Master Thesis Economics and Business (Policy Economics)

The Causal Impact of Grade Retention on Academic Achievement

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Department of Economics

Supervisor: Prof. dr. Dinand Webbink

Name: Sonny Kuijpers, BSc

Exam number: 345050

E-mail address: sonnykuijpers@gmail.com

Table of contents

Summary

Grade retention is still one of the most hotly debated topics in education, due to a lack of conclusive and credible evidence on its effects. This paper tries to find the causal impact of early grade retention on academic achievement in elementary school by employing a novel quasi-experimental strategy. This identification strategy is based on the empirical observation that the relationship between one's probability of being retained and his or her relative age in class appears to be nonlinear in some countries. This nonlinearity is argued to provide exogenous variation in the prevalence of grade retention, which is exploited by employing a difference-in-differences and instrumental variables approach. For this purpose, data on an international standardized math and science test (TIMSS 1995) are used.

The results are mixed across the countries studied. In Canada and Hong Kong, retained pupils perform significantly worse than their promoted age peers. The effect sizes of these estimates are more negative than in most other studies. The nature of the TIMSS test and, most notably, the potential violation of one or more of the IV assumptions can be seen as explanations for these results. For Korea and Norway, on the other hand, no significant relationship was found in any of the IV specifications. This suggests that the efficacy of grade retention as a method to address poor performance depends on the presence of additional country-specific educational policies to accommodate the retention policy. Policy makers should therefore consider other interventions when trying to remediate poor performance in elementary school.

1. Introduction

Grade retention, the practice of requiring a student who has been in a given grade level for a full school year to remain at that level for a subsequent school year (Jackson, 1975), is one of the most hotly debated topics in education. Despite the fact that it has been in place as an educational intervention for low-achieving pupils since the beginning of the 20th century, with fluctuating popularity over the years (Allen et al., 2009), there is still no consensus on its desirability in many countries as of today.

In a recent study ordered by the Dutch Ministry of Education, Culture and Science, for example, Driessen et al. (2014) note that grade retention is more common in the Netherlands than in most other OECD countries. At the same time, however, they mention that the Dutch Ministry is currently looking for effective ways to prevent grade retention, because there are questions regarding its efficacy.

One of the reasons why there is still a lot of uncertainty regarding the desirability of retaining poorly performing children is the fact that there is a lack of conclusive evidence on the effects of grade retention. Moreover, most studies only focus on the United States. Finally, and perhaps most importantly, many of the empirical studies suffer from methodological limitations.

While most "old" studies showed that grade retention has negative effects on the performance of pupils (see Jackson [1975] and Holmes and Matthews [1984], for example), Allen et al. (2009) suggest that many of these papers employed poor methodological controls for differences between retained and non-retained children. More recent papers try to overcome the limitations of the previous studies by employing quasi-experimental research designs. Jacob and Lefgren (2004) and Schwerdt and West (2013), among others, find less negative effects of grade retention in the United States. Still, these results are subject to uncertainties.

By using data on an international standardized math and science test, this paper tries to find the causal impact of early grade retention on academic achievement in elementary school in other, mostly European, countries. For this purpose, a novel strategy is used, based on relative age effects. In some countries, there appears to be a nonlinear relationship between one's relative age in class and his or her probability of being retained, as the youngest

children in a classroom are disproportionately more likely to be retained. This nonlinearity is assumed to be exogenous (i.e. not due to a nonlinearity in performance), and can hence be exploited by employing a difference-in-differences and an instrumental variables approach. The performance of retained children is compared with that of promoted children at the same age. As explained in the next section, it is then hypothesized that being retained has negative effects on academic achievement.

As will be shown below, this paper finds mixed results on the effects of grade retention. Some sizeable negative effects are found, while there is no significant relationship in other countries. The presence of additional country-specific educational policies seems to be a driving force behind this finding. The nature of the standardized test used in this paper and questions regarding the validity of the assumptions behind the identification strategy could explain the observation that some of the results deviate from those in most other empirical studies.

The remainder of this paper is organized as follows. The next section reviews the existing literature in more detail. Then, the identification strategy will be explained in Section 3. The dataset will be described in the fourth section. Afterwards, the results will be discussed in detail. Finally, the paper ends with a conclusion and discussion section, based on the results.

2. Literature review

As mentioned by Holmes (1989), each study on the impact of grade retention can be classified into one of two categories. The first set of studies compares the outcomes between retained and non-retained pupils while they are in the same grade. Other studies, however, compare the outcomes between retained and non-retained pupils that have the same age. It is important to distinguish these two categories, because both types of studies tend to generate different results. Theoretically, the studies that use same-grade comparisons may be more likely to find positive effects of grade retention, because the retained pupils are older and have received more schooling than their promoted classmates. Same-age studies, on the other hand, tend to focus on the negative consequences of retention, because retained pupils are in a lower grade and hence "miss out" on a more advanced year of schooling. This is explained more formally in Schwerdt and West (2013).

Over the years, Jackson (1975), Holmes and Matthews (1984), Holmes (1989), and Jimerson (2001), amongst others, have conducted meta-analyses, referring to hundreds of grade retention studies. They all conclude that grade retention appears to be an inferior method of remediating poor performance, because a vast majority of studies finds that retained pupils underperform their promoted age peers (in same-age studies) or do not outperform their classmates that had been promoted (in same-grade studies).

Allen et al. (2009: 493), on the other hand, note that "studies that do a better job of removing the effect of preretention differences on achievement yield a less negative picture of the effects of grade retention". As will be explained in the next section, it is very important to adequately control for differences between retained and non-retained pupils in order to find the causal impact of grade retention on pupil performance. Therefore, the remainder of this section will focus on studies that use a (quasi-)experimental design.

Johnson et al. (1990), for example, compare the outcomes on a set of standardized tests (the Metropolitan Achievement Tests) between fourth-grade children that had been retained at the K-1 level and those that were nominated for retention at the K-1 level but were promoted instead. They find no significant differences in achievement between the two groups, and hence conclude that early grade retention is not effective as an academic intervention. This study has at least two limitations, however. First of all, its sample size is

modest. The study compares only 20 retained pupils with 17 pupils in the "recommended-for-retention-but-not-retained" group. Secondly, it is not demonstrated whether the two groups of pupils are indeed comparable, and the author only controls for gender differences.

Schwerdt and West (2013) is an example of a same-age study. They exploit a discontinuity in the probability of being retained in Florida using a regression discontinuity design (RDD). The performance of students that had been retained in grade 3 because they had a reading score just below the cutoff is compared with the performance of students that scored just above this cutoff and hence had not been retained. The authors find that retained pupils score significantly higher on math and reading tests than their non-retained age peers in the first years following retention. This seemingly surprising result may be partially explained by the fact that retained pupils received supplemental interventions during the retention year. For example, they were assigned to smaller classes and a "high-performing" teacher.

Jacob and Lefgren (2004) also employ an RDD design. They exploit a Chicago Public Schools accountability policy that tied retention decisions in third and sixth grade to performance on the Iowa Test of Basic Skills (ITBS). It is found that pupils that had been retained in grade 3 perform significantly better than their promoted age peers in both math and reading on the ITBS in the first year following retention. However, as mentioned by the authors, this may be a result of the fact that the ITBS is a high-stakes test for third grade pupils, while performing poorly on the test has no direct consequences for fourth grade pupils. Thus, the results may be driven by incentive effects.

One similarity of this paper with the three studies mentioned above is its focus on the short-term effects of grade retention. Like Schwerdt and West (2013) and Jacob and Lefgren (2004), this paper compares the performance between retained and non-retained pupils that have the same age. A major difference between this study and most other empirical studies is the fact that this paper examines the effects of grade retention in other countries than the United States, as mentioned above. Contrary to most studies, this study takes a general approach by examining nationally representative samples of pupils, instead of focusing on a specific state or school district. Finally, unlike any other study to date, it tries to exploit exogenous variation in grade retention due to age effects, as will be explained below.

3. Identification strategy

In most empirical studies, the effect of grade retention on pupil performance is estimated by the following regression model:

(1) $$T_{ijk} = \beta_0 + \beta_1 R_{ijk} + \psi S_{jk} + \tau X_{ijk} + \varepsilon_{ijk} \ ,$$

where T_{ij} is a measure of the performance of child *i* (who goes to school *j* and lives in country *k*), R_{ijk} is a dummy variable indicating whether the child has been retained, S_{jk} is a vector of school fixed effects, and X_{ijk} is a vector of child-specific control variables. In this specification, however, the coefficient of interest, β_1, is unlikely to represent the causal impact of grade retention. This will be explained below.

In order to find the causal impact of a treatment (here: grade retention) on some outcome variable (here: pupil performance) through an OLS regression, it is necessary that the treatment variable is exogenous, i.e. treatment is assigned randomly. If that is the case, the treatment group will be very similar to the control group, on average, so a simple comparison between the two groups yields the causal effect of the treatment.

Pupils that have been retained are very likely to be different from children that are "on time", both in observable and unobservable ways, however. One can control for the observable differences (such as age and gender) through the vector X_{ijk} and for (un)observable differences at the school level through the vector S_{jk}, but not for the unobservable characteristics of individual pupils. Yet, grade retention is very likely to be affected by these unobservable characteristics, such as motivation and ability. Children that have been retained are probably less able and less motivated than average, so they have self-selected themselves into treatment (that is, treatment is not assigned randomly). Since motivation and ability are also important determinants of performance, retained pupils are likely to perform worse than on-time pupils, *even if they are not retained*. This is called selection bias. Assuming, for the moment, that grade retention has a negative impact on pupil performance, it means that estimating a "naïve" OLS equation probably leads to an overestimation of the "true" negative impact of grade retention (that is, the absolute value of the coefficient will be too high). More formally, it means that the error term, ε_{ijk}, is correlated with the treatment variable, R_{ijk}, so there is endogeneity. Consequently, the OLS regression will generate biased coefficients.

This paper takes a different approach, focusing on an important concept: relative age. As shown in Bedard and Dhuey (2006), age differences within a classroom can have significant effects on pupil performance in elementary school. In some countries, the oldest fourth grade children score up to twelve percentiles higher than their youngest classmates. The presence of age differences in a classroom is a result of the fact that most education systems have a single cutoff date for school eligibility. Children born just before this cutoff date are then about a year younger than their oldest classmates, that were born just after the cutoff date a year earlier. Especially in the first couple of years in elementary school, this age difference reflects a substantial difference in maturity, which is likely to explain the significant difference in performance between the youngest and oldest pupils.

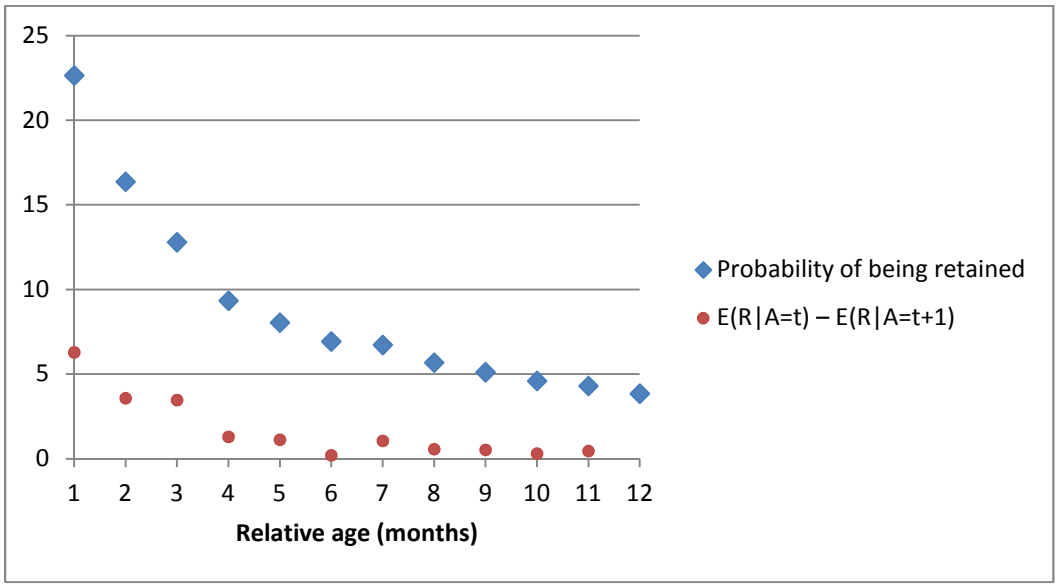
Bedard and Dhuey (2006) mention that younger pupils are also more likely to be retained. This is not surprising: these pupils tend to perform worse and are therefore more likely to be retained if grade retention is (at least partially) determined by pupil performance. If one assumes that there is a linear relationship between pupil performance and relative age (which means that for every one month increase in relative age, the pupil's performance improves by the same amount) and that the decision whether or not to retain a pupil is solely based on his or her achievement in class, then one would also expect the relationship between grade retention and relative age to be linear.

In this paper, the relative age measure is constructed as follows. The youngest children, who were born in the last month before the cutoff date, have relative age $A = 1$. The oldest children, who were born in the first month after the cutoff date a year earlier, have relative age $A = 12$. The measure thus indicates how many months the child was born before the cutoff date.

Figure 1 plots the probability of being retained and its change against relative age for a sample of eight countries (being Austria, Canada, Cyprus, Hong Kong, Korea, the Netherlands, Norway, and Portugal). It indicates that, contrary to what one might expect, there is no linear relationship between grade retention and relative age for this sample. Between $A = 4$ and $A = 12$, the relationship between the probability of being retained and relative age is approximately linear, suggesting a constant "maturity effect". For the youngest pupils, however, the pattern is different. Taking the trend between $A = 4$ and

A = 12 as the baseline, it can easily be seen that the youngest pupils (i.e. the pupils born in the last three months before the cutoff date) are disproportionately more likely to be retained.

Figure 1: Grade retention (pooled sample of countries)



There are several potential explanations for this observation. One could argue, for example, that there is an informal "rule" that specifies that in each classroom a certain fraction of pupils should be retained (on average). In that case, teachers may find it easier to retain the youngest pupils, because they can use their low relative age as an argument to justify this decision to the parents of the pupils. In the same vein, it may be the case that teachers also sincerely take psychological factors into account when deciding which pupils to retain. If they observe that the youngest pupils are more often bullied, or predict that they are more likely to be bullied in the future, for example, they may more easily decide to retain these pupils in order to "protect" them. Combined with the fact that the youngest children naturally perform worse on average than their older classmates, because of the maturity effect, they may indeed be disproportionately more likely to be retained.

It may also be the case that the youngest pupils perform disproportionately worse in class. In this paper, however, it is assumed that the maturity effect is constant for all children, even for the youngest ones. Consequently, the discontinuity in the probability of being retained for the youngest pupils should be uncorrelated with any (unobservable) factors that influence pupil performance. This also means that any nonlinearity in pupil performance in

the youngest age group is assumed to be due to grade retention. Intuitively, one can see a pupil's relative age as a random variable, as it is determined by a combination of the relevant cutoff date and his or her birth date[1]. As a consequence, the youngest children in a classroom should be inherently very similar to their older classmates. Their higher-than-expected probability of being retained is a result of "bad luck", i.e. randomness. This means that belonging to the youngest relative age group can be seen as a source of exogenous variation in grade retention, which can be useful in establishing the causal impact of grade retention on pupil performance. The empirical strategies will be described more formally below.

Difference-in-differences

The first possible way to exploit the nonlinearity in grade retention is to conduct a difference-in-differences (DID) analysis. In general, a DID equation compares a treatment group ($P = 1$) with a control group ($P = 0$) at two points in time. At the first point in time (at $t = 0$), both groups are not treated. At $t = 1$, only the treatment group is treated. As the control group is not treated in both periods, the change in the outcome variable for this group can be seen as the baseline change. The change in the outcome variable for the treatment group can then be compared with this baseline change: the difference can be seen as an approximation of the causal treatment effect.

The DID method relies on one major assumption. It is only valid if the observed change between $t = 0$ and $t = 1$ for $P = 0$ is indeed equal to the (hypothetical) change in the outcome variable for the treatment group that would have been observed, had the treatment group not been treated. This assumption is called the parallel trend assumption.

This paper uses a cross-section dataset, which means that there is only one observation for each individual. However, because of the assumption of a constant maturity effect, the econometric framework of the DID method can still be used. This will be explained below.

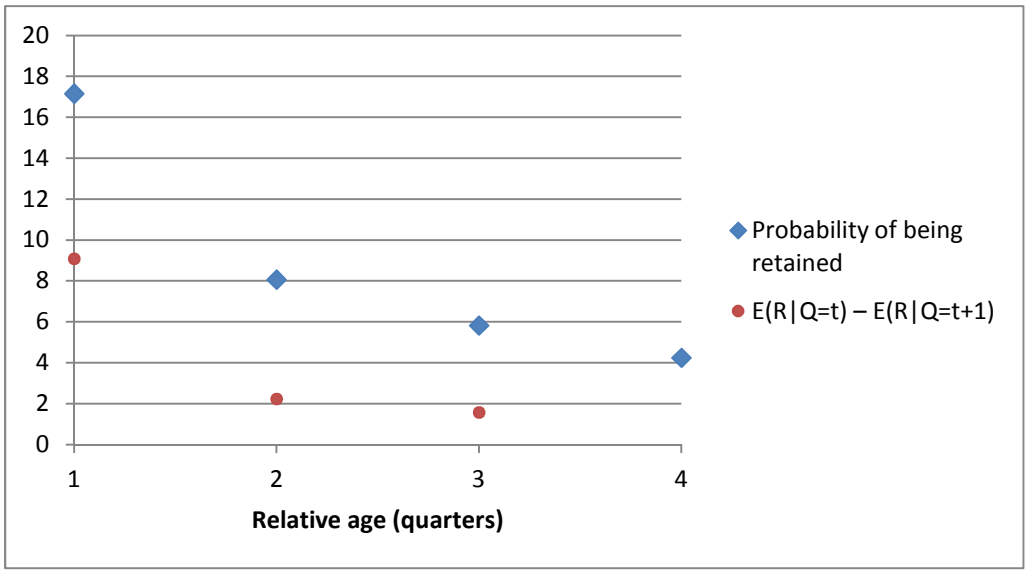
At first, each individual is assigned to one of four relative age groups. The youngest pupils (i.e. the ones with $A = 1$, $A = 2$, and $A = 3$) are assigned to the age group with $Q = 1$, called

[1] This is not the case if there is parental birth date targeting. For example, if more educated parents are more likely to manipulate the birth date of their child, in order to ensure that he or she is the oldest in class, the nonlinearity in the probability of being retained may be a result of the fact that the youngest children are less intelligent, on average. Bedard and Dhuey (2006), however, show that there is no evidence for the hypothesis that mothers who are more educated are more likely to have children with a higher relative age.

Q1. The children that are assigned to this group were born in the last three months before the cutoff date. Similarly, the pupils with relative age 4 – 6, 7 – 9, and 10 – 12 are assigned to Q2, Q3, and Q4, respectively.

Then, it is useful to plot the probability of being retained for the four age groups, for the same set of countries as in Figure 1. Again, as can be seen in Figure 2, there is a significant 'jump' in the probability of being retained for the youngest age group. Between Q2 and Q4, the relationship seems to be approximately linear.

Figure 2: Grade retention by age group (pooled sample of countries)



Now, it is possible to exploit this nonlinearity by estimating a DID regression for pupil performance. For this purpose, the children that belong to Q1 and Q2 are assigned to the "treatment" group, while the other children are assigned to the "control" group. Moreover, the pupils that belong to Q1 and Q3 are assigned to the "t = 1" group, and the children that belong to Q2 and Q4 are assigned to the "t = 0" group. In the "normal" DID framework, the time component is used to separate the observations before treatment from the observations after treatment, within both the treatment group and the control group. In the DID regression, this time component thus controls for the general trend in the outcome variable. In this paper, the "time" component separates the observations of the children belonging to the youngest age group from the observations of the children belonging to the oldest age group, within both the treatment group (Q1 and Q2) and the control group (Q3

and Q4). In the regression, then, it controls for the general effect of being younger, i.e. the (constant) maturity effect.

The DID equation looks as follows:

(2) $$T_{ijk} = \delta_0 + \delta_1 P + \delta_2 t + \delta_3 (P \times t) + \sigma S_{jk} + \theta X_{ijk} + \nu_{ijk} .$$

It can be estimated for each country separately, or for a pooled sample of countries. δ_3 is the coefficient of interest. It represents the impact of belonging to the youngest age group (Q1) on pupil performance. The model also contains school fixed effects and child-specific control variables.

In order to find the impact of grade retention on pupil performance, δ_3 needs to be divided by the additional increase in the probability of being retained for the youngest pupils. So:

(3) $$\beta_{DID} = \frac{\delta_3}{\left[E(R \mid Q=1) - E(R \mid Q=2)\right] - \left[E(R \mid Q=3) - E(R \mid Q=4)\right]} .$$

This means that the additional change in pupil performance for the youngest age group is entirely attributed to grade retention.

Again, this method is only valid if the parallel trend assumption holds. Here, it means that the change in pupil performance between Q3 and Q4 is a good approximation of the change between Q1 and Q2 that would have been observed, had it not been for grade retention. This amounts to assuming a constant maturity effect. The validity of this important assumption will be discussed in the results section.

Instrumental variables

A similar, but more sophisticated way to exploit the nonlinearity in the probability of being retained is to conduct an instrumental variables (IV) analysis. An instrumental variable is a variable that generates exogenous variation in the treatment variable. As mentioned before, belonging to the youngest relative age group might be such a source of variation in grade retention. The IV method works as follows. First, a first-stage equation is estimated by regressing the treatment variable on the instrument and the (exogenous) control variables. Then, a second-stage regression is estimated, in which the outcome variable is regressed on

the control variables and fitted values from the first-stage regression. The coefficient of the treatment variable should then represent the causal treatment effect.

The IV approach has a number of important assumptions. Most importantly, the independence assumption and exclusion restriction should hold. The independence assumption states that the instrument is randomly assigned. This means that the assignment of the instrument does not depend on potential (treatment) outcomes. The exclusion restriction says that the instrument should only have an effect on the outcome variable through the treatment variable. The instrument should not have any other direct impact on the outcome variable.

Looking at Figure 1, one can come up with several ways to use the IV method in order to exploit the nonlinearity in the probability of being retained. The main part of this paper will employ two different specifications. There are two reasons for doing so. First, not many papers exist in which a nonlinearity is used as an instrumental variable. Consequently, there is no standard or "best" way to do so at the moment. Second, by comparing the results of different specifications, it can be seen whether the results are sensitive to the construction of the instrument. In other words, the robustness of the results can be assessed.

The first specification exploits the discrete jump in the probability of being retained for the students with $Q = 1$ by using a Q1 dummy as instrument. The second specification focuses on the change in the trend between relative age and grade retention in the $Q = 1$ group, using an interaction term between Q1 and relative age as instrument. The first-stage regressions look as follows:

(4.1) $$R_{ijk} = \rho_0 + \rho_1 Q1_{ijk} + \rho_2 rel\_age_{ijk} + \gamma S_{jk} + \Omega X_{ijk} + \xi_{ijk}$$

(4.2) $$R_{ijk} = \rho_0 + \rho_1 (Q1_{ijk} \times rel\_age_{ijk}) + \rho_2 Q1_{ijk} + \rho_3 rel\_age_{ijk} + \gamma S_{jk} + \Omega X_{ijk} + \xi_{ijk}.$$

In these equations, $Q1_{ijk}$ is a dummy variable indicating whether the child has relative age $Q = 1$. In both models, rel_age_{ijk} is included as a control variable. It is a linear measure of the child's relative age in months. All equations also include school fixed effects and child-specific control variables. In order to avoid confounding changes in the trend with changes in the intercept, the model with an interaction term, equation (4.2), also includes the dummy

variable $Q1_{ijk}$. In order to further assess the robustness of the results, two additional specifications will be employed in the results section.

The second-stage regressions look as follows:

(5.1) $$T_{ijk} = \alpha_0 + \alpha_1 \hat{R}_{ijk} + \alpha_2 rel\_age_{ijk} + \mu S_{jk} + \varphi X_{ijk} + \eta_{ijk}$$

(5.2) $$T_{ijk} = \alpha_0 + \alpha_1 \hat{R}_{ijk} + \alpha_2 Q1_{ijk} + \alpha_3 rel\_age_{ijk} + \mu S_{jk} + \varphi X_{ijk} + \eta_{ijk}\,.$$

Again, these regressions can be estimated for each country separately, or for a pooled sample of countries. In both equations, α_1 is the coefficient of interest. It should represent the causal impact of grade retention on pupil performance.

As mentioned before, the IV approach relies heavily on the independence assumption and exclusion restriction. Here, the independence assumption states that a child's relative age (and, hence, the relative age group he or she belongs to) does not depend on his or her potential performance or *ex ante* probability of being retained. At hand, this assumption does not seem to be violated (see footnote 1). The exclusion restriction states that belonging to the youngest age group should only affect pupil performance through grade retention (recall that a linear measure of the child's relative age is included as a control variable in all models). It should not have any other direct impact on achievement. Again, this resembles the assumption of a constant maturity effect. The validity of both assumptions will be examined in the results section.

4. Data

Pupil performance

This paper uses the scores on an international standardized test as a measure of pupil performance. More specifically, data from the 1995 Trends in International Mathematics and Science Study (TIMSS) are used. As the name suggests, this test measures the performance of children in math and science in a number of countries, using nationally representative samples of pupils. The 1995 edition of this test is especially relevant for the purposes of this paper, because the sample of this study includes pupils that are enrolled in two adjacent grades. More specifically, it provides data on the test scores of children enrolled in grade 3 and 4, the grades containing the largest proportion of nine-year-olds. TIMSS 1995 measures the performance of third and fourth graders in 26 countries. Its test scores are standardized to have a mean of 500 and a standard deviation of 100[2].

For each child, there is data on his or her date of birth. Given the cutoff date in each country, it is then possible to predict the grade the child is supposed to be in. In Canada, for example, a five-year-old may start grade 1 if he or she turns six by December 31. Consequently, all Canadian pupils born in 1985 are predicted to be in grade 4 when the test is taken (May 1995). For each country, the sample is restricted to the birth cohort of children that "should" be in grade 4 at the time of the test (i.e. to the predicted grade 4 cohort).

Grade retention and relative age

The dataset does not contain a variable explicitly indicating whether a pupil has been retained. However, if the data indicates that a child is in grade 3 (or in an even lower grade), he or she is considered to have been retained[3].

As explained in the previous section, the cutoff date is used to construct the relative age measure in each country. It is a linear measure that indicates how many months the child was born before the relevant cutoff date.

[2] Since not all pupils have taken exactly the same test, the dataset contains five "plausible values" of both the math score and science score for each pupil. In this paper, the average of these five plausible values is taken, so as to obtain one composite measure of math performance and one composite measure of science performance for each child.

[3] If a child is behind his or her assigned grade, this may also be due to late entry. This means that a pupil is held back by his or her parents, which is called academic redshirting. Strictly speaking, it would therefore be more appropriate to speak of "being delayed" instead of "being retained". However, for the purpose of ease, the term "grade retention" will be used to refer to both mechanisms throughout most of this paper.

The cutoff dates are obtained from Webbink and Gerritsen (2013). They provide a list of cutoff dates in more than 30 countries, including their respective sources. For some countries, however, there is no source explicitly mentioning the cutoff date in 1995. These countries are dropped from the analysis. For all countries studied, it is also checked whether there is indeed a (sharp) discontinuity in average grade around the cutoff. The relevant cutoff dates can be found in Table A1 in the appendix. Only the countries where a significant first-stage effect is found in equation (4.1) or (4.2) are considered (see the next section).

Other variables

TIMSS 1995 also provides data on the individual characteristics of the pupils and the families they live in. In this paper, the following socioeconomic control variables are used: gender, child is born in the country of residence, household possesses a calculator, household possesses a computer, household owns more than 100 books, number of people living in the household, child has a native born father, child has a native born mother, and child lives with both parents. For each child, it is also indicated which school he or she goes to. Due to nonreporting for some of the socioeconomic controls, some observations are dropped from the sample. In the end, a total of 23,715 observations remains.

Table 1 presents summary statistics of the most important variables that were discussed in this section. It can be seen that there is quite some variation in the prevalence of grade retention. In Norway, for example, less than one percent of the pupils in the predicted grade 4 cohort have been retained. At the same time, more than sixteen percent of the Dutch children have been retained at least once. As mentioned in the introduction, this number is higher than in most other OECD countries. It is also the highest among the countries studied.

Table 1: Summary statistics

	Math score	Science score	% retained	Relative age	# of observations
Austria	556.8 (79.8)	559.8 (76.9)	13.9	6.3 (3.4)	2,123
Canada	525.6 (76.6)	535.9 (84.9)	7.3	6.6 (3.4)	6,951
Cyprus	515.5 (76.2)	486.8 (64.2)	3.3	6.8 (3.3)	2,557
Hong Kong	590.2 (73.2)	535.0 (69.8)	11.6	6.3 (3.5)	3,647
Korea	608.5 (65.9)	593.8 (60.1)	11.7	6.2 (3.5)	2,594
Netherlands	570.0 (74.0)	553.5 (62.5)	16.8	6.4 (3.4)	1,937
Norway	504.3 (65.6)	533.9 (76.6)	0.9	6.6 (3.3)	2,000
Portugal	489.7 (72.1)	491.7 (73.8)	7.0	6.6 (3.4)	1,906
Pooled sample	545.3 (82.8)	536.6 (79.6)	8.8	6.5 (3.4)	23,715

Notes: Standard deviations are in parentheses.

5. Results

Visual inspection

Before turning to the difference-in-differences calculations and instrumental variables estimates, it is useful to visually inspect the test score and grade retention patterns in the countries studied. Not only does this give an intuitive illustration of how both methods work, it also gives an indication of the results one can expect to find.

Figure 3: Math scores (pooled sample of countries)

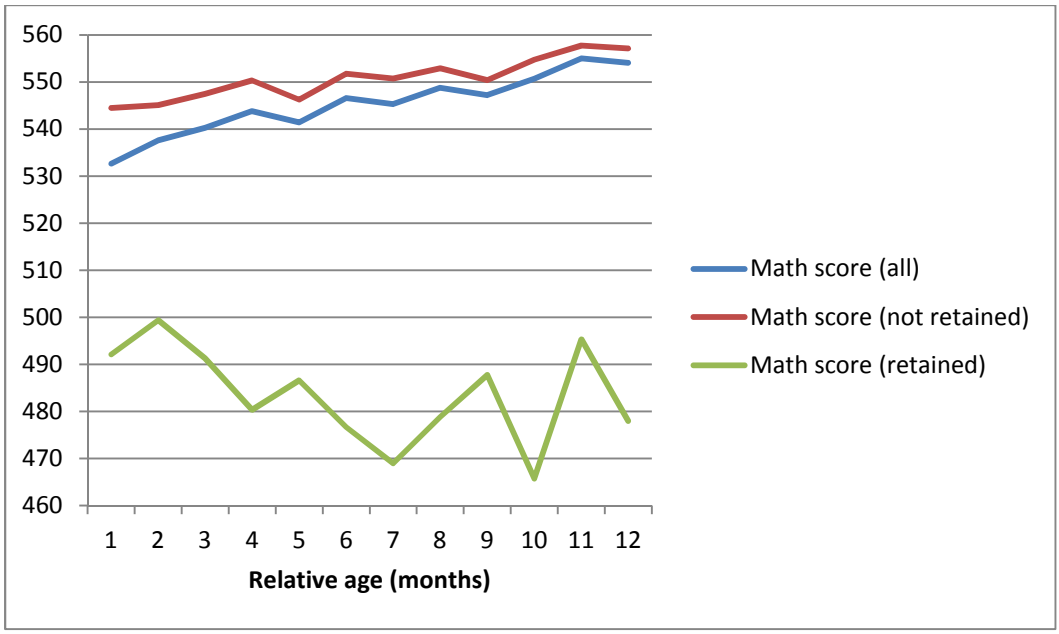


Figure 3 plots the math scores against relative age (in months) for the pooled sample of countries. The blue line represents the math scores of all pupils in the sample, while the red line and green line represent the math scores of non-retained and retained pupils, respectively.

Naïve analysis

First of all, it can be seen that the retained pupils perform significantly worse than the non-retained pupils. Ignoring the control variables for the moment, conducting a naïve analysis would amount to comparing the average math score between the two groups. An OLS regression is then expected to generate a β_1 coefficient of approximately –70.

However, as mentioned before, the two groups are unlikely to be similar. The figure also seems to indicate this. As expected, the red line is upward sloping. This confirms the belief

that older pupils tend to perform better on standardized tests because of the maturity effect. The green line, however, has a less consistent (but negative) trend. It can be seen that the youngest retained pupils outperform the retained pupils that are a bit older (e.g. the ones with relative age $A = 4$, $A = 5$, and $A = 6$). This observation can be explained by referring back to Figure 1, which plotted the grade retention pattern in the pooled sample of countries. As the youngest pupils are disproportionately more likely to be retained, one could argue that some of the youngest retained pupils in Figure 3 have been retained for other reasons than their poor performance in class. To see this, it is useful to theorize that normally a pupil is retained whenever his or her (composite) measure of performance in class is below a given threshold. In that case, all pupils that have been retained are assumed to have scored below this threshold. However, given the assumption of a linear relationship between performance in class and relative age, the nonlinearity in the probability of being retained seems to suggest that some of the youngest retained pupils have actually scored *above* the threshold. Consequently, it can be argued that older retained pupils are more likely to have self-selected themselves into treatment, or, similarly, that the retained pupils that belong to the Q1 group are more likely to have been "randomly" retained. The pattern of the green line in Figure 3 thus suggests that "partially randomly retained" pupils outperform older, "self-selected" retained pupils. This seems to indicate that pupils that self-select themselves into grade retention are less able and motivated than average. In other words, the potential outcomes of retained pupils are worse than those of on-time pupils.

Having established (informally) that retained pupils are indeed negatively selected, one can now get an idea of what the "true" results should look like. As explained in the section describing the identification strategy, OLS will generate coefficients that are lower (or more negative) than the "true" parameters if there is selection bias. Therefore, one would expect the causal impact of grade retention on pupil performance to be less negative than the naïve guess (–70). One could argue that the trend between relative age and performance for the group of retained pupils would be similar to the trend for the group of non-retained pupils, *had all retained pupils been retained randomly*. Assuming, for the moment, that all pupils who have relative age $A = 1$ have been randomly retained, this would mean that the "true" green line in Figure 3 (i.e. the line representing the relationship between relative age and performance for the group of randomly retained pupils) would have an intercept of 492 and

a positive slope. The estimate of the causal impact of being retained on math performance is then expected to equal about –50 points.

Differences-in-differences

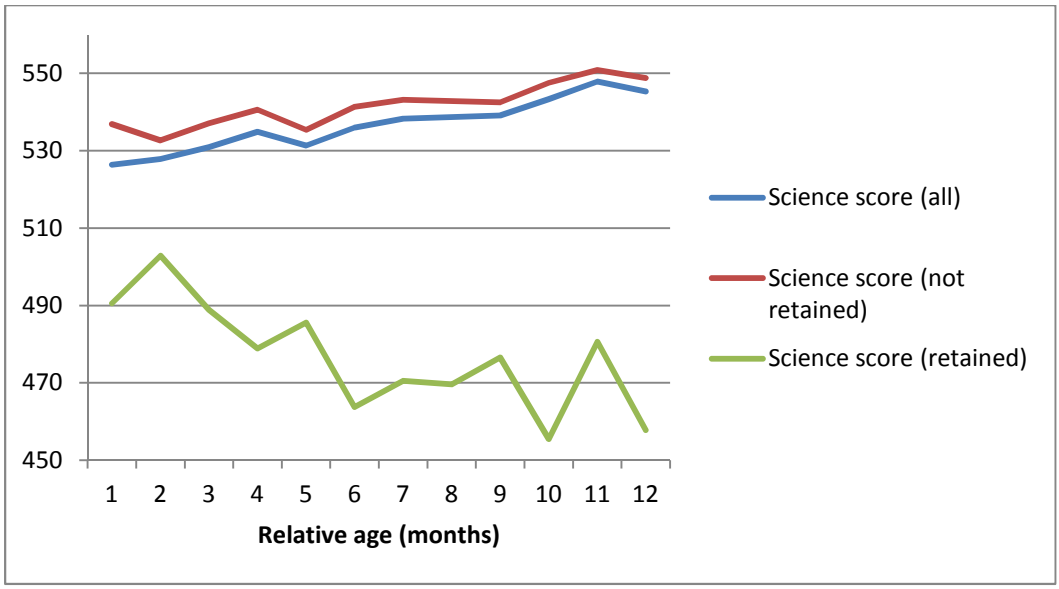
As explained in the identification strategy section, the difference-in-differences method attributes the additional change in pupil performance for the youngest age group (Q1) to grade retention. As can be seen by looking at the blue line in Figure 3 (and, in more detail, by looking at the red line in Figure A1), however, the change between Q1 and Q2 seems to be similar to the change between Q3 and Q4, which is seen as the baseline change when using the DID method. After adding the full set of controls, however, the change between Q1 and Q2 does appear to be bigger than the baseline change. This can be seen by looking at the blue dotted line in Figure A1. The additional (negative) change between Q1 and Q2 equals about 7.5 points. This number is then divided by the additional increase in the probability of being retained, which is about 7 percentage points (see Figure 2). The DID estimate of the causal impact of grade retention would then be approximately equal to $\frac{-7.5}{0.07} \approx -107.1$. Indeed, this yields an estimate bigger (in absolute value) than the naïve estimate, while one would expect it to be smaller, as explained in the previous paragraph. This apparent anomaly will be examined in detail below. In any case, one should be careful in interpreting the DID results and prefer the more sophisticated (and more flexible) IV approach.

Instrumental variables

In this paper, the IV method is very similar to the DID method. The main difference is the fact that now, the change in the intercept or trend (depending on which specification is used) for the Q1 group relative to the *overall* and *monthly* trend is analyzed. When the first IV specification is used, the (average) additional drop in performance for the Q1 group is divided by the additional increase in the probability of being retained for this group. If the second specification is used, one divides the change in the trend between relative age and performance in the Q1 group by the increase in the trend between relative age and grade retention in this group. Looking at Figure 3 and Figure 1, one can see that the IV analysis is then also expected to generate more extreme results than the naïve estimate. As will be shown in the next subsection, this is indeed the case.

One will reach similar conclusions after inspecting the science scores. Figure 4 plots these scores against relative age (in months) for the pooled sample of countries.

Figure 4: Science scores (pooled sample of countries)



A naïve OLS regression is now expected to generate a β_1 coefficient that is somewhat smaller in absolute value than before. Again, it seems to be the case that retained pupils are negatively selected, so estimates of the causal impact of grade retention are expected to be less negative.

In Figure 4, there do not seem to be any major discontinuities in performance for the youngest age group. However, as can be seen in Figure A2, the change between Q1 and Q2 does appear to be slightly bigger than the change between Q3 and Q4, after adding the full set of controls. The IV regressions are therefore also likely to find a significant change in the intercept or trend for the youngest age group. Both methods, however, will probably find smaller effects for the science scores than for the math scores. It could even be the case that the DID and IV analyses turn out to generate less extreme results than the naïve estimate. The regression results will be shown and discussed below.

Regression results

Difference-in-differences

Table 2: Difference-in-differences estimates of the effects of grade retention

	DID coefficient math	DID coefficient science	Difference retained (percentage points)	DID estimate math	DID estimate science
Austria	-4.1 (6.1)	-7.5 (5.7)	3.1	-132.4	-240.1
Canada	-7.9** (3.1)	-1.6 (3.3)	11.1	-70.8	-14.6
Cyprus	-6.9 (5.6)	-2.3 (4.6)	1.7	-413.1	-138.1
Hong Kong	-11.0*** (3.9)	-3.9 (3.8)	9.6	-114.7	-40.4
Korea	-3.1 (4.8)	2.9 (4.4)	6.9	-44.8	42.5
Netherlands	-13.2** (5.9)	-7.9 (4.9)	7.5	-175.6	-105.1
Norway	-10.2* (5.3)	-10.8* (6.3)	2.4	-417.5	-441.9
Portugal	3.6 (5.7)	-0.2 (5.7)	1.6	231.8	-12.4
Pooled sample	-7.8*** (1.7)	-3.8** (1.6)	7.5	-103.4	-51.1

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 2 shows the DID calculations. In the second and third column, the δ_3 coefficients from equation (2) can be found. They measure the causal impact of belonging to the Q1 group on math and science performance, respectively. The additional increase in the probability of being retained for the Q1 group $\left(\left[E(R\,|\,Q=1)-E(R\,|\,Q=2)\right]-\left[E(R\,|\,Q=3)-E(R\,|\,Q=4)\right]\right)$ in each country is shown in the fourth column. Finally, the DID estimates of the causal impact of being retained on math and science performance are shown in the fifth and sixth column, respectively. These are the β_{DID} coefficients from equation (3).

Not many of the DID regressions yield significant results. Belonging to the youngest age group only has a significant impact on one's math score in Canada, Hong Kong, the Netherlands, and Norway. For the science scores, the DID coefficient is only (marginally) significant in Norway. The β_{DID} coefficients are displayed for each country, but one should only interpret them if the DID coefficient is significant.

As mentioned before, the TIMSS test scores have a standard deviation of 100 by construction. Therefore, the estimates can be divided by 100 to obtain an estimate of the effect size. Moreover, it should be mentioned that, on average, one's TIMSS score is expected to increase with about 40 points between two adjacent grades (as noted by Grønmo and Onstad [2013] among others). This means that if one finds β_{DID} coefficients bigger in absolute value than 40, this indicates that being retained has more (negative) effects than just the effect of "missing out" on a more advanced year of schooling.

For Canada, the DID estimate of the causal impact of grade retention on math performance equals about –70.8 points. This estimate is smaller than the coefficient one would obtain when running a naïve regression (–77.8, see Table A2). Still, it is relatively large: it is more negative than –40, and the effect size is larger than in most other same-age studies (see Allen [2009]). The final part of this section will be aimed at explaining the size of the estimates.

In Hong Kong, being retained is estimated to reduce the math score by about 114.7 points, on average. Not only is this estimate large, it is also more negative than the naïve estimate (–66). This is also the case in the Netherlands: the DID estimate (–175.6) is way more negative than the naïve estimate of –106.7 points. This apparent anomaly will also be explained at the end of this section.

In Norway, both the math and science score are expected to decrease by more than 400 points when a pupil has been retained. Obviously, these results cannot be true. Most likely, it is a result of the fact that grade retention is very uncommon in Norway. In the dataset, only 17 children are observed to have been retained. Moreover, the DID coefficients are only significant at the ten percent level. Given the social promotion policy in Norway (Driessen et al., 2014), one should generally be careful in interpreting the results for this country, as they may be driven by extreme cases.

Finally, the pooled estimate of the effect of grade retention on math performance (–103.4) is also more negative than the naïve estimate of –78.1, whereas the effect on science performance (–51.1) is less extreme than the naïve estimate (–65.3). This is in line with what was observed when the test score and grade retention patterns were inspected visually.

The magnitude of the results will be discussed in detail below, but Table 2 provides some information about where it comes from. One can see that the DID coefficients in the second and third column are not necessarily "too big" in absolute value, as they never exceed 14 points. However, the additional increase in the probability of being retained for the Q1 group is relatively small in most countries. As seen in the fourth column, it never exceeds 12 percentage points. Consequently, if one attributes the change in performance for the Q1 group entirely to the small change in the probability of being retained for this group, big estimates of the causal impact of grade retention on achievement can be expected to be found. Doing so is inappropriate, however, if the change in performance is not solely a result of the change in the prevalence of grade retention. Given the big estimates, this cannot be ruled out. It is therefore possible that the assumption of a constant maturity effect is violated.

Instrumental variables

Table 3: Instrumental variables results – Math (1)

	RF	FS	IV	OLS	F-statistic instrument
Austria	-1.9 (5.3)	0.02 (0.02)	-100.9 (241.1)	-102.9*** (4.2)	0.6
Canada	-5.0* (2.7)	0.09*** (0.01)	-54.2** (27.5)	-75.7*** (3.1)	81.0
Cyprus	-3.1 (5.0)	0.02* (0.01)	-151.8 (241.5)	-77.0*** (8.9)	3.1
Hong Kong	-10.7*** (3.4)	0.08*** (0.02)	-139.4*** (43.9)	-64.3*** (3.2)	20.6
Korea	-3.8 (4.1)	0.07*** (0.02)	-58.3 (59.2)	-49.2*** (3.7)	9.2
Netherlands	-8.2 (5.1)	0.06** (0.03)	-137.3** (69.9)	-105.9*** (3.6)	4.6
Norway	-8.9* (4.7)	0.02*** (0.01)	-386.5* (223.4)	-64.7*** (14.3)	9.0
Portugal	2.5 (5.0)	0.02 (0.02)	125.9 (288.7)	-58.5*** (6.5)	1.2
Pooled sample	-5.6*** (1.5)	0.06*** (0.01)	-91.4*** (22.1)	-75.9*** (1.5)	102.2

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 4: Instrumental variables results – Science (1)

	RF	FS	IV	OLS	F-statistic instrument
Austria	-6.2 (5.0)	0.02 (0.02)	-333.3 (398.9)	-82.5*** (4.1)	0.6
Canada	-0.3 (2.9)	0.09*** (0.01)	-3.1 (30.2)	-69.6*** (3.3)	81.0
Cyprus	-1.1 (4.1)	0.02* (0.01)	-57.0 (195.9)	-59.6*** (7.3)	3.1
Hong Kong	-5.7* (3.3)	0.08*** (0.02)	-73.9* (40.9)	-52.5*** (3.2)	20.6
Korea	-0.8 (3.8)	0.07*** (0.02)	-12.1 (55.7)	-41.9*** (3.5)	9.2
Netherlands	-6.1 (4.3)	0.06** (0.03)	-102.4 (62.9)	-72.0*** (3.2)	4.6
Norway	-7.6 (5.6)	0.02*** (0.01)	-332.0 (250.8)	-57.3*** (16.9)	9.0
Portugal	-1.1 (5.1)	0.02 (0.02)	-58.6 (240.2)	-70.1*** (6.5)	1.2
Pooled sample	-3.1** (1.4)	0.06*** (0.01)	-50.9** (22.2)	-62.7*** (1.5)	102.2

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

The first IV specification exploits the discrete jump in the probability of being retained for the youngest pupils by using a Q1 dummy as an instrument for being retained. The results of this specification are shown in Table 3 (math) and Table 4 (science). In the second column of these tables (labeled “RF”), the reduced-form estimates are displayed. They measure the effect of belonging to the youngest age group on math performance. The third column contains the first-stage (“FS”) effects. These coefficients represent the size of the jump in the probability of being retained for the youngest pupils. In the fourth column, the IV estimates of the causal impact of grade retention are shown. They are compared with the estimates from a naïve (OLS) model that are shown in the fifth column. In order to avoid working with weak instruments, one should only interpret the IV coefficients if the F-statistic of the excluded instrument is bigger than 10 (this is the main reason why one should be careful in interpreting the DID results). The F-statistics can be found in the sixth, and last, column.

Only in Canada, Hong Kong, the Netherlands, and Norway (and in the pooled sample of countries), the effect of being retained on math performance turns out to be significant. For the Netherlands and Norway, however, the results may suffer from weak instrument issues.

In Canada, being retained is estimated to reduce the math score by about 54.2 points, on average. This estimate is less extreme[4] than the OLS estimate of –75.7 points.

For Hong Kong, again, the estimate of the causal impact of grade retention on math performance (–139.4) is quite big in absolute value. Moreover, the IV coefficient is more negative than the naïve estimate of –64.3 points. As expected, the estimate of the causal impact of grade retention on math performance (–91.4) is also more extreme than the naïve estimate (–75.9) for the pooled sample of countries.

The effect of being retained on science performance is only (marginally) significant in Hong Kong. Even though the estimate (–73.9) is less extreme than the estimated effect of grade retention on math performance, it is still more negative than the OLS estimate of –52.5 points.

For the pooled sample of countries, the IV coefficient is significant at the 5% level. As before, the estimate (–50.9) is smaller in absolute value than the naïve estimate (–62.7) when the science score is used as the outcome variable.

Table 5: Instrumental variables results – Math (2)

	RF	FS	IV	OLS	F-statistic instrument
Austria	8.2** (3.7)	-0.04** (0.02)	-209.23** (92.4)	-102.9*** (4.2)	5.1
Canada	3.9** (1.9)	-0.03*** (0.01)	-120.8** (57.6)	-76.0*** (3.1)	19.0
Cyprus	3.7 (4.0)	-0.02** (0.01)	-179.7 (189.0)	-76.9*** (8.9)	5.2
Hong Kong	7.5*** (2.3)	-0.06*** (0.01)	-128.9*** (39.0)	-63.9*** (3.2)	25.4
Korea	0.8 (2.9)	-0.07*** (0.01)	-10.6 (37.0)	-49.1*** (3.8)	24.6
Netherlands	-0.9 (3.7)	-0.02 (0.02)	42.5 (188.4)	-105.8*** (3.6)	1.2
Norway	-1.2 (3.4)	-0.02*** (0.01)	49.9 (137.2)	-63.2*** (14.3)	19.5
Portugal	6.1 (3.7)	-0.04*** (0.01)	-162.0 (99.6)	-58.6*** (6.5)	7.7
Pooled sample	3.9*** (1.0)	-0.04*** (0.00)	-95.8*** (23.7)	-75.8*** (1.5)	90.0

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

[4] Due to the inaccuracy (i.e. high standard errors) of the IV estimates, however, they are never significantly different from the OLS estimates. This is true for all IV specifications that are discussed in this section.

Table 6: Instrumental variables results – Science (2)

	RF	FS	IV	OLS	F-statistic instrument
Austria	4.1 (3.5)	-0.04** (0.02)	-104.6 (78.7)	-82.5*** (4.1)	5.1
Canada	2.1 (2.1)	-0.03*** (0.01)	-67.2 (60.7)	-70.4*** (3.4)	19.0
Cyprus	0.32 (3.3)	-0.02** (0.01)	-15.7 (152.6)	-59.6*** (7.3)	5.2
Hong Kong	4.9** (2.3)	-0.06*** (0.01)	-84.5** (37.3)	-52.4*** (3.2)	25.4
Korea	2.1 (2.7)	-0.07*** (0.01)	-28.7 (33.7)	-42.0*** (3.5)	24.6
Netherlands	-3.0 (3.0)	-0.02 (0.02)	142.1 (224.8)	-71.9*** (3.2)	1.2
Norway	-2.5 (4.0)	-0.02*** (0.01)	104.1 (163.9)	-56.0*** (17.0)	19.5
Portugal	6.3* (3.8)	-0.04*** (0.01)	-168.0* (99.6)	-70.1*** (6.5)	7.7
Pooled sample	2.2** (1.0)	-0.04*** (0.00)	-52.9** (23.7)	-62.8*** (1.5)	90.0

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

The second IV specification exploits the change in the trend between relative age and grade retention in the $Q = 1$ group by using an interaction term between Q1 and relative age as an instrument for being retained. The results are shown in Table 5 (math) and Table 6 (science). Here, the reduced-form coefficients should be interpreted as the additional effect of getting a month older on performance, when one belongs to the Q1 group. Similarly, the first-stage coefficients represent the additional change in a pupil's probability of being retained as he or she gets a month older, when he or she belongs to the youngest age group.

Using this specification, one avoids the weak instrument problem for Korea and Norway. However, in both countries, being retained appears to have no significant impact on both math and science performance. If one finds no significantly negative relationship between grade retention and performance in a same-age study, this means that the performance of retained pupils is comparable to that of their promoted age peers that are a grade ahead of them. So, retained pupils perform "as if" they had not been retained, on average. If this also applies to the children that are usually retained (i.e. the less able ones), it is very likely that grade retention actually has positive (long-run) effects in practice.

There are some potential explanations for not finding significantly negative effects on performance in some countries. It may be possible, for example, that retained pupils receive supplemental interventions in school during the retention year in these countries, as mentioned in Schwerdt and West (2013). It could also be the case that retained pupils receive extra lessons outside school (e.g. private education). Finally, it is possible that third and fourth grade pupils are in the same class in some schools (multi-grade classrooms), so there might be knowledge spillovers between retained and non-retained children.

In Canada and Hong Kong, being retained is estimated to reduce the math score by about 120.8 and 128.9 points, on average, respectively. Both estimates are bigger in absolute value than the naïve estimates. Moreover, for Canada, the IV coefficient in this specification differs substantially from the IV coefficient in the first specification (–54.2). The differences between the two specifications will be examined below.

For the pooled sample of countries, again, the IV estimate of the effect of grade retention on math performance (–95.8) is more negative than the naïve estimate of –75.8 points.

Being retained only appears to have a significant impact on science performance in Hong Kong, if a country is only considered when there does not seem to be a weak instrument problem. The IV coefficient (–84.5) is more negative than the OLS coefficient (–52.4), but for the pooled sample of countries, again, the IV estimate (–52.9) is smaller in absolute value than the naïve estimate (–62.8).

In sum, it can be concluded that the results are mixed across countries. In general, the results for math performance are more pronounced than the results for science performance. Below, it is assessed whether the IV results shown above are robust to the specification chosen.

Robustness checks

Given that one can only compare the results for Canada, Hong Kong, and the pooled sample of countries (because of the weak instrument problem), in general, the two IV specifications seem to yield more or less similar results. However, as mentioned before, the estimates of the effect of grade retention on math performance in Canada appear to differ substantially across the two specifications. Even though the differences are never statistically different at conventional levels, it should be noted that the IV coefficient in the second specification is

more than twice as large in absolute value as in the first specification. This difference can be explained by looking at the grade retention pattern in Canada. As can be seen in Figure A3, there is a big jump in the probability of being retained between *A* = 3 and *A* = 4. This jump is responsible for the relatively large first-stage coefficient in the first specification (0.09). It is not exploited in the second specification, however, as the Q1 dummy is only used as a control variable in equation (4.2). Consequently, the first-stage coefficient is much smaller in absolute value (–0.03), while the two reduced-form coefficients differ less substantially across the first and second specification (–4.99 and 3.86, respectively). Given the grade retention pattern in Canada, the first specification may be seen as more realistic.

In order to further assess the robustness of the results, two additional specifications will be employed. The first one focuses on the sharp trend increase in the probability of being retained between *A* = 2 and *A* = 1 by using an interaction term as an instrument for being retained. The second specification only uses an *A* = 1 dummy as instrument. The first-stage regressions look as follows:

(4.3) $$R_{ijk} = \rho_0 + \rho_1(youngest^2_{ijk} \times rel\_age_{ijk}) + \rho_2 youngest^2_{ijk} + \rho_3 rel\_age_{ijk} + \gamma S_{jk} + \Omega X_{ijk} + \xi_{ijk}$$

(4.4) $$R_{ijk} = \rho_0 + \rho_1 youngest^1_{ijk} + \rho_2 rel\_age_{ijk} + \gamma S_{jk} + \Omega X_{ijk} + \xi_{ijk}\,.$$

In these equations, $youngest^2_{ijk}$ takes on the value 1 if the child has relative age *A* = 1 or *A* = 2, while $youngest^1_{ijk}$ indicates whether the child has relative age *A* = 1. Again, *A* indicates how many months the child was born before the cutoff date. The second-stage regressions look as follows:

(5.3) $$T_{ijk} = \alpha_0 + \alpha_1 \hat{R}_{ijk} + \alpha_2 youngest^2_{ijk} + \alpha_3 rel\_age_{ijk} + \mu S_{jk} + \varphi X_{ijk} + \eta_{ijk}$$

(5.4) $$T_{ijk} = \alpha_0 + \alpha_1 \hat{R}_{ijk} + \alpha_2 rel\_age_{ijk} + \mu S_{jk} + \varphi X_{ijk} + \eta_{ijk}\,.$$

The results of these specifications can be found in Table 7–10 below.

Table 7: Instrumental variables results – Math (3)

	RF	FS	IV	OLS	F-statistic instrument
Austria	11.0 (7.2)	-0.06* (0.03)	-191.3 (118.1)	-102.7*** (4.2)	2.9
Canada	8.3** (4.0)	-0.05*** (0.02)	-175.4** (84.8)	-75.8*** (3.1)	9.8
Cyprus	-1.8 (7.7)	-0.04** (0.02)	40.4 (173.0)	-76.7*** (8.9)	6.3
Hong Kong	8.3* (4.6)	-0.09*** (0.02)	-89.3* (46.2)	-63.7*** (3.2)	16.4
Korea	-4.7 (5.5)	-0.00 (0.03)	8786.8 (449402.1)	-49.4*** (3.8)	0.0
Netherlands	9.6 (7.2)	-0.06 (0.04)	-171.9 (110.4)	-106.1*** (3.6)	2.1
Norway	3.3 (6.9)	-0.05*** (0.01)	-68.5 (137.0)	-64.7*** (14.3)	19.0
Portugal	-0.1 (7.1)	-0.06** (0.03)	1.7 (118.3)	-58.2*** (6.5)	5.0
Pooled sample	4.7** (2.1)	-0.05*** (0.01)	-91.5** (37.3)	-75.7*** (1.5)	36.0

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 8: Instrumental variables results – Science (3)

	RF	FS	IV	OLS	F-statistic instrument
Austria	11.4* (6.8)	-0.06* (0.03)	-197.9 (122.9)	-82.5*** (4.1)	2.9
Canada	4.1 (4.2)	-0.05*** (0.02)	-86.9 (84.6)	-70.0*** (3.4)	9.8
Cyprus	-4.9 (6.3)	-0.04** (0.02)	112.4 (152.6)	-59.5*** (7.3)	6.3
Hong Kong	8.2* (4.5)	-0.09*** (0.02)	-88.7* (46.5)	-52.3*** (3.2)	16.4
Korea	-0.0 (5.0)	-0.00 (0.03)	8.5 (9149.3)	-42.2*** (3.5)	0.0
Netherlands	-3.2 (6.0)	-0.06 (0.04)	56.7 (124.8)	-72.1*** (3.2)	2.1
Norway	-4.3 (8.2)	-0.05*** (0.01)	89.5 (165.7)	-57.2*** (17.0)	19.0
Portugal	4.5 (7.3)	-0.06** (0.03)	-77.1 (116.3)	-69.7*** (6.5)	5.0
Pooled sample	2.5 (2.1)	-0.05*** (0.01)	-49.3 (37.4)	-62.8*** (1.5)	36.0

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 9: Instrumental variables results – Math (4)

	RF	FS	IV	OLS	F-statistic instrument
Austria	-13.1** (6.1)	0.07** (0.03)	-191.7** (84.6)	-102.9*** (4.2)	5.6
Canada	-9.3*** (3.3)	0.08*** (0.01)	-109.4*** (36.9)	-75.7*** (3.1)	45.3
Cyprus	-3.2 (6.6)	0.05*** (0.02)	-67.7 (132.6)	-77.0*** (8.9)	10.1
Hong Kong	-14.1*** (3.9)	0.12*** (0.02)	-116.0*** (30.8)	-64.3*** (3.2)	39.1
Korea	0.5 (4.7)	0.08*** (0.02)	5.7 (54.9)	-49.2*** (3.7)	11.6
Netherlands	-6.7 (6.0)	0.07** (0.03)	-101.0 (71.7)	-105.9*** (3.6)	4.2
Norway	-3.6 (5.7)	0.05*** (0.01)	-68.0 (103.7)	-64.7*** (14.3)	33.2
Portugal	-4.0 (6.1)	0.07*** (0.02)	-59.3 (85.2)	-58.5*** (6.5)	9.3
Pooled sample	-7.7*** (1.8)	0.08*** (0.01)	-93.3*** (19.7)	-75.9*** (1.5)	129.4

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 10: Instrumental variables results – Science (4)

	RF	FS	IV	OLS	F-statistic instrument
Austria	-11.2* (5.8)	0.07** (0.03)	-164.7** (81.6)	-82.5*** (4.1)	5.6
Canada	-4.0 (3.5)	0.08*** (0.01)	-47.5 (39.3)	-69.6*** (3.3)	45.3
Cyprus	1.9 (5.5)	0.05*** (0.02)	39.7 (113.2)	-59.6*** (7.3)	10.1
Hong Kong	-10.2*** (3.8)	0.12*** (0.02)	-84.0*** (30.0)	-52.5*** (3.2)	39.1
Korea	-2.0 (4.3)	0.08*** (0.02)	-24.3 (49.1)	-41.9*** (3.5)	11.6
Netherlands	2.0 (5.0)	0.07** (0.03)	30.2 (80.3)	-72.0*** (3.2)	4.2
Norway	1.6 (6.7)	0.05*** (0.01)	30.1 (123.8)	-57.3*** (16.9)	33.2
Portugal	-7.8 (6.2)	0.07*** (0.02)	-116.3 (87.0)	-70.1*** (6.5)	9.3
Pooled sample	-4.2** (1.7)	0.08*** (0.01)	-51.2*** (19.7)	-62.7*** (1.5)	129.4

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Most of the results that do not suffer from the weak instrument problem are comparable to the main estimates. More specifically, grade retention has no significant impact on both math and science performance in Korea (in the fourth specification) and in Norway (in both specifications). For Hong Kong, the coefficients are always significant and more negative than the naïve estimates. For both math and science performance, the pooled estimates of the causal impact of grade retention appear to be very similar across all four specifications (even though the coefficient is not significant in Table 8).

It can therefore be concluded that the results are more or less robust. Only for Canada, the size of the coefficients seems to depend heavily on the specification. This is due to the fact that there is a lot of variation in the grade retention pattern in Canada.

Anomalies

Some of the results that were found in this section are surprising. Most notably, the fact that the estimate of the effect of grade retention in Hong Kong is more negative than the naïve (OLS) estimate in all eight IV regressions (and in the DID specification) seems to be counterintuitive. This subsection is aimed at trying to explain the apparent anomalies in the results.

Selection bias

Firstly, one could argue that the results are true, even though some of them are surprising. In order to explain the observation that for some countries and in some specifications the IV coefficient is more extreme than the OLS coefficient, one should then argue that self-selected retained pupils outperform "normal", randomly retained pupils. Even though it is very likely that self-selected retained pupils are less able than on-time pupils and would be less motivated than their non-retained age peers (had they not been retained), it could theoretically be the case that *randomly* retaining "normal" children has such detrimental effects on their motivation, that these motivational effects on performance actually dominate the *ex ante* differences in ability and motivation between retained and "normal" pupils.

When visually inspecting the test score patterns of retained children, one would then expect to observe that the youngest children are outperformed by the pupils that are a bit older, because the youngest age group contains more randomly retained children. As shown at the

beginning of this section, however, it is actually the other way around for the pooled sample of countries. In Canada and Hong Kong, the countries for which counterintuitive results were (sometimes) found, the youngest retained pupils also seem to outperform the retained pupils belonging to the Q2 group, as can be seen in Figure A4 and A5. Therefore, this potential explanation is implausible.

Violation of assumptions

The second potential explanation of the fact that some of the results seem to be "wrong" is that not all assumptions are satisfied. If could be the case, for example, that the exclusion restriction is violated. This would mean that belonging to the youngest relative age group (i.e. being youngest in class) has a direct impact on performance, other than through grade retention. In other words, there could be a non-constant maturity effect. It may be possible, for instance, that being youngest in class has negative psychological effects.

It is also possible that the independence assumption is violated. Since one's relative age is directly affected by one's birth date, it could be the case that nonlinear relative age effects on performance actually reflect direct season of birth effects on performance, especially in the country-specific analyses. Bound and Jaeger (1996), for example, refer to studies finding that performance in school, health factors, and personality traits, among other factors, all vary by season of birth. Bedard and Dhuey (2006) find no evidence for the hypothesis that mothers who are more educated are more likely to have children with a higher relative age (see footnote 1), but it is possible that other (unobservable) factors affecting performance are indeed correlated with a pupil's relative age.

Even though the assumptions cannot be formally tested, there is a way to get an idea of whether it makes sense to think that they are violated. If the nonlinearity in pupil performance is only attributable to grade retention, then one would not expect to find any direct nonlinear relative age effects on the performance of on-time pupils. In order to test whether this is the case, the DID and main reduced-form regressions are estimated for the sample of pupils that have *not* been retained.

It should be noted, however, that the sample of on-time pupils used in these regressions might not be a good representation of the population of grade 4 pupils. The simple reason for this is the fact that this sample only includes the children that have never been retained,

which means that they are probably more able and motivated than average (especially in the countries where grade retention is common). More importantly, the sample of Q1 pupils is even more restricted. Due to the jump in the probability of being retained for this group, only the most intelligent and motivated Q1 children are likely to be included in this subsample. It is therefore possible that there is a nonlinear relationship between relative age and the performance of on-time pupils. In that case, one would expect to find positive coefficients of the Q1 dummy, and negative coefficients of the interaction term.

In order to avoid these sample selection issues, the regressions are also run for two countries where grade retention is very uncommon and where no significant first-stage effects were found: Iceland and Singapore.

Table 11: Coefficients for the sample of non-retained pupils

	DID math	DID science	RF IV math (1)	RF IV science (1)	RF IV math (2)	RF IV science (2)
Austria	0.1	-4.6	-0.2	-6.2	4.6	-1.5
	(5.7)	(5.4)	(5.0)	(4.8)	(3.7)	(3.5)
Canada	-3.1	-3.4	-1.1	2.8	1.0	1.0
	(3.1)	(2.2)	(2.7)	(2.9)	(2.1)	(2.2)
Cyprus	-6.2	-2.3	-2.2	-0.3	4.3	2.0
	(5.6)	(4.6)	(5.1)	(4.1)	(4.1)	(3.3)
Hong Kong	-4.7	1.3	-5.6*	-2.0	6.4***	2.4
	(3.8)	(3.8)	(3.3)	(3.3)	(2.4)	(2.4)
Korea	0.8	6.6	-1.5	2.1	-3.3	-0.2
	(5.0)	(4.5)	(4.3)	(3.9)	(3.1)	(2.8)
Netherlands	-4.1	-6.1	-1.2	-5.4	-1.7	-3.3
	(5.4)	(4.7)	(4.8)	(4.2)	(3.6)	(3.2)
Norway	-8.1	-9.0	-7.2	-6.1	-2.1	-3.7
	(5.3)	(6.3)	(4.7)	(5.6)	(3.4)	(4.0)
Portugal	3.3	-1.1	2.8	-1.7	2.1	2.6
	(5.7)	(5.7)	(5.0)	(5.0)	(3.9)	(3.8)
Pooled sample	-3.7**	-1.4	-2.4	-1.2	1.6	0.3
	(1.7)	(1.6)	(1.5)	(1.5)	(1.1)	(1.1)
Iceland	1.8	5.5	1.5	4.3	2.2	6.5
	(6.2)	(7.1)	(5.5)	(6.3)	(3.8)	(4.4)
Singapore	1.4	0.9	3.1	0.1	0.9	1.3
	(3.5)	(3.2)	(2.9)	(2.7)	(2.0)	(1.8)

*Notes: Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*

Table 11 shows the coefficients from these regressions. Even though most coefficients are not significantly different from zero, some of them are relatively big in absolute value. More disturbingly, for math performance, the reduced-form coefficient is significant in both IV specifications for Hong Kong, and the DID coefficient is significant for the pooled sample of countries. The signs of these coefficients are not in line with what one would expect (see above). These results therefore suggest that it is indeed possible that not all assumptions are satisfied in each country, which could explain the sometimes counterintuitive results. Especially for Hong Kong, for which the most surprising results were found, the large effects are unlikely to represent the true effects of grade retention.

Size

Finally, it has been shown above that all significant estimates are bigger in absolute value than 40 points, which is the one-grade-level equivalent. Moreover, in general, the significant estimates of the impact of grade retention on performance are more negative than in most other same-age studies. These two observations can potentially be explained by considering the nature of the TIMSS test.

The TIMSS test is a low-stakes test, which means that the results do not have direct consequences for the children concerned. As a consequence, the effort exerted on the TIMSS test depends heavily on the intrinsic motivation of the child. Therefore, as mentioned by Figlio and Loeb (2011: 398), "students may not take a low-stakes test sufficiently seriously to do their best work". If being retained has a negative (causal) effect on someone's motivation, then retained pupils will exert less effort on the TIMSS test, on average, than on-time pupils. Consequently, the estimates of the causal impact of grade retention on TIMSS scores may be more negative than the estimates from studies that look at other tests, and (way) more negative than the assumed "upper bound" effect of –40 points.

At the same time, however, it seems unrealistic to think that the size of the results can be entirely explained by the fact that the TIMSS test is a low-stakes test. Given the findings from the previous subsection, the potential violation of one or more of the IV assumptions (in one or more of the countries studied) seems to be the most important explanation for the magnitude of the significant results. Yet, this cannot be stated with certainty.

6. Conclusion and discussion

This paper has tried to establish the causal impact of grade retention on performance in elementary school. By exploiting a nonlinearity in the probability of being retained, it was attempted to find unbiased estimates of the effect of being retained on the math and science scores on a standardized test (TIMSS 1995). The international character of the TIMSS test made it possible to look at multiple countries.

The main results are mixed. In Canada and Hong Kong, being retained is found to have a significantly negative impact on performance. These estimates are bigger in absolute value than in most other studies. This observation could partially be explained by the fact that the TIMSS test is a low-stakes test, combined with the presumption that grade retention has a negative impact on motivation. However, it is also possible that not all IV assumptions are satisfied, especially in Hong Kong. For Norway (where grade retention is very uncommon) and Korea, no significant relationship could be found in any of the preferred IV specifications. The results for the other countries are more difficult to interpret, due to the weak instrument problem.

The fact that the results are mixed seems to suggest that the effects of grade retention depend heavily on the specific situation in each country. More specifically, the efficacy of grade retention as a method of remediating poor performance seems to depend on the presence of additional educational policies to accommodate the retention policy. The results for Canada, for example, suggest that grade retention in itself has a tendency to have negative consequences for the performance of children. The case of Korea, however, seems to indicate that the net effects of grade retention do not necessarily have to be negative, as long as there are additional interventions in place. This is in line with the findings of Schwerdt and West (2013). Whether or not these other interventions alone would be more effective in improving performance than in combination with grade retention, remains a question open for debate.

Education professionals, and policy makers in general, are therefore advised to consider other interventions when trying to remediate poor performance in elementary school. These interventions can be valuable alternatives or necessary additions to retaining poorly performing pupils.

This paper has some limitations that could be addressed in future research. First of all, it has only looked at the short-term effects of being retained. It would have been interesting to look at long-term outcome variables, such as performance in high school and earnings. Second, this paper has only examined the impact of being retained on academic achievement. Grade retention, however, is also likely to affect other factors, like motivation and self-esteem. Third, only a limited number of countries could be studied for data reasons. Fourth, this paper has looked at a low-stakes test that was taken about twenty years ago. One could try to replicate this study with a high-stakes and more recent test to see if the results are any different. Fifth, this paper has actually investigated the causal impact of being delayed. Future research could address the potentially different effects on performance (and other outcome variables) of grade retention versus academic redshirting.

Finally, there are two potential improvements to the specific instrumental variables approach that has been taken in this paper. One could gain more insight into the validity of the IV assumptions by investigating whether there is a direct relationship between being youngest in class and performance (other than through grade retention), and, if so, through which mechanisms. In addition, one could try to find a stronger instrument, related to (relative) age, in order to avoid the weak instrument problem.

References

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, 31(4): 480-499.

Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics*: 1437-1472.

Bound, J., & Jaeger, D. A. (1996). On the validity of season of birth as an instrument in wage equations: A comment on Angrist & Krueger's "Does compulsory school attendance affect schooling and earnings?". NBER working paper 5835.

Driessen, G., Leest, B., Mulder, L., Paas, T., & Verrijt, T. (2014). *Zittenblijven in het Nederlandse basisonderwijs: een probleem?*. Nijmegen: ITS.

Figlio, D., & Loeb, S. (2011). School accountability. In Hanushek, E. A., Machin, S., & Woessmann, L. (2011). *Handbook of the Economics of Education*, 3: 383-421. Amsterdam: North-Holland.

Grønmo, L. S., & Onstad, T. (2013). *The significance of TIMSS and TIMSS Advanced: Mathematics education in Norway, Slovenia and Sweden*. Oslo: Akademika Publishing.

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

Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. In Shepard, L. A., & Smith, M. L. (1989). *Flunking grades: research and policies on retention*: 16-33. London: Falmer.

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of Educational Research*, 613-635.

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

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

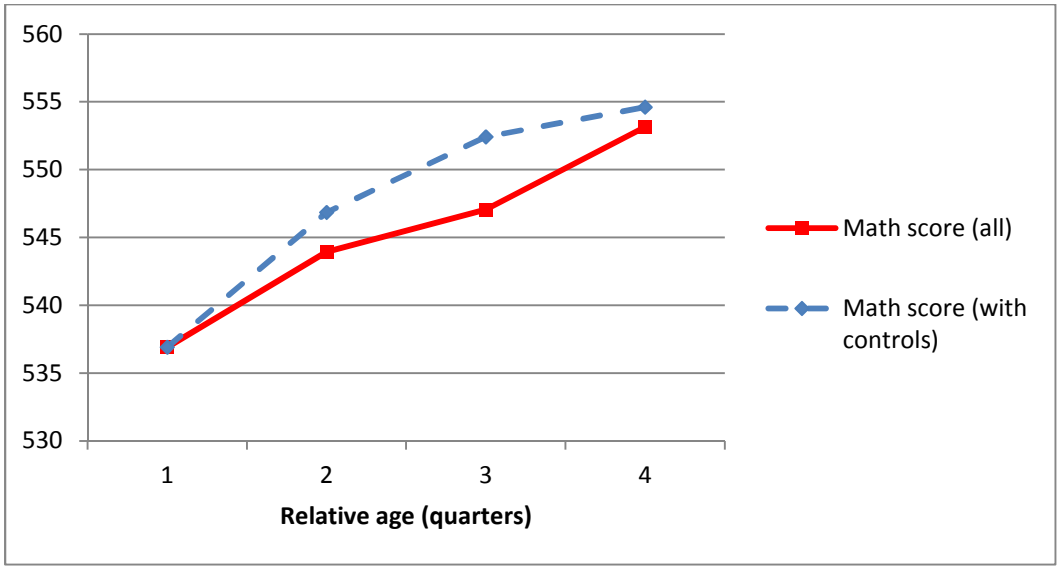
Johnson, E. R., Merrell, K. W., & Stover, L. (1990). The effects of early grade retention on the academic achievement of fourth-grade students. *Psychology in the Schools*, 27(4): 333-338.

Schwerdt, G., & West, M. R. (2013). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. Working paper. Harvard Kennedy School.

Webbink, D., & Gerritsen, S. (2013). How much do children learn in school? International evidence from school entry rules. CPB discussion paper 255.

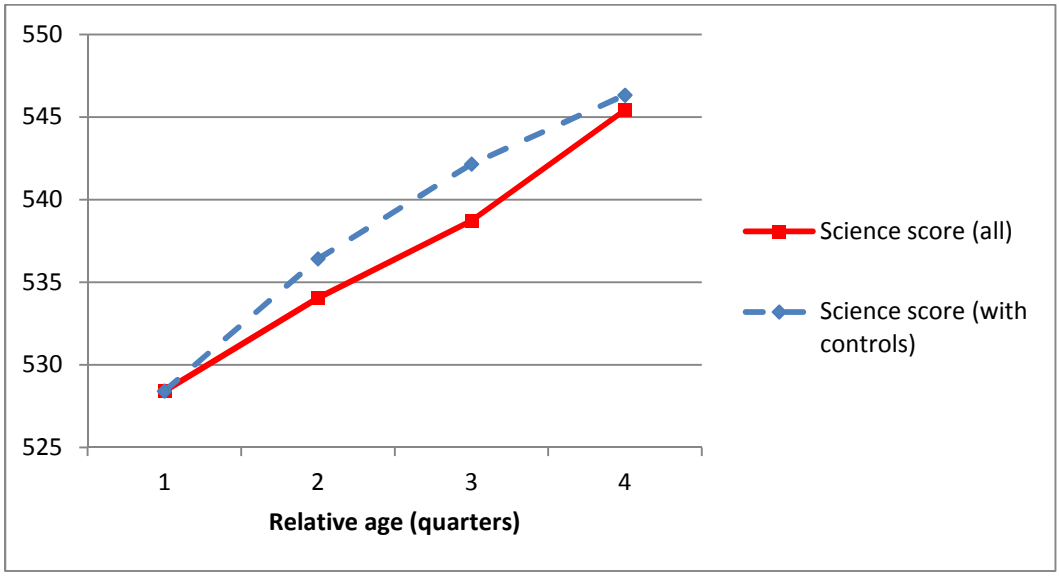
Appendix

Figure A1: Math performance by age group (pooled sample of countries)



Notes: When constructing the blue dotted line, all control variables are held constant (at zero). The line is then shifted to have the same intercept as the red solid line.

Figure A2: Science performance by age group (pooled sample of countries)



Notes: When constructing the blue dotted line, all control variables are held constant (at zero). The line is then shifted to have the same intercept as the red solid line.

Figure A3: Grade retention (Canada)

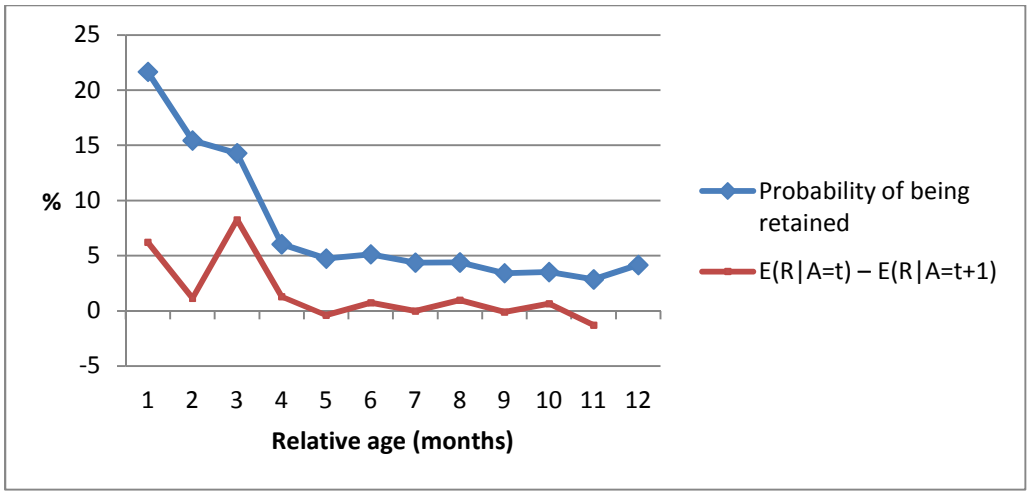


Figure A4: Math performance of retained pupils

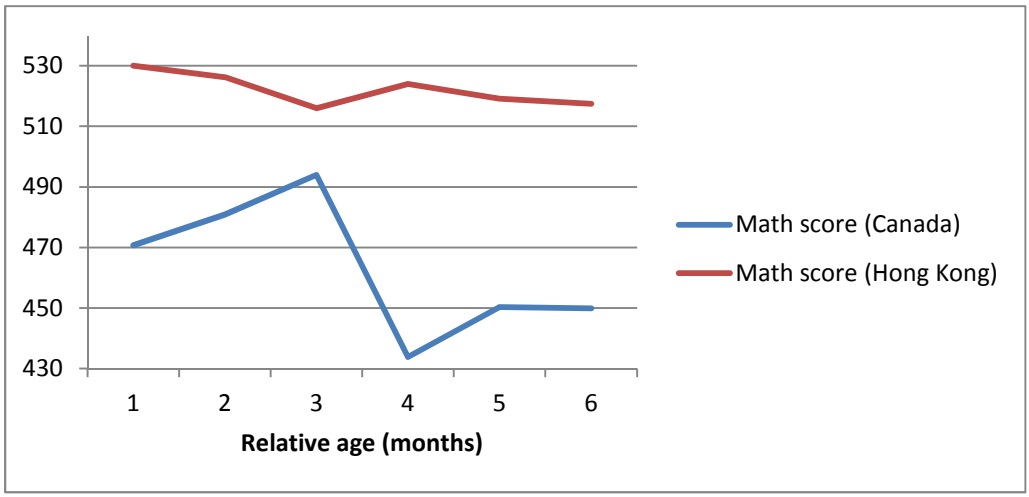


Figure A5: Science performance of retained pupils

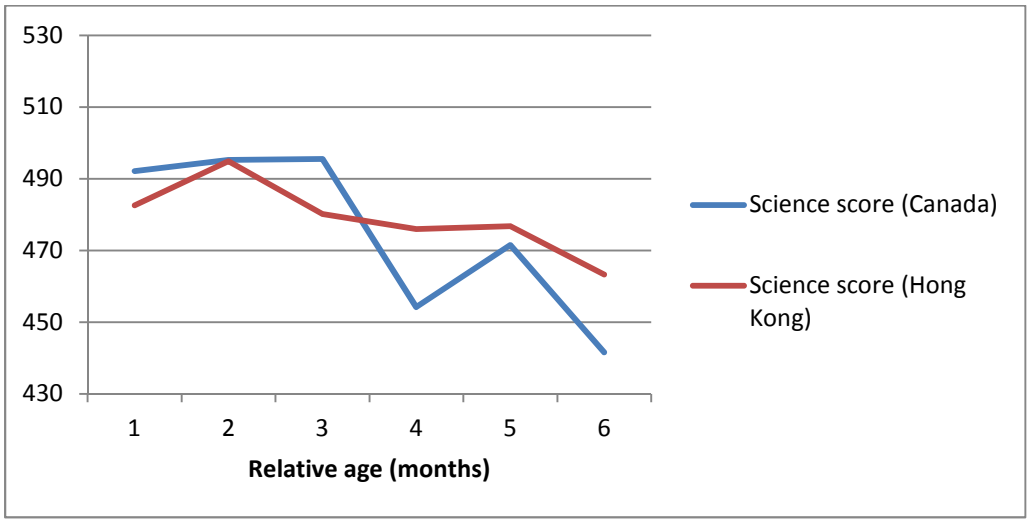


Table A1: Cutoff dates

	Cutoff date
Austria	September 1
Canada	January 1
Cyprus	March 1
Hong Kong	January 1
Korea	March 1
Netherlands	October 1
Norway	January 1
Portugal	January 1

Notes: Based on Webbink and Gerritsen (2013)

Table A2: Naïve estimates of the effects of grade retention

	Math	Science
Austria	-103.6***	-83.4***
	(4.1)	(4.0)
Canada	-77.8***	-72.2***
	(3.1)	(3.3)
Cyprus	-80.7***	-63.0***
	(8.9)	(7.4)
Hong Kong	-66.0***	-54.3***
	(3.1)	(3.1)
Korea	-50.5***	-43.7***
	(3.7)	(3.4)
Netherlands	-106.7***	-72.5***
	(3.5)	(3.1)
Norway	-73.7***	-68.7***
	(14.3)	(16.9)
Portugal	-60.0***	-72.2***
	(6.4)	(6.5)
Pooled sample	-78.1***	-65.3***
	(1.5)	(1.5)

Notes: In these regressions, math and science performance are regressed on a dummy variable indicating whether the child has been retained. Only S_{jk} *and* X_{ijk} *are included as control variables. Standard errors are in parentheses. *: significant at 10%. **: significant at 5%. ***: significant at 1%.*