and Heckman 2008). Also, it is possible that retention during kindergarten and retention during primary school have a different impact on future inputs and effort levels. I find some evidence that delayed school entry improves student self-esteem and work ethic, but more evidence is needed to make more conclusive statements on the potential importance of such mechanisms in relation to retention. It should be emphasized that the difference in the estimated treatment effects of delayed school entry versus primary school retention could also be due to differences in the characteristics of students who start school late and students who repeat a grade in primary school. As such, the results do not automatically imply that early retention is relatively better for a given student but rather tell us that students for whom a low relative age in class induces primary school retention are relatively worse off than students for whom a low relative age in class induces delayed school entry.

This study does not focus on changes in peer effects for those who are not being retained, as well as the possible effect of the sole threat of retention on nonretained students. Empirical evidence on the effect of retention on the untreated is very scarce; a rare exception is provided by Babcock and Bedard (2011) who find that higher statewide retention rates are related to higher hourly wages. Assessing such effects also for school achievement is an interesting topic for future research. Additionally, in light of the theoretical framework of this study, it would be very valuable if robust estimation approaches can be combined with yearly longitudinal data that also contain rich information on school and parental inputs, so that the exact mechanisms behind retention effects can be uncovered.

ACKNOWLEDGMENTS

I thank Lex Borghans, Bart Golsteyn, Gabriele Marconi, Tyas Prevoo, the participants of the EALE 2014 conference in Ljubljana, seminar participants at Maastricht University and KU Leuven, and two anonymous referees for their valuable comments.

REFERENCES

Abadie, Alberto, Joshua D. Angrist, and Guido Imbens. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70(1):91–117. doi:10.1111/1468-0262.00270.

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber. 2002. *On the success of failure: A reassessment of the effects of retention in the primary grades*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511500091.

Babcock, Philip, and Kelly Bedard. 2011. The wages of failure: New evidence on school retention and long-run outcomes. *Education Finance and Policy* 6(3):293–322. doi:10.1162/EDFP_a_00037.

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

Borghans, Lex, and Ron Diris. 2014. *An economic analysis of the optimal school starting age*. Available http://papers.ssrn.com/abstract=2022360. Accessed 24 August 2016.

Buckles, Kasey, and Daniel M. Hungerman. 2013. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* 95(3):711–724. doi:10.1162/REST_a_00314.

Chernozhukov, Victor, and Christian Hansen. 2005. An IV model of quantile treatment effects. *Econometrica* 73(1):245–261. doi:10.1111/j.1468-0262.2005.00570.x.

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

Crawford, Claire, Lorraine Dearden, and Costas Meghir. 2010. *When you are born matters: The impact of date of birth on child cognitive outcomes in England*. Available www.ifs.org.uk/docs/born_matters_summary.pdf. Accessed 24 August 2016.

Cunha, Flavio, and James J. Heckman. 2007. The technology of skill formation. *American Economic Review* 97(2):31–47. doi:10.1257/aer.97.2.31.

Cunha, Flavio, and James J. Heckman. 2008. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* 43(4):738–782.

Cunha, Flavio, James J. Heckman, Lance J. Lochner, and Dimitriy V. Masterov. 2006. Interpreting the evidence on life cycle skill formation. In *Handbook of the economics of education*, edited by Eric A. Hanushek and Frank Welch, pp. 697–812. Amsterdam: North-Holland.

Eide, Eric R., and Mark H. Showalter. 2001. The effect of grade retention on educational and labor market outcomes. *Economics of Education Review* 20(6):563-L-576. doi:10.1016/S0272-7757(00)00041-8.

Garcia-Perez, J. Ignacio, Marisa Hidalgo-Hidalgo, and J. Antonio Robles-Zurita. 2011. Does grade retention affect achievement? Some evidence from PISA. *Applied Economics* 46(12):1373–1392. doi:10.1080/00036846.2013.872761.

Greene, Jay P., and Marcus A. Winters. 2007. Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy* 2(4):319–340. doi:10.1162/edfp.2007.2.4.319.

Greene, Jay P., and Marcus A. Winters. 2012. The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy* 7(3):305–330. doi:10.1162/EDFP_a_00069.

Grissom, James B., and Lorrie A. Shepard. 1989. Repeating and dropping out of school. In *Flunking grades: Research and policies on retention*, edited by Lorrie A. Shepard and Mary Lee Smith, pp. 34–63. New York: The Falmer Press.

Holmes, Thomas C., and Kenneth M. Matthews. 1984. The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research* 54(2):297–315. doi:10.3102/00346543054002225.

Imbens, Guido W., and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–475. doi:10.2307/2951620.

Jacob, Brian A., and Lars Lefgren. 2004. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics* 86(1):226–244. doi:10.1162/003465304323023778.

Jacob, Brian A., and Lars Lefgren. 2009. The effect of grade retention on high school completion. *American Economic Journal. Applied Economics* 1(3):33–58. doi:10.1257/app.1.3.33.

Jimerson, Shane R., Gabriella E. Anderson, and Angela D. Whipple. 2002. Winning the battle and losing the war: Examining the relation between grade retention and dropping out of high school. *Psychology in the Schools* 39(4):441–457. doi:10.1002/pits.10046.

Koenker, Roger, and Gilbert Bassett, Jr. 1978. Regression quantiles. *Econometrica* 46(1):33–50. doi:10.2307/1913643.

Manacorda, Marco. 2012. The cost of grade retention. *Review of Economics and Statistics* 94(2):596–606. doi:10.1162/REST_a_00165.

Organization for Economic Co-operation and Development (OECD). 2004. *Learning for tomorrow's world: First results from PISA 2003*. Available www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34002216.pdf. Accessed 24 August 2016.

Organization for Economic Co-operation and Development (OECD). 2009. *PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes (volume II)*. Available www.oecd.org/pisa/pisaproducts/48852584.pdf. Accessed 6 September 2016.

Organization for Economic Co-operation and Development (OECD). 2013. *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science (volume I)*. Available www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf. Accessed 6 September 2016.

Özek, Umut. 2015. Hold back to move forward? Early grade retention and student misbehavior. *Education Finance and Policy* 10(3):350–377. doi:10.1162/EDFP_a_00166.

Roderick, Melissa, and Jenny Nagaoka. 2005. Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis* 27(4):309–340. doi:10.3102/01623737027004309.

Schwerdt, Guido, and Martin R. West. 2013. The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. IZA Discussion Paper No. 7314.

Shepard, Lorrie A., and Mary Lee Smith. 1986. Synthesis of research of school readiness and kindergarten retention. *Educational Leadership* 44(3):78–86.

Tomchin, Ellen M., and James C. Impara. 1992. Unraveling teachers' beliefs about grade retention. *Educational Research Journal* 29(1):199–223.