Econometric Game 2024

Late to the Party: Inequalities Resulting from Delayed Primary School Starting

Team 8

Abstract

This paper investigates the causal effect of school-starting-age on educational outcomes in math and reading, using Pisa data. Specifically, we investigate three countries that are known to start primary education late (Latvia, Finland and Korea). Through regression discontinuity design, we find a significantly negative effect of starting late on educational outcomes, even though the consensus finding in the literature is that older children perform better. We argue that this is caused by the already high average age at which children start school in the investigated countries. We decompose the treatment effect to show that the negative effect of starting later is primarily driven by boys, especially boys from lower socio-economic backgrounds.

Keywords: returns to education, regression discontinuity design, school-starting-age

1 Introduction

Education is beneficial to not only the individual but also to society as a whole. Therefore, it is crucial for countries to determine how they design their education system optimally. One of the key factors in this policy design is determining the starting rule, which specifies when children start to go to school.

Despite the importance of the school-starting-age (SSA), there is very strong variation in it, even within Europe. Some countries like the United Kingdom opt to start very early around age 5, hoping that teaching children early will have positive long-term outcomes. Proponents of a low starting age argue that teaching students earlier will be especially beneficial for those with a lower socio-economic status. This policy is not without controversy however, as multiple pedagogical experts have stressed the importance of enough play for children Whitebread (2013). Other countries like Finland start much later, at age 7, exactly to allow children more play. Still, the question remains whether starting even later and allowing for more play could benefit students.

These differences between countries not only cause debates in the public domain, but also in the academic literature. To determine the effect of SSA on educational outcomes, much of the literature makes use of the fact that in grades some students are (almost) a year older than other students. For example, in Denmark students start school in the year in which they turn 6. This means that children born December 31 start a full year earlier with school than children born January 1, despite being born only one day apart. It is then possible to compare the oldest and youngest children in a grade: they have received the same level of education, but just started at a different age.

Many studies have compared these groups and a widely replicated finding in the literature is that these older children generally perform better at school than their younger peers Valdés (2023) Zhang et al. (2017). Not only do they perform better, they also have less behavioral issues and are diagnosed with learning disabilities less often Balestra et al. (2020). Even though the differences are clear at a young age, how much of the effect remains when students are older is unclear. Bahrs and Schumann (2020) find that in Germany students that start older with school smoke less and live healthier lifestyle, also in adulthood. On the other hand, some papers do not find an effect on the final degree of education attained Black et al. (2011) Oosterbeek et al. (2021). Both those papers also find that students who start school younger end up earning more on the labor market, as they likely enter the labor market sooner and therefore have more experience.

Using this cut-off point, where being born one day later or earlier can affect the school starting year by a full year, much of the literature considers either regression discontinuity design (RDD) or instrumental variable (IV). RDD considers students born just around the cut-off date and assumes groups of students before or after the cut-off date are essentially the same. This then means that whether students start relatively early or late is essentially designed randomly, as parents cannot control the birth of their child to the day. Hence, this allows for causal inference. IV is similar and also uses this cut-off point. It then assumes that school entry rules are exogenous and SSA can be used as an instrument to predict educational outcomes.

When investigating the effect of SSA on educational outcomes, many papers consider one specific region. For example, they make use of changes in the cut-off date in specific jurisdiction to employ RDD Peña (2017), Cook and Kang (2020), who consider Tlaxcala, Mexico and North Carolina, USA respectively. Other papers use country-specific datasets to draw conclusions Oosterbeek et al.

(2021). Still, cross-country analysis is performed less often, as comparison can be complicated by differing datasets.

Luckily, the OECD has collected data on educational performance in several countries since 2000 using standardised tests, known as the Program for International Student Assessment (PISA). This allows for comparisons between countries. It not only tests students performance in the subjects math, reading and science, but also asks them about their background, including questions about when they started school and their parents' socio-economic background.

Using this dataset, this paper first looks for countries that have clear starting rules and similar. This means looking for countries where we find clear cut-off points: for example everyone born in September starts with school early and all October births start late. This is essential to ensure that RDD is possible. Considering the modal starting age for every birth month, we identify three different countries that have the same starting rule in December/January and have similar average starting ages: Korea, Latvia and Finland. The set of countries is heterogeneous in terms of culture and student characteristics, but quite homogeneous in terms of policy, allowing us to clearly distinguish the effect of a one-year difference in SSA and reading and math aptitude.

For these countries, we find on average that starting school later decreases educational performance. At first glance, this seems contradictory with the existing literature, as consensus there is that older children (that have started later) perform better. However, the countries we consider have relatively high starting ages, indicating that starting school later only has a positive effect on younger students.

Next, we explore the mechanisms through which SSA affects educational performance by investigating heterogeneity. To this end, we perform subgroup analysis and investigate non-linear interaction effects for performance on reading tests. We find that the positive effect starting earlier has on educational performance is primarily driven by boys, especially those whose parents have received relatively little education and whose parents income is relatively low. We hypothesize that parents from a lower socio-economic background are less able to provide their children with a suitable education in the absence of formal schooling than higher educated and wealthier parents.

Section 2 deals with the data, provides descriptive statistic and explains our method of finding the countries with clear starting rules. Then Section 3 discuss our methodology, first to estimate country-specific effect and then to obtain global treatment effects. Section 4 gives our results and then Section 5 rounds off this paper by providing the conclusion and discussion.

2 Data

2.1 Data transformations

As PISA surveys take a sample, observation must be reweighed to ensure the sample is representative of the entire population. To this end, PISA provides weights for every observation, ω_{ijt} for observation i in country j and wave t as follows. We obtain standardised weights $\widetilde{\omega_{ijt}}$ by dividing by the sum of weights for that country and wave, as given below

$$\widetilde{\omega}_{ijt} = \frac{\omega_{ijt}}{\sum_{i=1}^{n} \omega_{ijt}}. \tag{1}$$

Next, PISA provides its reading and math scores in 10 plausible values, since not every student receives the same exact questions. We handle this by taking the average of these 10 plausible values for every student.

Furthermore, we delete observations that contain missing data for one of the following variables:

- Start age of school (*ST126Q01TA*)
- Birth month and year (*ST003D02T*, *ST003D03T*)
- plausible scores (*PV1MATH* to *PV10MATH*, *PV1READ* to *PV10READ*).

In total, these criteria leads to us delete about 8.5% of observations from out dataset. This leaves us with more than enough observations, but the percentage greatly differs between countries. For example, Norway does not have any data on the birth month and year of students in 2022 and no data on the birth month in 2018, so we disregard the entire 2018 and 2022 surveys for Norway.

One of the crucial variables in our research is the age at which students started school in a country. However, some students started schooling in another country. This can be derived from information about when students arrived in the country where they are currently enrolled in education. For simplicity, we delete all observation corresponding to students that arrived in the country of testing after they were born. Lastly, there are some outliers in this variable, as some students report having started at age 3 or younger or at age 9 or later. Here we simply state that these students started at exactly age 3 or exactly age 9 respectively.

Lastly, PISA tests students in three different subjects: reading, math and science. As reading and math are the most fundamental skills, we do not consider science scores.

2.2 Descriptive Statistics

In this section we present descriptive statistics for a set of selected countries. We present the data for the countries we will use in out later analysis, which are Latvia, Finland and Korea (our selection procedure will be explained later). Furthermore, we will show the data for several other countries located in diverse regions to highlight the variation between countries and over the years.

Our first graphs plots the average math and reading score against the variation in those scores. For every country we take the average score over all students across the years, meaning every observation is counted equally years with fewer observation are not assigned larger weights. We consider every student's math and reading scores, denoted by y^s_{ijt}, where s denotes the subject, meaning $s \in \{reading, math\}$ and then multiply these values with their standardised weights $\widetilde{\omega_{ijt}}$. We then take the average over all these scores for every country j and wave t and finally rescale by multiplying by the sum of the weights before standardisation to obtain the average weighted math/reading score μ^s_{jt}

$$\mu^s_{jt} = \left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\omega}_{ijt} \times y^s_{ijt}\right)\sum_{i=1}^{n} \omega_{ijt}. \tag{2}$$

We want to compare the average weighted score to the average weighted variation. We consider the average of the weighted sum of squared differences and then take the square root to obtain the standard deviation. Again, we do this per country and per wave

$$\sigma^s_{jt} = \sqrt{\sum_{i=1}^{n} \widetilde{\omega}_{ijt} \times (y^s_{ijt} - \mu^s_{jt})^2}. \tag{3}$$

Figure 6 shows that there is strong variation between countries for both values. For the weighted average score, Singapore and Korea score the highest, while developing countries such as Indonesia perform considerably worse. Unsuprisingly, higher income countries such as Korea and Finland have higher scores in both reading and math than lower income countries such as Indonesia and Brazil. There is also strong variation in the standard deviation where countries with higher scores generally have more variation in test scores.

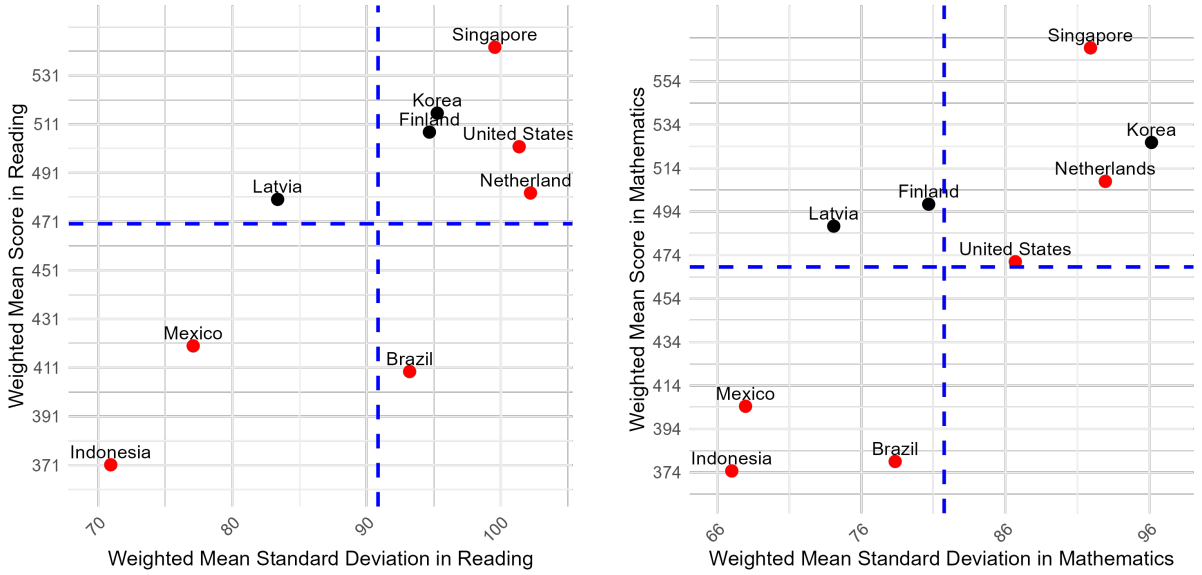


Figure 1: Average and Variance of Performance in Reading and Math

Since we consider the relation between the SSA and academic performance, it is also important to gauge the starting age across countries because the difference between starting at age 5 instead of 6 might not be the same as starting at age 6 instead of 7. Figure 2 shows that starting ages differ as much as 1.2 years across the ten countries we selected, indicating that there are considerable differences in policy (and perhaps policy compliance) across the countries in the sample. Overall, the heterogeneity in performance and policy between countries indicates it might be prudent to avoid pooling them to model the relation between students' characteristics and performance.

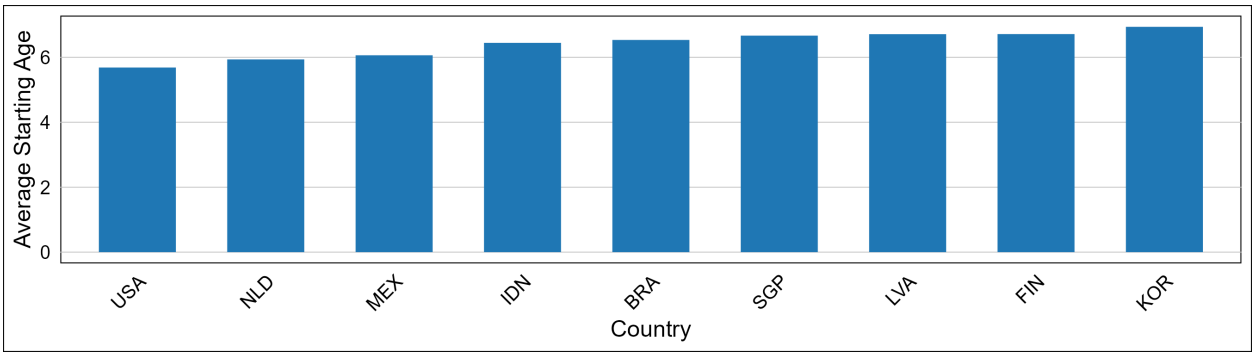


Figure 2: Average SSA across Countries.

There is also variance in how test scores have developed over time. We consider three years in which PISA scores where measured: 2015, 2018 and 2022. Countries generally perform consistently in both math and reading across, but for some countries there are some significant exceptions. For example, the Netherlands scored 503 in reading in 2015 and 459 in 2022. Scores generally increased between 2015 and 2018 and decreased from 2018 to 2022, likely caused by school closures due to the COVID pandemic, although the drop in scores differs considerably between countries. Overall the differences across time within countries are not very large, so we chose to pool country results across waves.

Figure 3: Average Performance in Reading and Math Over Time

Test performance also varies with other student characteristics. Figure 4 displays that girls outperform boys in reading on average but boys do better in math, though differences are not very large. The large differences between the 10th and 90th (denoted by the black and red dotted lines respectively) are indicative of large discrepancies between the best and worst students. Furthermore, there is slightly lower variance in performance for girls than for boys in both reading and math.

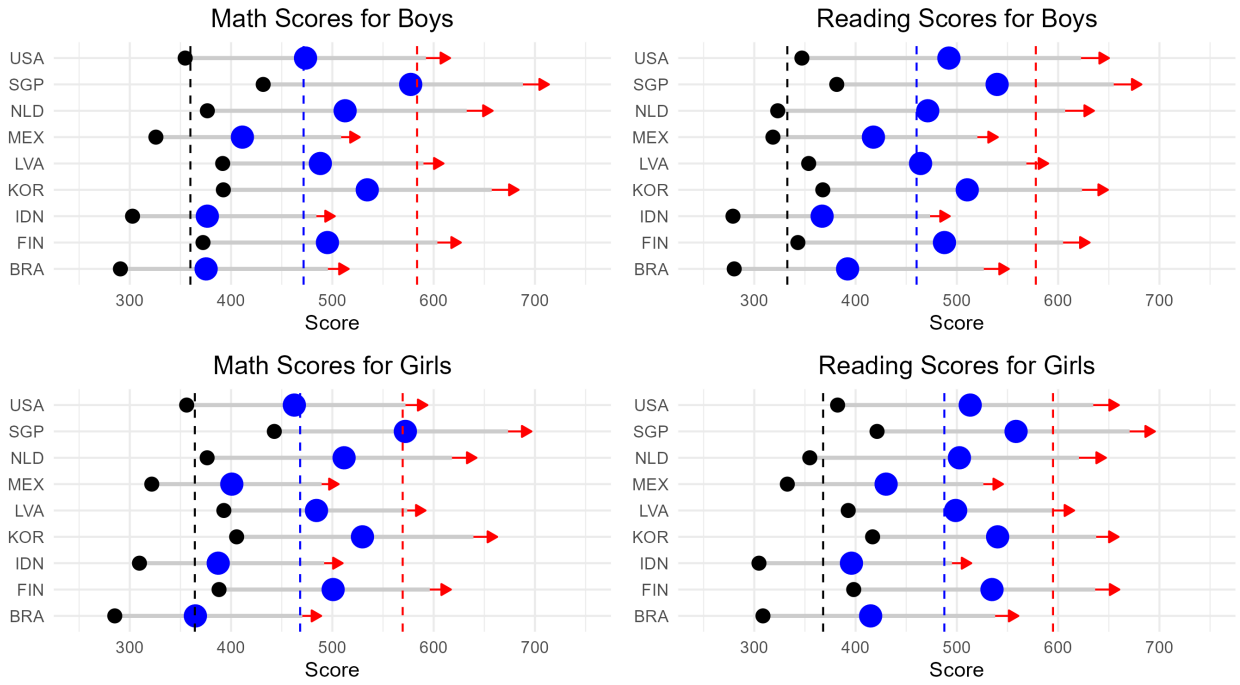


Figure 4: Gender and Average Performance in Reading and Math. The black dots represent the 10th percentile, the blue dots the 50th percentile, and the red arrows denote the 90th percentile of scores per country. The black, blue and red dotted lines denote the average 10th, 50th, and 90th percentile across countries in a plot.

Next, we consider class differences between students to study intergenerational mobility. We first consider the education level of the students' parents, proxied by the highest number of years that either parent has completed. The relation between the average reading and math scores and parents' level of education is clearly positive, as shown in Figure 5. The relation appears particularly strong for countries where students perform relatively well. The relations are not completely stable, with large spikes at times, though this might be due to underrepresentation of parents in certain categories. For example, the amount of parents in Finland that have completed 3 years of education or less is low. The bottom part of Figure 5 shows the the relation between home possessions, such as a desk, a quiet place to study, and books is very clearly positively related to student performance. Students in a higher decile for the variable, i.e., those that have more home possessions, perform better in both math and reading, with a monotonic and quite smooth relation.

Overall, the dependence between performance and student characteristics such as gender and parents' education implies it might be interesting to incorporate them in the models as interaction terms or covariates.

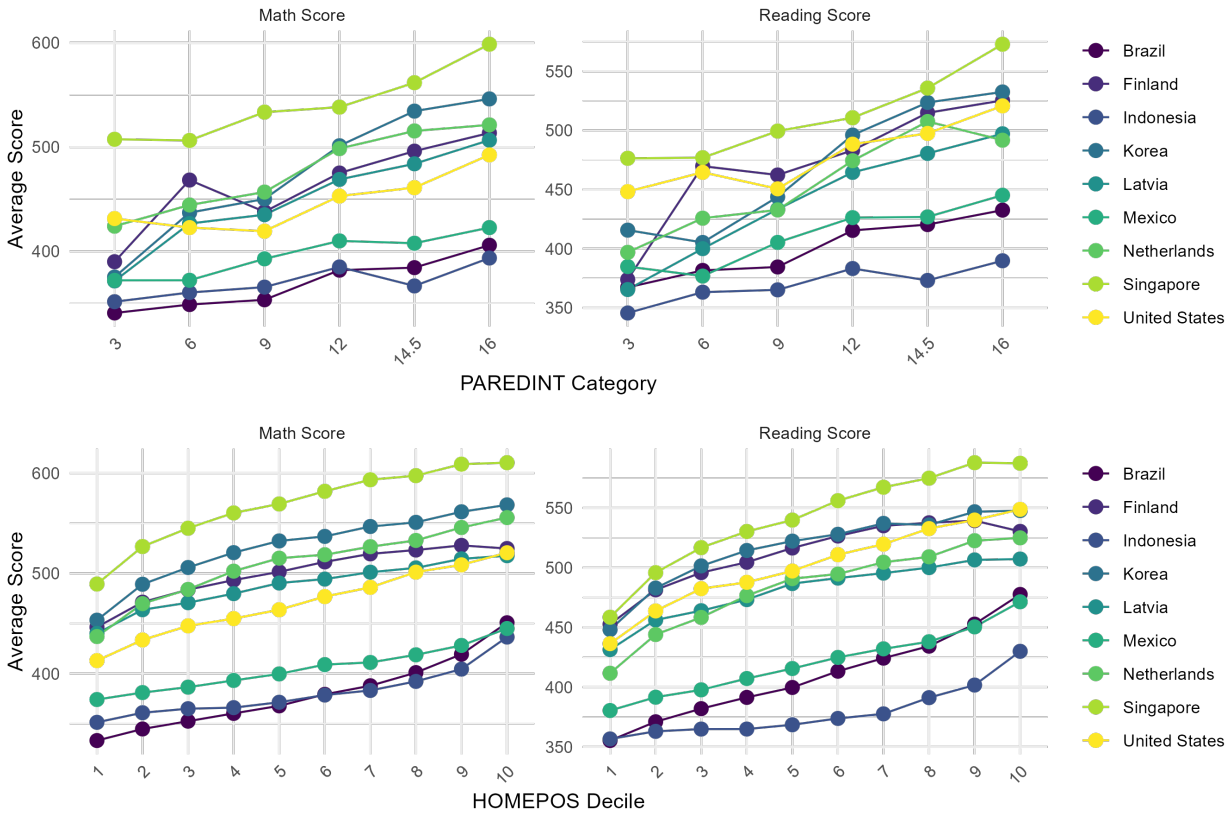


Figure 5: Intergenerational Mobility and Average Performance in Reading and Math. PAREDINT indicates the years of education received by the most educated parent. HOMEPOS indicates the income decile using home posessions as proxy

2.3 Correlation between SSA and educational outcomes

In this section we explore the correlation between the age at which students enter school and their math and reading scores. We did this by considering the average SSA per country and per year of testing. We performed weighted least squares (WLS) according to the standardised sample weights, as explained in section 2.1. In WLS, you attempt to minimize the following error.

$$\sum_{i=1}^{n}(\widetilde{\omega}_{ijt}(y_i - (\alpha + \beta\mu_{it}^{SSA})))^2. \tag{4}$$

Here, y_i^s denotes the score of student i in either math or reading, μ_{jt}^{SSA} the average SSA in country j and time of testing t , α denotes the intercept and β is the WLS coefficient.

The results for the regressions for both math and reading scores are shown in table 1. The coefficient for average starting age shows that starting a year later on average with schooling corresponds to a decrease of -22.37 and 38.33 in the math and reading scores respectively. Both coefficients are highly significant. This corresponds to a decrease in standard deviation of 0.4 and 0.71 for math and reading respectively.

Still, this number just gives an indication of the degree of correlation and says nothing about causation, as in this regression we do not control for student-characteristics of country fixed effects.

Table 1: Weighted least squares regression results

	Depedent Variable	
Variable	Average math score	Average reading score
Intercept	579.17***	683.08***
	(1.03)	(1.08)
Average starting age	-22.37***	-38.33***
	(0.17)	(0.17)
Observations	1631566	1631566

***:p<0.001, standard errors indicated in brackets

2.4 Country Selection

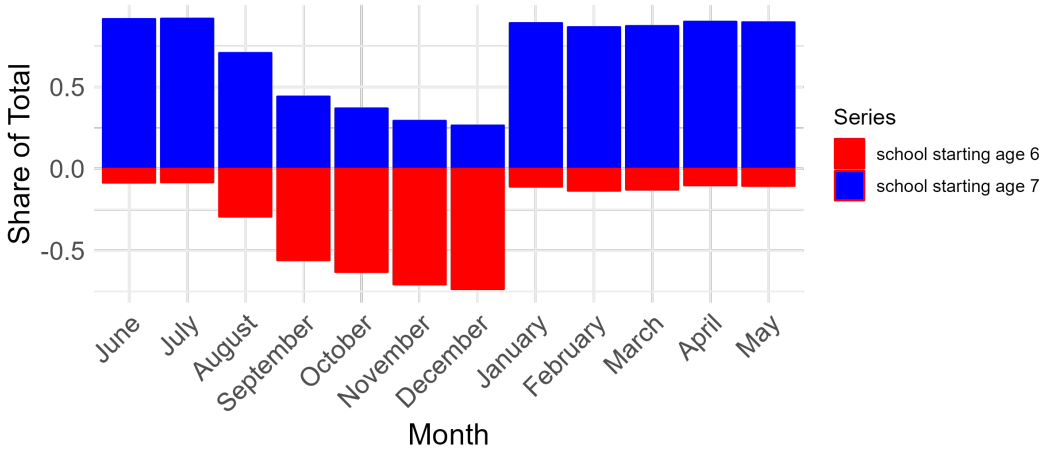
Facilitating a good causal inference exercise requires careful consideration of the circumstances. The compared groups must exhibit substantial differences in the treatment of interest but should be very similar in all other relevant characteristics. Since we focus on the effects of SSA on academic performance, it is essential to select countries with clearly defined school starting rules that can serve as the cutoff for our analysis. We select this point empirically by identifying a point where SSA significantly differs between students born before and after a certain date or policy change. It is also preferable for this policy to be consistent across the three analyzed PISA measurement waves to ensure the robustness and reliability of the results. Furthermore, determining the correct cutoff point is important for the validity of the regression discontinuity design. Such a cutoff ensures that units on both sides of the cutoff are comparable, minimizing the risk of confounding variables influencing the estimated treatment effects.

To select countries that meet these criteria, we group the data per country and measurement wave, and analyse the SSA for each month. We calculate the mode of SSA in each month, and select countries which have two unique modes over the course of the year, as there is potential for a clear cut-off when such a shift occurs. We drop all observations that have a SSA different from these two modes because we are interested in the difference starting school a single year earlier makes. Subsequently, we calculate the share of students that belongs to the higher of the two modes (old students), and choose countries for which there is a clear increase in this share. The month of this clear increase logically serves as the cut-off.

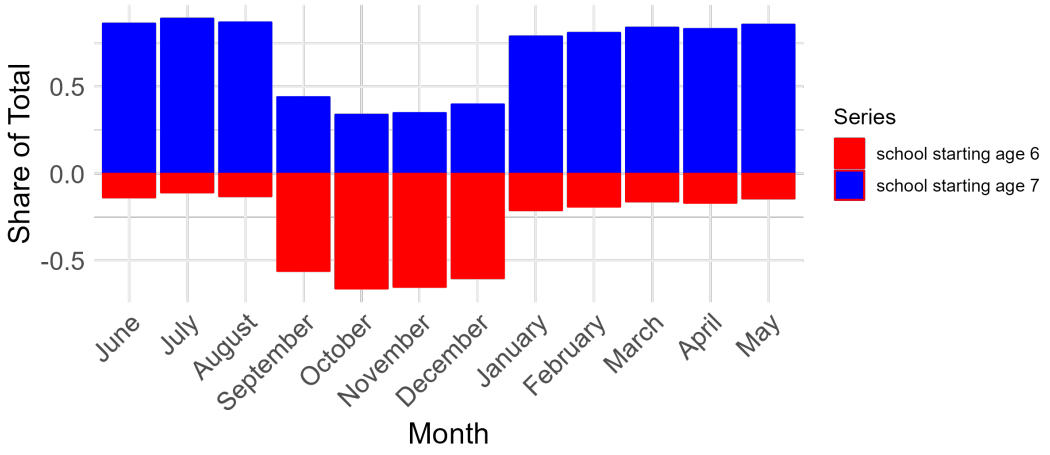
It is probably helpful to provide an example to illustrate the selection procedure. Panel a in Figure 6 presents the share of students with a SSA of 6 and 7 per birth month in Finland for the PISA measurement wave of 2022 (students with a SSA below 6 or above 7 are excluded from the analysis). The share of students with a SSA of 7 increases substantially between December and January, is stable between January and July, and decreases steadily between August and December. The PISA measurement waves of 2015 and 2018 in Finland contain very similar patterns as those observed in Figure 6. Accordingly, we judge Finland as a valid candidate for our analysis and identify December/January as the school starting rule.

Figure 6 also shows there is a significant increase in the share of students with a later SSA from

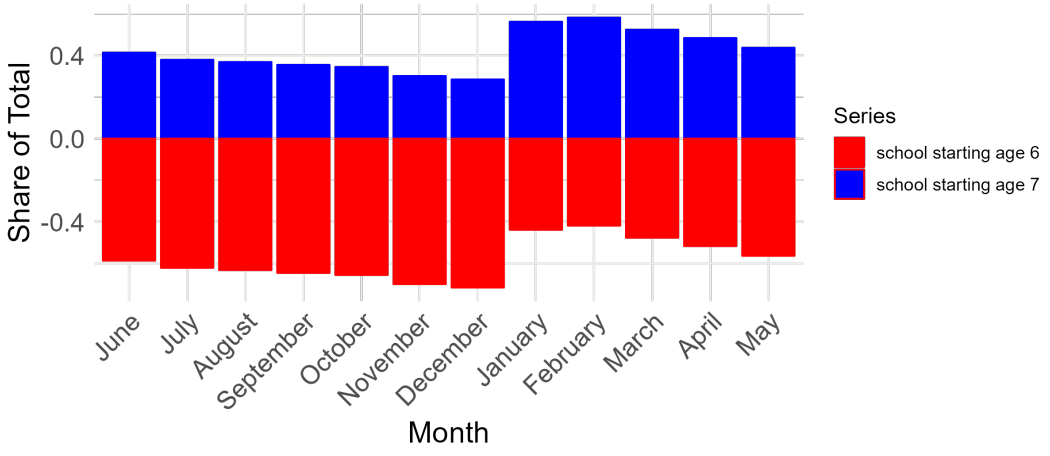
December to January for Latvia and South-Korea. Importantly, Finland, Latvia, and South-Korea all have a relatively high average SSA (see Figure 2) with modes of 6 and 7 across the months. The countries also have the same cut-off point in December/January. By selecting Finland, Latvia, and South-Korea, we obtain a group of countries that is fairly heterogeneous in terms of culture and student characteristics, but quite homogeneous in terms of policy.



(a) Share of students with a SSA of 6 and 7 per birth month in Finland



(b) Share of students with a SSA of 6 and 7 per birth month in Latvia



(c) Share of students with a SSA of 6 and 7 per birth month in Korea

Figure 6: The figure presents the relative share of students with a SSA of 6 and 7 for Finland, Latvia, and South-Korea. Students with a SSA below 6 or above 7 are excluded from the analysis.

3 Methodology

In this section, we will first focus on procedures to estimate country-specific effects of educational policies. In the end, we discuss a method for pooling the results and obtaining global treatment effects and corresponding assumptions.

3.1 Regression discontuinity (RD)

For the methodology, we will mostly focus on regression discontinuity (RD) design. This is currently one of the most popular research designs in multiple domains of science for doing causal inference and program evaluation in the case where a randomly assigned treatment is not available (Imbens and Lemieux, 2008; Cattaneo and Titiunik, 2022). In such designs, there is a particular cutoff rule for the treatment and under few assumptions, we can still identify the treatment effect. This is specifically relevant for this study as we expect that we students who have been born just before a particular date and just after are comparable in other unobserved characteristics. We will discuss several frameworks and extensions of the RD design in this section.

In general, RD designs involve three key elements: a score (also known as a running variable, index, or forcing variable), a cutoff, and a treatment rule. Every unit in the data is assigned a score. The treatment rule then uses a known cut-off point on the score to decide if a unit receives treatment or not. This setup creates a sharp change in the likelihood of receiving treatment right at the cutoff. Under the condition that units cannot influence their own score and do not differ in any unobserved characteristics, we can compare the outcomes of units just before the threshold and just after.

3.2 Sharp regression discontinuity design (SRD)

We will now elaborate on the most standard RD design which is the case where the cuttoff rule is a deterministic function of the score, that is, the cutoff rule is "sharp". In this paper, the cutoff rule (i.e. the school-starting-age) is based on the month of birth. We now introduce some notation to model this. Let Y_{ij} denote the test outcomes of student i in country j (either maths or reading). Next, let X_{ij} denote the month of birth that individual (i.e. the running variable). Now employing the SRD framework, we assume that the treatment variable is a deterministic function of the running variable. That is, we take the treatment variable to be

$$T_{ij} = \mathbb{I}\{X_{ij} \geq c_j\}.$$

Note that we allow c_j to differ per country as different countries have different starting months. Now formulating our framework in terms of the Rubin Causal Model (RCM) (Rubin, 1974), let $Y_i(0)$ and $Y_i(1)$ denote the pair of potential outcomes for unit i, where $Y_i(0)$ is the outcome without the treatment and $Y_i(1)$ is the outcome with the treatment. The main issue in causal inference is that we never observe both of these. We assume that the observed outcome is equal to one of these based on the treatment, so we write

$$Y_{ij} = (1 - T_{ij})Y_{ij}(0) + T_{ij}Y_{ij}(1).$$

The goal is now to identify and subsequently estimate the average causal effect of the treatment at the cutoff point for country j

$$\tau_{SRD,j} = \mathbb{E}(Y_{ij}(1) - Y_{ij}(0)|X_{ij} = c_j).$$

To do this, we will need identifying assumptions which we will discuss in the following subsection.

There is one important issue that we need to investigate first when applying the SRD framework. This is whether the cutoff rule is in practice a deterministic function of the month of birth. It could be that there are many units in the dataset that not follow this rule which could invalidate this approach. For this paper, this means that many children go to school too early or too late. In the literature, we then talk about a low "compliance" rate (Angrist and Imbens, 1995). However, it is known that most countries have high-compliance (Givord, 2020). Moreover, we specifically based our decision of countries in Section 2.4 based on a clear cutoff rule. Therefore, the SRD framework seems reasonable to employ for this study.

3.2.1 Identification restrictions

There are two common frameworks for analysis and interpretation of RD designs with corresponding identifying assumptions. These are known as the continuity framework and the local randomization framework. For now, we will focus on the continuity framework as this is the most standard and we elaborate on the assumptions belonging to this framework. We discuss the local randomization approach in Section 3.4.

Assumption C1 (*Continuity of conditional regression functions*)

$$\mathbb{E}(Y_{ij}(0)|X_{ij} = x) \quad \text{and} \quad \mathbb{E}(Y_{ij}(1)|X_{ij} = x)$$

are both continuous functions of x.

Assumption C2 (*Continuity of conditional distribution functions*)

$$F_{Y_{ij}(0)|X_{ij}}(y|x) \quad \text{and} \quad F_{Y_{ij}(1)|X_{ij}}(y|x)$$

are continuous functions in x *for all* y.

These assumptions are based on the idea that units just below and just above the cutoff c would have similar average outcomes if their treatment status were the same. Therefore, any observed difference in average outcomes between the treated and control groups at the cutoff is due to the treatment. This difference can be seen as the causal average effect of the treatment for units with a score of $X_{ij} = c_j$. Under these assumptions we can write

$$\tau_{SRD,j} = \mathbb{E}(Y_{ij}(1) - Y_{ij}(0)|X_{ij} = c_j) = \lim_{x \downarrow c_j} \mathbb{E}(Y_{ij}|X_{ij} = c_j) - \lim_{x \uparrow c_j} \mathbb{E}(Y_{ij}|X_{ij} = c_j). \tag{5}$$

This result now establishes that the SRD can be identified by the vertical distance between the conditional expectations just to the left $\lim_{x \uparrow c_j} \mathbb{E}(Y_{ij}|X_{ij} = c_j)$ and just to the right which is $\lim_{x \downarrow c_j} \mathbb{E}(Y_{ij}|X_{ij} = c_j)$. This gap can estimated from the data by estimating two regression models and then taking the difference of the intercepts at the cutoff value.

The above assumptions also immediately showcase when RD designs could fail. One such case is where the potential outcomes are not continuous functions of the running variable. We have a discrete running variable (month of birth) so this assumption is violated per definition (Kolesár and Rothe, 2018). In pratice, this might be a minor problem when the number of mass points is large and therefore approximates a continuous variable well. In our case, we have only 12 months so this seems problematic. For now, we ignore this problem but discuss a more appropriate approach in Section 3.4. Another possible violation of the above assumptions happens when units could strategically adjust their scores to qualify for their preferred treatment condition (Lee, 2008; McCrary, 2008). This behavior could create a sharp shift in both their observable and unobservable characteristics around the cutoff point, e.g., parents who illegaly alter the birth certificate of their children to make them go to school at more favorable time points. As also mentioned by Peña (2017), "The manipulation of student age—through redshirting, grade retention or selection into gestational seasons—could bias Ordinary Least Squares estimates of the effect of relative age." In such case (5) does not hold anymore and we cannot even identify the SRD.

3.2.2 Estimation

Estimating the SRD treatment effect involves estimating two regressions functions which is a standard nonparametric regression problem (Härdle, 1990). Nonetheless, there are two nonstandard features in this problem. The first is that we are only interested in the vlue of the regression function at a specific point. Additionaly, this point is a boundary point. These two issues make standard nonparametric kernel regression less effective. At boundary points, such estimators demonstrate a slower convergence rate compared to their performance at interior points. The standard approach to deal with this in the literature is by using local linear regression (Fan and Gijbels, 1996). This technique approximates the regression functions on both sides of the cutoff using weighted polynomial regressions, typically first or second order. The weights are calculated using a kernel function based on how close on how close each running variable is to the cutoff. Mathematically, this would mean that if we want to fit a polynomial of order p, using kernel K and bandwith h, the SRD can be found by fitting two weighted least-squares regression. If we want country-specific estimates, this boils down to the following two regressions

$$\widehat{\boldsymbol{\beta}}_{-j} = \arg\min_{b_0,\dots,b_p} \sum_{i=1}^{n} 1(X_{ij} < c)\big(Y_{ij} - b_0 - b_1(X_{ij}-c) - b_2(X_{ij}-c)^2 - \cdots - b_p(X_{ij}-c)^p\big)^2 K\left(\frac{X_{ij}-c}{b}\right)$$

$$\widehat{\boldsymbol{\beta}}_{+j} = \arg\min_{b_0,\dots,b_p} \sum_{i=1}^{n} 1(X_{ij} \geq c_j)\big(Y_{ij} - b_0 - b_1(X_{ij}-c) - b_2(X_{ij}-c)^2 - \cdots - b_p(X_{ij}-c)^p\big)^2 K\left(\frac{X_{ij}-c}{b}\right),$$

where $\widehat{\boldsymbol{\beta}}_{-j} = (\widehat{\beta}_{-j,0}, \widehat{\beta}_{-j,1}, \dots, \widehat{\beta}_{-j,p})'$ and $\widehat{\boldsymbol{\beta}}_{+j} = (\widehat{\beta}_{+j,0}, \widehat{\beta}_{+j,1}, \dots, \widehat{\beta}_{+j,p})'$ denote the least-squares estimates for the group to the left of the threshold and to the right of the threshold, respectively. The SRD treatment effect of country j, $\tau_{SRD,j}$ is calculated as the estimated vertical distance at the cutoff specifically, which is the difference in intercepts:

$$\widehat{\tau}_{SRD,j}(h) = \widehat{\beta}_{+j,0} - \widehat{\beta}_{-j,0},$$

where we assume the data is normalized so that $c_j = 0$ for all countries. Under typical assumptions, $\widehat{\tau}_{SRD,j(h)}$ gives a consistent estimate of $\tau_{SRD,j} = \mathbb{E}[Y_{ij}(1) - Y_{ij}(0)|X_{ij} = c_j]$. It is standard to use $\tau_{SRD,j}(h_{MSE})$, where h_{MSE} is taken to be the bandwidth that minimizes the mean squared error (MSE). This provides an estimator that is not only consistent but also MSE-optimal.

3.2.3 Inference

Although it is common to use bandwidht that minimize the MSE in the literature, one should be cautious when performing subsequent inference. When selecting the bandwidth based on minmization of the MSE, we are balancing the squared-bias and variance of the corresponding RD estimator. However, these often result in "large" bandwidth choices which to significant bias in the approximations. Consequently, the confidence intervals derived from such estimators for RD treatment effects might be unreliable, leading substantial undercoverage. This suggests that conventional confidence intervals might frequently incorrectly reject the null hypothesis of no treatment effect. To overcome this limitation, Calonico et al. (2014) propose new confidence intervals for RD treatment effects that are robust against the bias introduced by "large" bandwidths, such as those minimizing MSE. These improved intervals aim to provide more reliable and accurate inference in RD studies. The resulting intervals will have the following from

$$I_{robust} = \left(\tau_{SRD,j}(h_{MSE}) - \widehat{B}_j \pm 1.96\sqrt{\widehat{V}_j + \widehat{W}_j} \right)$$

where $\widehat{B}_j$ is the estimated bias correction, $\widehat{V}_j$ is the estimated variance and $\widehat{W}_j$ denotes the adjustment made in standard errors for country j. For the exact form of these estimators, we refer to original paper. Next to the standard non-robust confidence intervals, we also examine these robust confidence intervals for inference.

3.3 Fuzzy regression discontinuity design

In SRD, we treatment variable is a deterministic function of the running variable. However, in practice, this is often not the case and treatment assigned does not align with the treatment received. In this paper, this would either mean that units go to school before the threshold condition is satisfied or units would not go to school while the threshold condition is satisfied. One could think of many cases where this would be the case. Mathematically, this would mean that

$$\lim_{x \downarrow c} \mathbb{P}(T_{ij} = 1 | X_{ij} = x) - \lim_{x \uparrow c} \mathbb{P}(T_{ij} = 1 | X_{ij} = x) \neq 1,$$

that is, the jump in probability of being treated around the threshold is not equal to 1.

To accommodate for this, the fuzzy regression discontinuity (FRD) design has been developed. Under Assumptions 1 and 2 and an extra monotonicity condition that is similar to one in LATE framework by Angrist and Imbens (1995) one can then identify the FRD. However, to estimate this, one needs to know *both* whether someone was assigned treatment and whether someone complied to treatment. For our application, treatment assignment refers to being born in specific month, which we know in the dataset. However, in the dataset, there is no information on whether someone in fact complied to treatment, i.e. one went to school when one was supposed to. A naive approach would be to use the grade variable to check whether someone is in the right class and thereby a possible complier. However, it is possible that students start school too late but skip a class and thus end up in a class with students who complied and did not skip a class. In the same way, one could think about students starting too early and repeating a class. As we do not have information on treatment compliance, we do not consider FRD in this study.

3.4 Local randomization approach

In the continuity framework that was discussed before, we assumed that the running variable is a continuous random variable. Theoretically, the moment of birth can be regarded as such a variable. However, in our dataset, we only observe the month of birth. This means that our running variable is discrete and therefore the assumptions required assumptions in the continuity framework are violated. A more appropriate alternative for dealing with such a discrete running variable is to use local randomization (Cattaneo et al., 2015). This approach refines the model by treating it similarly to a randomized experiment around the cutoff. Specifically, it assumes that units within a small range around the cutoff are comparable. This mimics the idea of a random assignment to a treatment or a control group. Using this method, we only focus on units whose scores are within this narrow window around the cutoff. One important difference with the continuity approach is that we now assume that the value of the running variable is unrelated to the potential outcomes. For this study, we choose the smallest window possible which is a month before and a month after the threshold.

3.4.1 Identification restrictions

Let us properly outline the required assumptions for this framework. Firstly, let us denote $\mathcal{W}_j = [c_j - \omega, c_j + \omega]$ as the window of interest around the cutoff, where we set $\omega = 1$ for our study, meaning that we look at a window of two months. The assumptions we need are.

Assumption L1 *(Unconfoundedness and known treatment)*
The distribution of the running variable X_{ij} is unconfounded and distribution of the treatment assignment is known within $\mathcal{W}$.

Assumption L2 *(Constant outcomes within group)*
The potential outcomes are not influenced by the running variable within $\mathcal{W}$.

Under assumptions L1 and L2, for all students where X_{ij} falls within the window $\mathcal{W}$, whether they are placed above or below the cutoff is independent of their potential outcomes. Additionally, these potential outcomes do not depend on the running variable. As a result, the regression functions within this interval $\mathcal{W}$ remain constant and we can therefore estimate them simply by taking an average. Note that these assumptions might fail under the same scenario discussed for the continuity framework where the running variable is confounded. Moreover, we now have the extra assumption of constant potential outcomes within $\mathcal{W}$. This means students that we assume that students who have been born in this two month window do not differ in unobserved characteristics. This assumption is perhaps not entirely realistic, but as two months is still relatively short, we think it is a reasonable assumption.

3.4.2 Estimation and inference

In our paper, the treatment assignment mechanism is taken to be Bernoulli, i.e. we assume equal probabilities of being born one month before or one month after the threshold. This means we use the standard difference-in-means estimator to estimate the effect. Let $\theta_{SRD,j}$ denote the SRD

effect using local randomization for country j. Then we run the following regression

$$Y_{ij} = T_{ij}\theta_{SRD,j} + w'_{ij}\delta_j + \epsilon_{ij}$$

for all units in window $\mathcal{W}$ and where w_{ij} denote other possible regressors, δ_j denotes other regression parameters for country j and ϵ_{ij} denotes an error term. We use standard least-squares methods to obtain estimates and corresponding standard errors.

3.5 Conditional average treatment effect (CATE)

To detect possible heterogeneity among subgroups in our sample, we add interactions terms. These interactions terms allow us to investigate the conditional average treatment effect (CATE) for different characteristics. Investigating CATE is crucial because it allows us to understand how the impact of the treatment varies across different segments of the population. This differentiation can help in tailoring interventions more effectively and can provide deeper insights into the mechanisms behind the treatment effects.

We find the CATE by extending on the local randomization framework that was employed earlier. Let z_{ij} be a particular regressor of interest on which we want to condition for the treatment effect. We consider the following regression

$$Y_{ij} = T_{ij}\theta_{SRD,j} + T_{ij}z_{ij}\beta_j + w'_{ij}\delta_j + \epsilon_{ij},$$

where β_j denotes the coefficient corresponding to the interaction term. Under the assumptions outlined above and assuming no other regressors w_{ij}, we obtain the CATE using

$$\mathbb{E}(Y_i(1) - Y_i(0)|z_{ij} = z) = \theta_{SRD,j} + z\beta_j.$$

Therefore, we can estimate it by estimating $\theta_{SRD,j}$ and β_j using standard least-squares regression techniques.

3.6 Machine learning methods for mechanism detection

Our dataset contains a multitude of variables and a large number observations. One could think about many possible mechanisms driving the test scores of the students. It is likely that there are also many interacting effects between these variables and that these exhibit nonlinear patterns. An extremely popular and natural approach to capture such effects is by exploiting the power of machine learning models. Unlike traditional models that need clear rules about relationships, machine learning can find interactions and non-linear relationships as it allows for a highly flexible model specification. This allows it to reveal important but hidden patterns that simpler models might miss, making it invaluable for analyzing large and complex datasets like the PISA dataset.

For the purpose of this study, we will mainly focus on the eXtreme Gradient Boosting (XGBoost) model. This is a popular machine learning algorithm known for its efficiency and effectiveness in building predictive models, having been the winning model in many data science competitions (Chen and Guestrin, 2016). It is constructed by a sequence of decision trees, each one correcting errors made by the previous, through a technique called gradient boosting. Next to that, it

also incorporates regularization—techniques that help reduce overfitting (making the model too complex) to improve its performance on unseen data. It is also highly scalable and capable of handling large datasets and different datatypes, making it suitable for our problem.

Using XGBoost, we try to predict math and reading scores as accurately as possible. For the explanatory variables, we use the treatment variable and several other variables in our dataset such as home possession, gender, age, parental education and a few others. We then train the model parameters by employing cross-validation, which makes sure we avoid cross-fitting and the results generalize to other data. Finally, by adjusting parameters such as tree depth and learning rate, we fine-tune the model to achieve the best possible accuracy. As this model is able to capture complex interactions effects, we would like to have an idea about which interactions are driving these results and their corresponding strengths. It is known that machine learning models are harder to interpret than classical econometric methods, but a common way to assess these is by using Shapley values (Sundararajan and Najmi, 2020). In particular, this method decomposes a prediction into the sum of effects of each feature being introduced into the regression model.

3.7 Pooled average treatment effect (PATE)

For the last part of the methodology, we discuss how we deal with global / pooled treatment effects. The methods discussed above could just as well be applied to a pooled dataset and we only wrote it in terms of country-specific effects for generality. Furthermore, note that for all countries in our study, the cutoff date is from December to January and kids go to school either aged 6 or 7. This means that there can be no heterogeneity in terms of having the cutoff on a different moment or children being a different age. We now aim to estimate the pooled average treatment effect (PATE) of going to school one year later based on students who have been born just before New Year and just after and therefore start school at either age 6 or 7. However, the number of observations per country in our dataset is not the same. This could lead to inbalances when pooling the data. To guard against this, we reweight the data based on the number of observations per country. Therefore, we use another level of sampling weights on top of the weights from Section 2.1. This allows us to pool the data and use any of the methodology described above and interpret the effect as an average over three countries. When the effects are the same over all countries, i.e. it is homogeneous, this should lead to a more efficient estimate. This would be one of the reasons to also estimate a pooled model. Next to that, in case of heterogeneity across countries, it is of interest to see what the average effect is and especially how different effects add up or cancel each other.

4 Results

In this section, we first look at the causal effects of staring school later using the treatment effect estimators discussed above. Furthermore, we consider the CATE while conditioning on different explanatory variables. We select these explanatory variables using Shapley values resulting from the XGBoost model.

4.1 Causal effect of starting school later

Firstly, we are interested in the direction and strength of the effect of being born after the cut-off point. To study this, we first employ the estimation procedure for the SRD design under the standard continuity framework. One useful aspect about these estimates is that they can easily be displayed graphically by plotting both estimated regression functions to the left and right of the cutoff. Using first-order polynomials (that is $p = 1$), we have estimated these for all three countries separately in our dataset and also for the pooled dataset (with corresponding sampling weights). These results can be found in Figure 7 where the reading scores are in the left column of the plot and the maths scores in the right column.

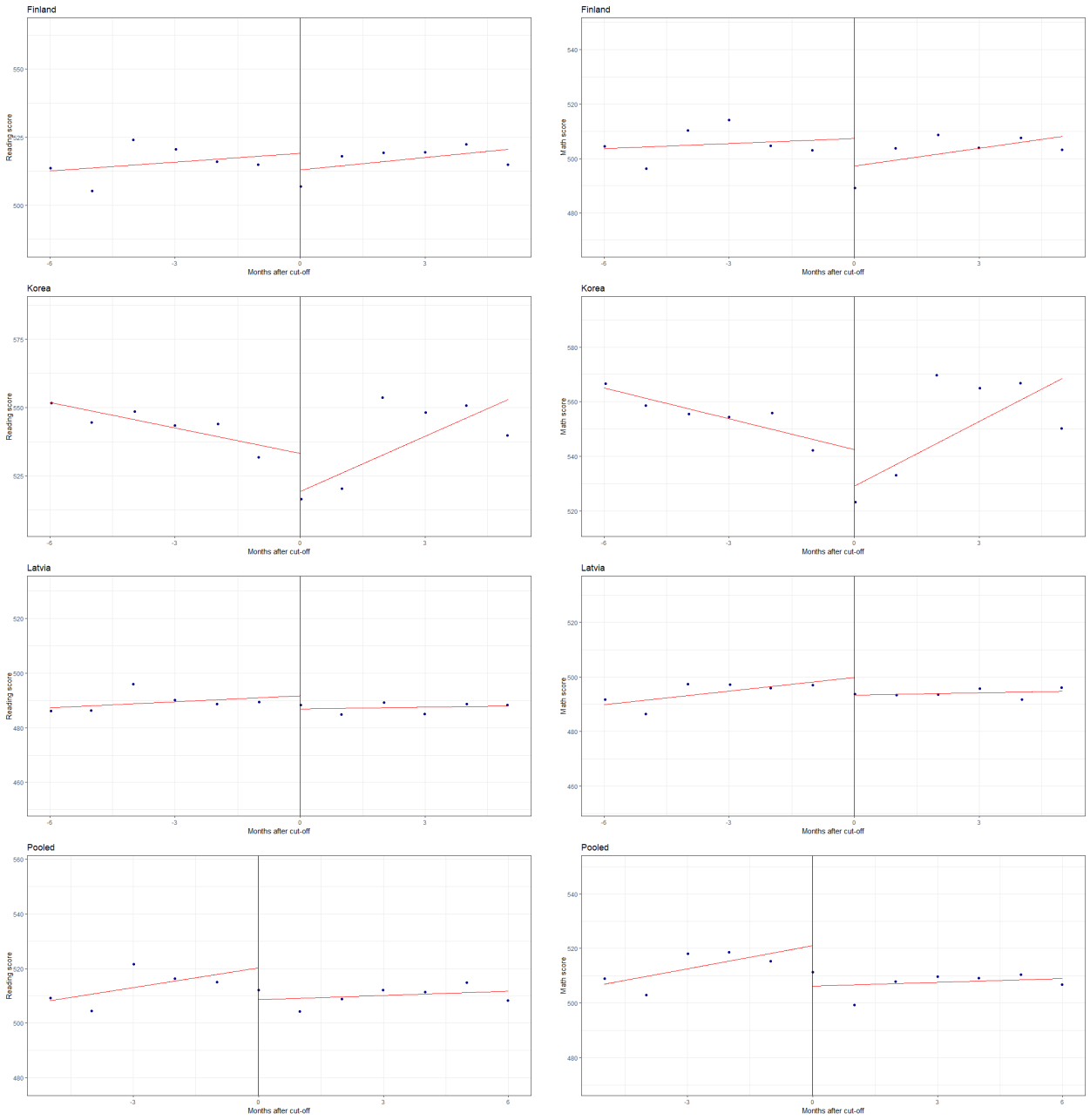


Figure 7: SRD design treatment effects for Reading and Math (left and right column, respectively) in Finland, Korea, Latvia (first three rows), and the three countries pooled together (last row).

We first consider only the country-specific effects which can be found in the first three rows of the figure. The first thing to notice is that all effects are negative. Students who have been born in January just after the threshold, start school one year later than their peers who have been born in December. These students who start school later also perform worse on both reading and writing when 15 years old. This demonstrates that, in these countries, starting school earlier is beneficial for later school outcomes. This might seem in contrast to Oosterbeek et al. (2021) who finds that older students do better in terms of academic performance. The difference is that they compare these students within a classroom where the age varies over students. In this study, we compare

students of the same age and find that students who started school later (and are therefore generally older than their peers) perform worse. Another difference is that they look at a country with a lower cut-off point, the Netherlands.

Moreover, although all effects have the same sign in all three countries, there is clearly some heterogeneity among them. Overall, the effects seem the weakest for Latvia, slightly stronger for Finland and the largest for Korea. Since the cut-off is at the same time and age for these countries, the difference is likely to be in the countries' characteristics. Next to this, we have also shown the PATE in the last row of Figure 7. As expected, the pooled effect is of the same sign and magnitude as the individual countries.

The difference in treatment effects across these countries already highlights the heterogeneity in treatment effects. In the next section, we attempt to leverage this heterogeneity to better understand what drives the observed causal effect.

4.2 Underlying mechanisms

To study the underlying mechanisms, we decompose the treatment effect on reading. We first look at subgroups within the regression discontinuity design. For this, we run a local linear regression. Then, we look at possible non-linearities in these mechanisms through XGBoost by estimating interaction effect between the treatment and other regressors. We will subsequently use these to estimate the CATE which lends itself for policy recommendations. It is important to note that these conditional expectations do not provide evidence for a causal mechanism beyond the RDD from the previous section. We merely use these different variables to form hypothesis on possible causes and suggest further investigation into these.

The results from the subgroup analysis are presented in 2. We investigate the subgroup of gender, and highly educated parents ($\geq$ 12 years). These can be seen as conditional average treatment effects, conditional on the subgroups. Model (2) shows that there is a big different in treatment effects between boys and girls. Boys tend to suffer from starting late, whereas girls do not. Model (3) shows that children with less educated parents suffer most from starting late. These insights are combined in Model (4) where we see that it is mostly boys from less educated families that suffer most from starting late within these countries.

A possible explanation for this finding is that people with higher educated parents have more learning opportunities at home. This combined with the hypothesized 'maturity gap' between young boys and girls might explain why this group suffers much. We further investigate heterogeneity by going beyond subgroups and looking at non-linear interactions.

Table 2: Subgroup analysis local linear regression

	Dependent variable:			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Treatment	-11.64***	-28.58***	-37.10***	-52.63***
	(2.54)	(3.26)	(3.54)	(4.03)
Treatment: Woman		29.47***		28.23***
		(3.59)		(3.56)
Treatment: High Educ Parent			38.19***	37.15***
			(3.76)	(3.73)
Intercept	540.27***	540.27***	540.27***	540.27***
	(1.80)	(1.78)	(1.78)	(1.77)
Observations	4,797	4,797	4,797	4,797
R^2	0.004	0.02	0.03	0.04
Adjusted R^2	0.004	0.02	0.02	0.04
Note:				*p<0.1; **p<0.05; ***p<0.01

The Shapley values that correspond to the interaction terms in the XGBoost model can be found in Figure 8 where they have been ordered in strength from left to right. Here, we firstly note that the strongest interaction effect comes from the home possession variable. This demonstrates that effect of starting school later is different for children coming from a wealthy background compared to children from poorer backgrounds. Next to that, we also observe a strong interaction between the treatment and gender, which shows that starting school later is also different for boys and girls.

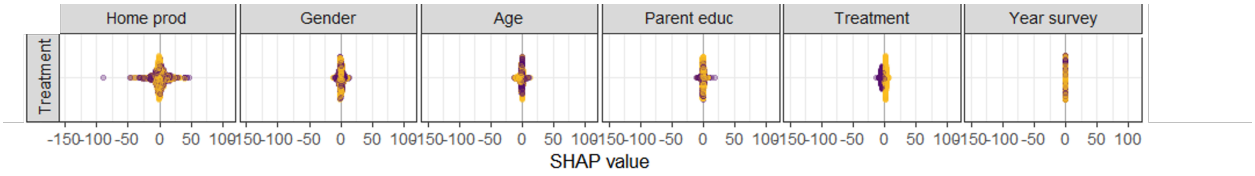


Figure 8: Shapley values corresponding to interactions between the treatment effect and other regressors ordered by absolute value from left to right.

5 Discussion and recommendations

In this paper, we investigated the causal effect of going to school later on subsequent educational outcomes. In particular, we used the PISA dataset which consists of test outcomes for several subjects for 15 year old students.

We find on average that starting school earlier increases educational performance in Finland, Latvia and South-Korea. This seems contradictory with the existing literature, which suggests that older children (who start school later) tend to perform better, but our analysis focuses on countries with relatively high starting ages, implying that the relation is not equal for children of younger and older ages. We also analyse the mechanisms underlying the impact of school-starting-age (SSA) on educational performance with subgroup analyses and by examining non-linear interaction effects. Our findings indicate that the beneficial effect of starting school earlier on educational performance mainly concerns boys, particularly those from families with lower levels of parental education and income. We hypothesize that parents from a lower socio-economic background are less able to provide their children with a suitable education in the absence of formal schooling than higher educated and wealthier parents.

On the basis of the results we recommend caution for countries considering to raise their SSA if their SSA is already relatively high, i.e., around or above 6 years of age, as it may be bad for students. The fact that a high starting ages particularly disadvantages children with a lower socio-economic background is concerning for equity reasons.

One important issue with RD designs is that they have limited external validity. This is because it only provides estimates for a subpopulation, that is, the part of the population with the score equal to the cutoff value. Therefore, our results do generalize to countries start primary school at a different age or countries that use different months as a threshold.

Another important limitation of this study is that we use several RD design estimators which only identify the treatment effect under particular assumptions. As has been mentioned before, the continuity assumption is per definition violated as we look at a discrete running variable with a small number of point masses. A more appropriate option is to use the local randomization approach here as this does not require continuity. Nonetheless, here we need to make the assumption that students born in a two month window do not differ in unobserved characteristics which is not completely realistic.

Furthermore, in this study we only considered the SRD design and not the FRD. In practice, we know that the compliance rate to these starting-school-age rules is not perfect. However, in the countries in our study we empirically found a relatively clear starting-school-age rule and therefore expect that the compliance rate in these countries is high which is also in line with the literature. Seeing how these results differ when using an FRD design is an interesting avenue for further study.

References

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

Bahrs, Michael and Mathias Schumann (2020). Unlucky to be young? the long-term effects of school starting age on smoking behavior and health. *Journal of Population Economics 33*(2), 555–600.

Balestra, Simone, Beatrix Eugster, and Helge Liebert (2020). Summer-born struggle: The effect of school starting age on health, education, and work. *Health economics 29*(5), 591–607.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes (2011). Too young to leave the nest? the effects of school starting age. *The Review of Economics and Statistics 93*(2), 455–467.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Cattaneo, Matias D, Brigham R Frandsen, and Rocio Titiunik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference 3*(1), 1–24.

Cattaneo, Matias D and Rocio Titiunik (2022). Regression discontinuity designs. *Annual Review of Economics 14*, 821–851.

Chen, Tianqi and Carlos Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Cook, Philip J and Songman Kang (2020). Girls to the front: How redshirting and test-score gaps are affected by a change in the school-entry cut date. *Economics of Education Review 76*, 101968.

Fan, Jianqing and Irene Gijbels (1996). *Local polynomial modelling and its applications*. Chapman Hall.

Givord, Pauline (2020). How a student's month of birth is linked to performance at school: New evidence from pisa.

Härdle, Wolfgang (1990). *Applied nonparametric regression*. Number 19. Cambridge university press.

Imbens, Guido W and Thomas Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142*(2), 615–635.

Kolesár, Michal and Christoph Rothe (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review 108*(8), 2277–2304.

Lee, David S (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics 142*(2), 675–697.

McCrary, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics 142*(2), 698–714.

Oosterbeek, Hessel, Simon ter Meulen, and Bas van Der Klaauw (2021). Long-term effects of school-starting-age rules. *Economics of Education Review 84*, 102144.

Peña, Pablo A (2017). Creating winners and losers: Date of birth, relative age in school, and outcomes in childhood and adulthood. *Economics of Education Review 56*, 152–176.

Rubin, Donald B (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Sundararajan, Mukund and Amir Najmi (2020). The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR.

Valdés, Manuel T (2023). The effect of the month of birth on academic achievement: heterogeneity by social origin and gender. *European Societies*, 1–27.

Whitebread, David (2013, Sep). School starting age: The evidence.

Zhang, Shiying, Ruoyu Zhong, and Junchao Zhang (2017). School starting age and academic achievement: Evidence from china's junior high schools. *China Economic Review 44*, 343–354.