Evaluating Educational Policies in Practice

Vedant Bhardwaj, María Valkov,
Alejandro Puerta-Cuartas

Abstract

This chapter evaluates two education policies using empirical methods. First, it examines the Texas Top Ten Percent (TTP) rule, which increased access to selective universities for high-performing students from underrepresented schools while displacing others. "Pulled In" students benefited from higher enrollment and graduation rates, while "Pushed Out" students compensated by enrolling in other institutions. Second, using regression discontinuity, the chapter analyzes the effect of school-starting age, finding that older starters generally perform better in standardized tests. The findings highlight policy trade-offs and underscore the importance of rigorous evaluation to ensure equitable and effective educational interventions.
Keywords: Selective college admissions, Education policy evaluation, Labor market outcomes, Regression discontinuity design (RDD), School-starting age, Higher education access.

1 The effect of accessing selective colleges on education and labor market outcomes

There is an inherent trade-off associated with selective college admission. On the one hand, admitting a student to a high-quality institution can be beneficial, as it can improve student graduation and earnings outcomes (Hoekstra, 2009; Zimmerman, 2014), and boost the economic outcomes of low-income and underrepresented racial minorities (Kozakowski, 2020). On the other hand, admitting one student means rejecting another, so selective rules can be detrimental to those rejected, as they lose access to the benefits of attending a high-quality institution. Accordingly, evaluating an admissions policy change requires assessing its positive and negative impacts.

In this section, we replicate the results in Black et al. (2023), who study the effect of the Texas Top Ten Percent rule- hereafter "TTP"— that guaranteed admission to any Texas public university to anyone in the top ten percent of their high school class. The introduction of this policy had positive and negative effects arising from the enrollment dynamics of the Texas higher education system. While it increased the chances for top students from disadvantaged high schools, it also reduced access to the flagship campuses for some students from "feeder" high schools who were not in the top ten percent. Accordingly, the authors exploit the introduction of TTP to identify the positive and negative effects of access to a selective institution on education and labor market outcomes.

As previously discussed educational policies benefit some students, while potentially harming others (Pushed Out students). The authors find positive effects of the TTP on the benefited students, for whom they find an improvement in enrollment and graduation, while earnings either stay the same or increase. Furthermore, they find no substantial harm to the students that could have been negatively affected by the educational policy.

We first briefly describe institutional details on the Texas higher education system and the Top Ten Percent rule. We then present the setting the authors use to evaluate the educational policy. Subsequently, we describe the available data, and finally, we present their empirical strategy and main results.

1.1 The Texas Higher Education System and the Introduction of the Top Ten Percent rule

Texas has a large public higher education system consisting of over 30 four-year universities and over 60 two-year colleges. Among them, the two most selective institutions are the University of Texas at Austin (UT) and Texas A&M University, which constitute the focus of analysis of Black et al.

(2023). Most students attend public universities since only 18.8 percent of Texas students attend private colleges, and 9.8 percent attend out-of-state colleges.

Until 1997, admission to UT was mainly determined by continuous high school class rank and SAT or ACT scores, with affirmative action preferences for students from underrepresented groups. The SAT is a standardized test widely used for college admissions in the United States. It is a multiple-choice, computer-based test created and administered by the College Board. Conversely, the ACT is a curriculum-based education and career planning test for high school students that evaluates the mastery of college readiness standards.

In May 1997, the Texas legislature created the TTP, which guaranteed admission to any public university in Texas for students in the top decile of their high school class. The student's performance is calculated by each high school and measured at the end of the junior year. Except for the UT, the vast majority of top decile students would have been admitted in the absence of the TTP rule. Thus, this educational policy mainly concerned admission to the UT. For students outside the top decile, an Academic Index (AI), and Personal Achievement Index (PAI) were used as an admission rule. However, given the large number of top-ten-percent applicants, UT admissions were very competitive for non-TTP applicants. Accordingly, the TTP rule ensured entrance to high-performing students who otherwise would not have been admitted to the UT, while raising the bar considerably, albeit without racial preferences, for the lower-performing students that would have been admitted before the law.

The main purpose of this educational policy was to maintain diversity among students without the need to explicitly consider race in the admission process. The rationale is that basing admissions on SAT or ACT scores disproportionately favored students from high-income, and primarily white high schools. Moreover, explicitly considering race in admission decisions, permitted applicants to take advantage of the substantial racial and economic segregation.

Admission by high school yields a more diverse class than statewide admissions. In their sample, While the average student before the TTP attends a school that is 30 percent Hispanic and 12 percent Black, the average Hispanic student attends a school that is 60 percent Hispanic, and the average Black student attends a school that is 37 percent Black. Furthermore, although 22 percent of students get free or reduced-price lunches, the average student who receives free lunches goes to a school where this percentage is 42 percent. Thus, if the top decile of each high school is demographically representative of the school as a whole, admission by high school yields a more diverse class than statewide admissions, due to the presence of segregation across high schools.

The TTP had a substantial impact on the students' perceived admission chances. Since students had information regarding their high school rank, they could assess their prospects more easily than in the previous regime. This was enhanced by the fact that the TTP law mandated that every high school post a sign explaining the law and that a letter be sent to every parent of a qualifying student. Furthermore, this policy was widely covered in the media.

The top ten percent rule transformed the enrolment dynamics of the most selective campus in Texas. In the year of its introduction, 41 percent of freshmen from Texas high schools were admitted under the TTP. This number rose to 70 percent by 2003. More importantly, the composition of the admitted students became less concentrated at traditional feeder schools. Therefore, this educational policy gave access to high-performing students who, before the new regime, would not have been admitted (or even applied) to the UT.

1.2 The setting

Our object of interest is the effect of the TTP on students' education and labor market outcomes. Due to the inherent trade-off associated with selective college admission, we are interested in both TTP's positive and negative effects. While some students were "Pulled In" by this educational policy since they became more likely to attend UT Austin, others were "Pushed Out", as they became less likely to attend a high-quality institution. To estimate the effect of the TTP, the authors compare these two treatment groups with a control group.

Figure 1 illustrates how the "Pulled In", "Pushed Out", and control groups were defined. The left panel depicts that, before the TTP, students admitted to the two most selective campuses were the top students from feeder schools. After the TTP (see the right panel in Figure 1), the students in the top ten percent of a school that had traditionally sent few if any students to the University of Texas flagship campus in Austin were "Pulled In". Conversely, students from "feeder" schools that do not qualify for the top ten percent, but performed above average in their schools were "Pushed Out" .

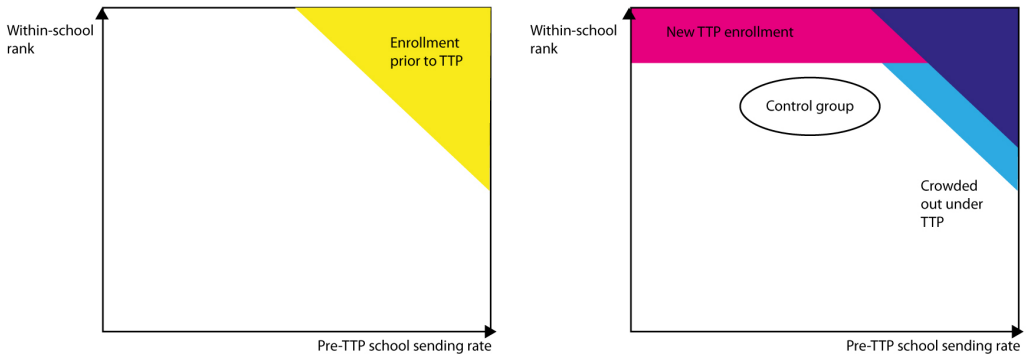


Figure 1: Schematic description of Top Ten Percent plan effect on enrollment

Finally, the control group consists of students who were above average in achievement but unlikely to be admitted to UT Austin under both regimes.

Ideally, we would like to know what the impact of the TTP for each student who was either pulled in or pushed out by the policy. Setting the time when the TTP took place at $t = 0$, we can express these individual effects as

$$\tau_{i,t}^{PI} = Y_{i,t}^{PI}(1) - Y_{i,t}(0), \quad t > 0, \tag{1}$$

$$\tau_{i,t}^{PO} = Y_{i,t}^{PO}(1) - Y_{i,t}(0), \quad t > 0, \tag{2}$$

where $Y_{i,t}^{PI}(1)$ is an education or labor market outcome at time t for a student i who was pulled in by the TTP, $Y_{i,t}(0)$ is the same education or labor market outcome for the student if the TTP had not taken place, and $Y_{i,t}^{PO}(1)$ is the outcome for the $i-$th pulled out student at time t. Accordingly, we can express the observed outcomes for the pulled in students as $Y_{i,t} = Y_{i,t}^{PI}(1)\text{PulledIn}_i + Y_{i,t}(0)\,(1 - \text{PulledIn}_{i,t})$, and those of the pulled out students as $Y_{i,t} = Y_{i,t}^{PO}(1)\text{PushedOut}_{i,t} + Y_{i,t}(0)\,(1 - \text{PushedOut}_{i,t})$, where $\text{PulledIn}_{i,t} = 1$ for the students that were "Pulled In" by the TTP and 0 for the control group, and $\text{PushedOut}_{i,t} = 1$ for the students that were "Pushed Out" by the TTP and 0 for the control group.

In our context, $\tau_{i,t}^{PI}$ and $\tau_{i,t}^{PO}$ provide the causal effect of the TTP at time $t > 0$ on the $i-$th student who was pulled in and pushed out by the educational policy, respectively. Thus, the former object provides the positive effect of the TTP, while the latter its negative effect.

The main limitation to estimate equations (1) and (2) is that we do not observe the counterfactual $Y_i(0)$ for any student affected (treated) by the TTP. In other words, we do not observe the education or labor market outcome if the TTP had not taken place for any student who was either pulled in or pushed out by this policy. However, we observe a control group, whose likelihood of attending a high-quality campus remained unchanged with the TTP. Accordingly, we can exploit this information to estimate the average treatment on the treated (ATT) of the TTP at time t, which can be expressed as

$$\tau_t^{PI} := \mathbb{E}\left[Y_{i,t}^{PI}(1) - Y_{i,t}(0) | \text{PulledIn}_{i,t} = 1\right], \quad t > 0, \tag{3}$$

$$\tau_t^{PO} := \mathbb{E}\left[Y_{i,t}^{PO}(1) - Y_{i,t}(0) | \text{PushedOut} = 1\right] \quad t > 0. \tag{4}$$

On the one hand, τ_t^{PI} measures how much the pulled in students' outcomes changed on average due to the TTP.[1] On the other hand, τ_t^{PO} measures how much the pushed out students' outcomes changed on average because of the TTP. Accordingly, we will henceforth refer to τ_t^{PI} and τ_t^{PO} as the causal positive and negative effects of the top ten percent rule, respectively.

A natural approach to estimate these causal effects would be to compare the average outcomes of

[1]For instance, if Y is post-college earnings, τ_t^{PI} corresponds to the average change in post-college earnings of the students who were pulled in by the educational policy.

pulled in and pushed out students to those of the control group in the subsequent years

$$\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right], \quad t>0, \tag{5}$$

$$\mathbb{E}\left[Y_{i,t}|\text{PushedOut}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PushedOut}_{i,t}=0\right], \quad t>0. \tag{6}$$

For instance, if our object of interest (Y) is post-college earnings, equation (5) corresponds to the difference between average post-college earnings of the pulled in students and the average of the control group. However, as we now show, equations (5) and (6) do not identify the causal effects of the TTP. In other words, they are not equivalent to τ_t^{PI} and τ_t^{PO}

$$\tau_t^{PI}=\mathbb{E}\left[Y_{i,t}^{PI}(1)-Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]\neq\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]$$
$$\tau^{PO}=\mathbb{E}\left[Y_{i,t}^{PO}(1)-Y_{i,t}(0)\text{PushedOut}_{i,t}=1\right]\neq\mathbb{E}\left[Y_{i,t}|\text{PushedOut}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PushedOut}_{i,t}=0\right].$$

To see why, consider the case of the pulled in students

$$\begin{aligned}
&\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]\\
=&\mathbb{E}\left[Y_{i,t}^{PI}(1)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]\\
=&\mathbb{E}\left[Y_{i,t}^{PI}(1)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]\\
+&\mathbb{E}\left[Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]\\
=&\mathbb{E}\left[Y_{i,t}^{PI}(1)-Y_i(0)|\text{PulledIn}_{i,t}=1\right]+\mathbb{E}\left[Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]\\
=&\tau_t^{PI}+\underbrace{\mathbb{E}\left[Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]}_{\text{Selection Bias}},
\end{aligned} \tag{7}$$

where in the first equality we have used the definition of $Y_{i,t}=Y_{i,t}^{PI}(1)\text{PulledIn}_{i,t}+Y_{i,t}(0)\left(1-\text{PulledIn}_{i,t}\right)$, and in the second we have added and subtracted the term $\mathbb{E}\left[Y_{i,t}(0)|\text{PulledIn}_{i,t}=1\right]$.

Equation (7) shows that the difference in average outcomes of the pulled in and control group students in a subsequent year ($\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t}|\text{PulledIn}_{i,t}=0\right]$) can be expressed as the sum of two objects, i.e., the average treatment on the treated (ATT) and the selection bias. While the former is our object of interest, the latter term captures that students who were pulled in by the TTP are systematically different from those who were not.

The selection bias in equation (7) would be zero if

$$\mathbb{E}\left[Y_i(0)|\text{PulledIn}_{i,t}=1\right]=\mathbb{E}\left[Y_i|\text{PulledIn}_{i,t}=0\right],$$

that is, if the average counterfactual outcome[2] of the pulled in students were the same as that of the students in the control group. In our framework, the treated students are the top performers in schools that had traditionally sent few if any students to the University of Texas flagship campus in Austin. Conversely, the control group consists of students who were above average in achievement who studied at high school with a high sending rate. Furthermore, these two groups of students are systematically different since Texas is highly segregated based on socioeconomic status. Since both academic performance and socioeconomic status affect expected earnings, one would expect that if the TTP had not taken place, the average educational and labor market outcomes of the treated students would be different from those of the control group. Thus, the selection bias is different from zero, so we cannot estimate the causal effects by comparing the outcomes of pulled in and pushed out students to those of the control group (see equation (7)). This gives rise to the question of how can we estimate the causal effects of the TTP? The answer is making reliable assumptions.

Assumptions are what allow us to learn something from the data. For instance, if we assume that the counterfactual average outcome of the pulled in students is the same as that of the students in the control group, the selection bias is zero. Thus, we could estimate the causal effects of the TTP by comparing the outcomes of pulled in and pushed out students to those of the control group. However, we have argued that this assumption is not plausible in our setting.

What reliable assumption can we make that allows us to estimate the causal effects of the TTP? In our framework, we can assume that outcomes for the three groups of students (pulled in, pushed

[2]The average counterfactual outcome refers to the average outcome of the pulled in students if the TTP would not have taken place.

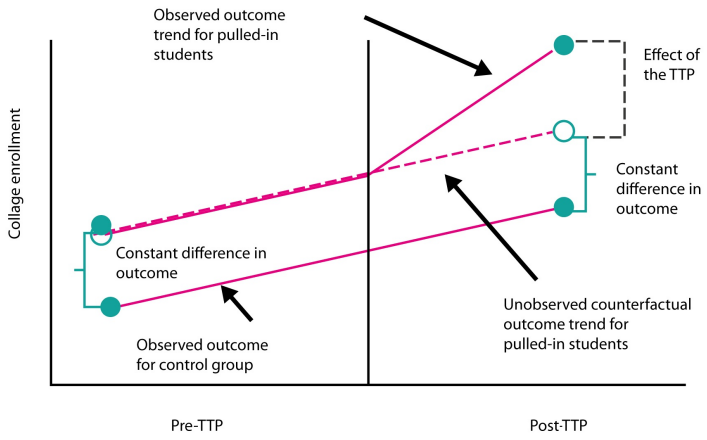


Figure 2: The effect of the TTP Under the Parallel Trends Assumption

out, and control) would have evolved similarly between the 1996 and 2002 cohorts had admissions policies been held stable. This is called a parallel trend assumption. We now illustrate the meaning and implication of this assumption in our framework, and how it allows us to estimate the causal effects of the TTP.

Figure 2 illustrates the effect of the TTP on college enrollment under the parallel trends assumption. The solid red and green lines depict the evolution of college enrollment before and after the TTP of the pulled in students and the control group, respectively. The dotted red line indicates the unobserved counterfactual college enrollment for the pulled in students. In other words, the outcome of these students if the TTP would not have taken place. According to Figure 2 and equation (3), the causal effect of the intervention (average treatment on the treated) corresponds to the difference between the observed college enrollment of the pulled in students (A) minus its counterfactual outcome (B). However, we do not observe B, so we cannot estimate the causal effect without further assumptions.

The parallel trends assumption states that outcomes for the pulled in and control groups would have evolved similarly between the 1996 and 2002 cohorts had the TTP would not have been implemented. This can be formalized as

$$\mathbb{E}\left[Y_{i,2002}(0)|\text{PulledIn}_{i,t}=1\right]-\mathbb{E}\left[Y_{i,1996}|\text{PulledIn}_{i,t}=1\right]=\mathbb{E}\left[Y_{i,2002}|\text{PulledIn}_{i,t}=0\right]-\mathbb{E}\left[Y_{i,1996}|\text{PulledIn}_{i,t}=0\right]. \quad (8)$$

To see how this assumption allows us to express the counterfactual outcome B in terms of observables, notice that according to Figure 2, point A corresponds to the expected value of college enrollment of pulled in students evaluated at time $t = 2002$, which can be expressed as $A = \mathbb{E}\left[Y^{PI}_{i,2002}(1)|\text{PulledIn}_{i,t}=1\right]$. Similarly, we have

$$\begin{aligned}
B &= \mathbb{E}\left[Y_{i,2002}(0)|\text{PulledIn}_{i,t}=1\right]\\
C &= \mathbb{E}\left[Y_{i,2002}|\text{PulledIn}_{i,t}=0\right]\\
D &= \mathbb{E}\left[Y_{i,1996}|\text{PulledIn}_{i,t}=1\right]\\
E &= \mathbb{E}\left[Y_{i,1996}|\text{PulledIn}_{i,t}=0\right].
\end{aligned}$$

Thus, equation (8) can be expressed as

$$\begin{aligned}
B-D &= C-E\\
B-C &= D-E,
\end{aligned}$$

which corresponds to the constant difference in outcomes in Figure 2. Finally, the last display can be expressed as

$$B = C+D-E,$$

so that the counterfactual outcome (B) can be expressed in terms of objects we observe, which are the college enrollment of the control group in 2002 and 1996 (C and E, respectively), and the college enrollment of the pushed in students in 1996 (D). Thus, the parallel trends assumption allows us to write what we do not observe (B) in terms of objects we do observe (C, D, and E), so the causal effect of the TPP (A-B) can be computed with observables (A-(C+(D-E)).

We have illustrated that under the parallel trends assumption, the positive causal effect of the TTP in equation (3) can be computed (is identified) as

$$\begin{aligned}
\tau_t^{PI} &\coloneqq \mathbb{E}\left[Y_{i,t}^{PI}(1) - Y_{i,t}(0) | \text{PulledIn}_{i,t} = 1\right] \\
&= \mathbb{E}\left[Y_{i,t}^{PI}(1) | \text{PulledIn}_{i,t} = 1\right] - \mathbb{E}\left[Y_{i,t}(0) | \text{PulledIn}_{i,t} = 1\right] \\
&= A - B \\
&= A - (C + (D - E) \\
&= \mathbb{E}\left[Y_{i,2002}^{PI}(1) | \text{PulledIn}_{i,t} = 1\right] - \mathbb{E}\left[Y_{i,2002} | \text{PulledIn}_{i,t} = 0\right] \\
&- \mathbb{E}\left[Y_{i,1996} | \text{PulledIn}_{i,t} = 1\right] + \mathbb{E}\left[Y_{i,1996} | \text{PulledIn}_{i,t} = 0\right]. \qquad (9)
\end{aligned}$$

Similarly, the ATT for the pushed out is identified as

$$\begin{aligned}
\tau_t^{PO} &\coloneqq \mathbb{E}\left[Y_{i,t}^{PI}(1) - Y_{i,t}(0) | \text{PushedOut}_{i,t} = 1\right] \\
&= \mathbb{E}\left[Y_{i,2002}^{PI}(1) | \text{PushedOut}_{i,t} = 1\right] - \mathbb{E}\left[Y_{i,2002} | \text{PushedOut}_{i,t} = 0\right] \\
&- \mathbb{E}\left[Y_{i,1996} | \text{PushedOut}_{i,t} = 1\right] + \mathbb{E}\left[Y_{i,1996} | \text{PushedOut}_{i,t} = 0\right]. \qquad (10)
\end{aligned}$$

The main challenge to estimate τ_t^{PI} and τ_t^{PO}, is identifying which students were pulled in and pushed out by the policy. This arises from the fact that the student's class rank is not measured in the statewide pre-TTP data and only limited information is available after TTP. Thus, we now turn to describe the ideal data one would require to estimate the causal effects of the TTP, how the available data differs from it, and the challenges this issue poses to assessing the causal effect of the TTP.

1.3 Data

To evaluate the effect of a policy, practitioners need information before and after its introduction. In the case of the TTP, the ideal dataset would consist of the universe of students who graduated from Texas public high schools before and after 1997. The more years available for each pre- and post-treatment period, the better. In particular, data availability for several years after the policy took place allows the assessment of the policy's short and long-term effects. Moreover, information on students for several years before the implementation of the policy provides a solid benchmark and accounts for potential confounders. For instance, if some other policy was implemented the year before the TTP, its impact could be underestimated.

To estimate the effect of the TTP, the ideal dataset set will comprehend information on class rank or high school GPA before and after this educational policy took place. This way, we would identify the pulled in and pushed out students, along with the students serving as a control group. Furthermore, since the objective is to assess the effect on education and labor market outcomes, the ideal dataset will comprehend information on them as well.

The available data usually differs from the ideal one, which poses a significant challenge in assessing the effect of policies. In this case, the main data limitation is that the student's class rank was not measured in the statewide pre-TTP data and only limited information is available after TTP. Thus, the authors can not deterministically identify the pulled in and pushed out students. We now turn to describe the data available to the authors for estimating the impact of the TTP, and how they circumvent not observing the ideal one.

The authors draw from administrative data covering the entire State of Texas, which consists of students who graduated from Texas public high schools between 1996 and 2002 for whom they have grade scores in the TAAS. This is a standardized test constituting the Texas high school exit exam, which is taken in 10th grade. It is offered several times, so that students failing the test or being absent when it is taken, could re-take the exam at later dates.[3] The main limitation of the available data is that before the TTP took place, data on class rank were not systematically collected. However,

[3]The authors use the student's first recorded score.

beginning with the graduating class of 1999, they observe for each student who applied to any Texas public higher education institution whether he/she was in the top ten percent of his/her high school class. The authors exploit this information to impute each student's probability of being in the top ten percent, before and after TTP, regardless of whether he or she applied to college (for a detailed description of the imputation method see Section 1.4).

The available data also contains information on college enrollment in the year after high school graduation, college completion, and labor market outcomes. For the first measure, the authors distinguish between community colleges, four-year campuses, and the two most selective campuses, i.e., Texas A&M and UT Austin. College completion is measured as whether an individual graduated with a bachelor's degree from a Texas public institution within six years of high school graduation, and attainment of a bachelor's degree in a scientific (STEM) field. Finally, the data includes information on earnings nine to eleven calendar years after high school graduation, at ages roughly 27 to 29, as well as thirteen to fifteen years after graduation. They also have access to a variable indicating whether each individual ever appears on the data, which accounts for long-term unemployment and absence from the state. Andrews et al. (2020) find that, in Texas, wages of students who leave the state are not systematically different from students who do not. Thus, not observing income for students leaving Texas does not entail a concern for the validity of the results.

1.4 Empirical Approach

To estimate τ_t^{PI} and τ_t^{PO} simultaneously, we can run the following linear regression

$$\begin{aligned} Y_{i,t} = \beta_0 + \beta_1 \text{PulledIn}_{i,t} + \beta_2 \text{PushedOut}_{i,t} + \beta_3 \text{PulledIn}_{i,t} \times \text{Post}_{i,t} + \beta_4 \text{PushedOut}_{i,t} \times \text{Post}_{i,t} \\ + Z_{i,t}\theta + \delta_t + \epsilon_{i,t}, \quad \mathbb{E}\left[\epsilon_{i,t} | Z_{i,t}, \text{PulledIn}_{i,t}, \text{PushedOut}_{i,t}\right] = 0, \end{aligned} \tag{11}$$

where $Z_{i,t}$ is a vector of individual characteristics, $\text{Post}_{i,t}$ is an indicator for the cohorts affected by TPP, from 1998 onward, δ_t are year-fixed effects, and $\epsilon_{i,t}$ is an idiosyncratic shock. The rationale to control for Z_{it} in the regression is that some students could be pulled it because of factors such as race, gender, or ethnicity. Thus, if we do not include them, $\text{PushedOut}_{i,t}$ could be correlated with the unobserved component (idiosyncratic shock), entailing misleading (biased) estimates for the causal effects. Furthermore, the inclusion of covariates helps to estimate the causal effects more precisely (reducing the estimates' variance). On the other hand, the year-fixed effects are included to account for potential events that could affect the outcomes of the treatment and control groups, avoiding over- or under-estimating the causal effects of the TTP. Finally, the variable $\text{Post}_{i,t}$ enables us to identify the causal effects for a given t. We now turn to illustrate how the parameters of interest relate to Figure 2, to motivate their inclusion.

Consider the hypothetical case in which there were no pushed out students by the TTP, and the students' characteristics do not affect their treatment status. Furthermore, we are only interested in the causal effect for the year 2002. This way, equation (11) boils down to

$$Y_{i,t} = \beta_0 + \delta_{2002} + \beta_1 \text{PulledIn}_{i,t} + \beta_3 \text{PulledIn}_{i,t} \times \text{Post}_{i,t} + \epsilon_{i,t}, \quad \mathbb{E}\left[\epsilon_{i,t} | \text{PulledIn}_{i,t}\right] = 0. \tag{12}$$

To map the parameters in equation (12) to the elements in Figure 2, consider evaluating equation (12) in the following cases:

$$\begin{aligned} A &= \mathbb{E}\left[Y_{i,2002}^{PI}(1) | \text{PulledIn}_{i,t} = 1\right] = \beta_0 + \delta_{2002} + \beta_1 + \beta_3 \\ C &= \mathbb{E}\left[Y_{i,2002} | \text{PulledIn}_{i,t} = 0\right] = \beta_0 + \delta_{2002} \\ D &= \mathbb{E}\left[Y_{i,1996} | \text{PulledIn}_{i,t} = 1\right] = \beta_0 + \beta_1 \\ E &= \mathbb{E}\left[Y_{i,1996} | \text{PulledIn}_{i,t} = 0\right] = \beta_0. \end{aligned}$$

The constant β_0 captures common characteristics of the individuals in the sample and corresponds to point E in Figure 2. By subtracting E from C, we get δ_t, which estimates the outcome trend for the control group, corresponding to the solid green line in Figure 2. Conversely, β_1 estimates the difference in outcomes of the baseline period (1996), which is D-E in Figure 2). Finally, β_3 estimates the causal effect of the policy, which is A-B, or equivalently A-(C+(D-E) under the parallel trends assumption. Thus, the parameters β_3 and β_4 in equation (11) estimate τ_t^{PI} and τ_t^{PO}, respectively. Motivated by equations (9) and Figure (9), β_3 and β_4 are called difference-in-differences estimators, because they

estimate the causal effects as the difference in differences of the outcomes for the treatment and control groups.

The main challenge in estimating equation (11) is that the authors do not identify which students were pulled in and pushed out by the policy, so that $\text{PulledIn}_{i,t}$ and $\text{PushedOut}_{i,t}$ are unobserved. Accordingly, the authors adopt a three-step procedure to estimate the effects of the TTP. First, they estimate the likelihood that each student is in the top 10 percent of his or her high school class. Subsequently, they define the tree comparison groups (pulled in, pushed out, and control group) based on this likelihood. Finally, they estimate equation (11) using the treatment status imputed in the previous step.

The limitation of this three step procedure is that the interpretation of the difference-in-difference estimators is not causal. The estimated Pulled In and Pushed Out indicators are imperfect proxies for the groups affected by TTP. In particular, they both include some students who were eligible for TTP guaranteed admission and some who were not. Furthermore, the estimated treatment groups also include students who would have been admitted to UT under both the pre-TTP and post-TTP admissions rules. Consequently, the estimates of β_3 and β_4 in their setting can be interpreted as the intention-to-treat (ITT) estimates of the effect of changes in access to selective colleges.

We now describe the three-step procedure and illustrate the concept of intention-to-treat and its relationship with the average treatment on the treated in equations (3) and (4).

First Step

The purpose of this step is to estimate the probability of each student being in the top ten percent of their high school. Let W be an indicator for a student from the 1999-2002 cohorts who applied to at least one public college in Texas, and $T = 1(r > 0.9)$ an indicator for being in the top ten percent of the high school class. The main data limitation is that only T is observed for the $W = 1$ subsample. Finally, let X be a vector of student characteristics, that includes characteristics that are measured consistently throughout our sample period, which include (i) TAAS exit exam scores in reading, writing, and math, measured both in statewide percentiles and as the percentile within the school; (ii) indicators for math and science course-taking (e.g., advanced math in 11th grade); (iii) the number of foreign language courses taken in high school;(iv) the number of courses failed in high school; (v) the number and percentage of school days absent in 12th grade; (vi) an indicator for being 18 upon graduation; (vii) the school's racial, gender, and socioeconomic (free & reduced-price lunch) composition; and (viii) the share of students at the school who are classified as special education.

We want to estimate the probability[4] of being in the top 10 percent of an individual's high school given student characteristics

$$p\left(T = 1|X\right) = \mathbb{E}\left[T|X\right]. \tag{13}$$

As we mentioned earlier, reliable assumptions are what allow us to learn something from the data. To impute (estimate) the conditional probability of being in the top 10 percent of an individual's high school, the authors assume that conditional on X, T is independent of W. That is, once we have taken into account student characteristics, applying to at least one public college in Texas is independent of whether the student was in the top ten percent of her high school. In doing so, we have

$$\mathbb{E}\left[T|X\right] = \mathbb{E}\left[T|X, W = 1\right].$$

This equation tells us that we can estimate the probability in equation (13) for the whole sample, using only the subsample of students from the 1999-2002 cohorts who applied to at least one public college in Texas.

Differences between applicants and non-applicants in the relationship between covariates and top 10 status would violate the assumption that conditional on X, T is independent of W. However, the authors argue that the evidence in Long and Tienda (2008) supports the validity of this assumption.

To estimate $\mathbb{E}\left[T|X, W = 1\right]$ the authors perform Random Forest Classification. The main appeal of this approach is allowing for non-linearities and interactions between student characteristics while reducing over-fitting.[5]

[4]Since T is a binomial variable, the probability of its occurrence equals its expected value.

[5]In this framework, overfitting occurs when the top 10 status is almost perfectly classified for $W = 1$, so that the Random Forest would not make accurate predictions for the $W = 0$ (see Hawkins (2004) for an overview on over-fitting).

After estimating the conditional probability of being in the top ten percent, the authors find that 23 percent of students have zero predicted probability of being in the top 10 percent of their high school. Accordingly, they construct a variable $\hat{q}$ that takes 40 different values. In particular $\hat{q}_{it} = 1$ if $\hat{p}_{it}\left(T_{it} = 1|X_{it}\right) \leq 0.23$, $\hat{q}_{it} = 12$ if $0.23 < \hat{p}_{it}\left(T_{it} = 1|X_{it}\right) \leq 0.24$, and the remaining 38 value are assigned according to sets of two percent of the sample, e.g., $\hat{q}_{it} = 13$ if $0.24 < \hat{p}_{it}\left(T_{it} = 1|X_{it}\right) \leq 0.26$,..., $\hat{q}_{it} = 50$ if $0.98 < \hat{p}_{it}\left(T_{it} = 1|X_{it}\right) \leq 1$.

Second Step

From the right panel in Figure 1 we have that the treatment and control groups were defined according to the school rank and the school sending rate. The former object was estimated in the First Step, while the latter is measured as the share of students from each high school in the 1996 and 1997 (pre-TTP) cohorts who enrolled at UT Austin. Subsequently, high schools are categorized into deciles (s) according to their sending rate. Figure 3 depicts the share of students from schools in each decile attending UT Austin, pre- and post-TTP. While pre-TTP the bottom five deciles each send less than two percent of students, the top decile of schools sends thirteen percent of students to Austin. Even though post-TTP this pattern remains, the schools that previously sent few students to Austin send slightly more, while the share of students from the top-decile schools who attend Austin falls to eleven percent.

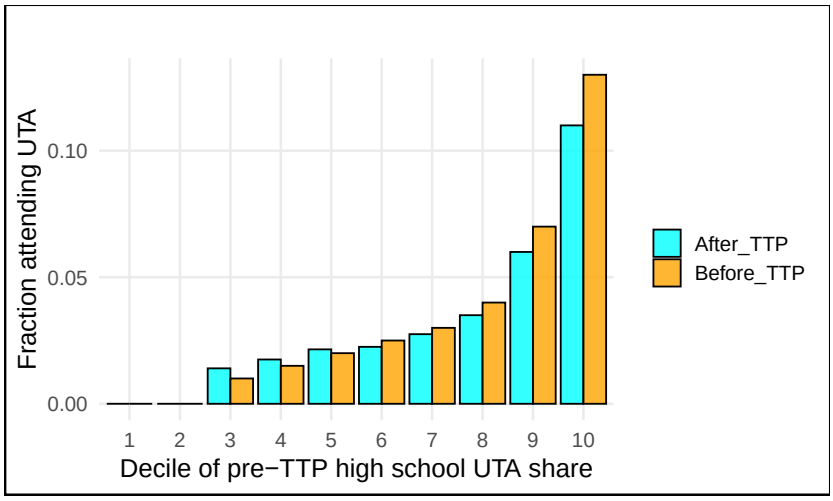


Figure 3: Simulated histogram constructed to replicate the share of students from high school attending UT Austin, by school pre-TTP decile (s) in the dataset of Black et al. (2023)

The authors use the estimated top-ten-percent probability categories from the First Step, and school sending rate deciles (s) to define the Pulled In, Pushed Out, and control groups. These three groups are defined according to the change in the sending rate of cells composed of $(\hat{q}, s)$ pairs. In particular, any cell where the share attending UT Austin rose (respectively, fell) by more than 3 percentage points is included in the Pulled In (Pushed Out) group.[6] Thus, consistent with Figure 1, the Pulled In students are those from schools with low s and high $\hat{q}$ (top-ranked students), whose school sending rate increased post-TTP. Conversely, the Pulled Out groups consist of students whose school has high s and moderately high $\hat{q}$, that is, students with a moderate likelihood of being in their school's top decile. As regards the control group, the authors use two blocks of cells that are close the the treated groups but present small changes in the probability of enrolment at UT Austin. These two groups of cells are selected for $\hat{q} \in [25, 40]$, and $s \in [6, 8]$, and $\hat{q} \in [40, 45]$, and $s \in [3, 5]$.[7]

[6] cells $\hat{q} = 50, s \geq 9$ are excluded from the Pulled In group, since they correspond to students in the top ten percent, attending "feeder" schools.

[7] For a detailed selection of these three groups, see Figure 2 in Black et al. (2023)

Third Step

The third step of the estimation approach consists of estimating equation (11) with the treatment status that was imputed in the Second Step. The treatment is defined in a multi-valued manner, which reflects each of the different colleges that students might attend, and that the TTP might affect it in complex ways. In some cases, the counterfactual of not being admitted to UT Austin is attending A&M, while for others it is a community college, a less selective UT campus, or no college at all. Furthermore, TTP may have affected students who do not attend UT Austin under either admissions policy, in their choices among other alternatives. For these reasons, along with the fact that the treatment status is not observed but imputed, the interpretation of β_3 and β_4 is not causal. Conversely, it can be interpreted as the intention-to-treat estimates of the effect of changes in access to selective colleges.

The intention-to-treat analysis considers every subject in the treatment and control groups ignoring noncompliance, protocol deviations, withdrawal, and anything that happens after the treatment status has been assigned (Gupta, 2011). Accordingly, the setting to evaluate the TTP falls into this definition. Firstly, the imputed Pulled In and Pushed Out indicators are imperfect proxies for the groups affected by TTP. They both included some students who were eligible for TTP guaranteed admission and some who were not. Moreover, the estimated treatment groups also include students who would have been admitted to UT under both the pre-TTP and post-TTP admissions rules. Given their ITT interpretation, the estimated effect of the TTP will likely substantially understate the impact on students who were actually affected by TTP.

1.5 Main Results

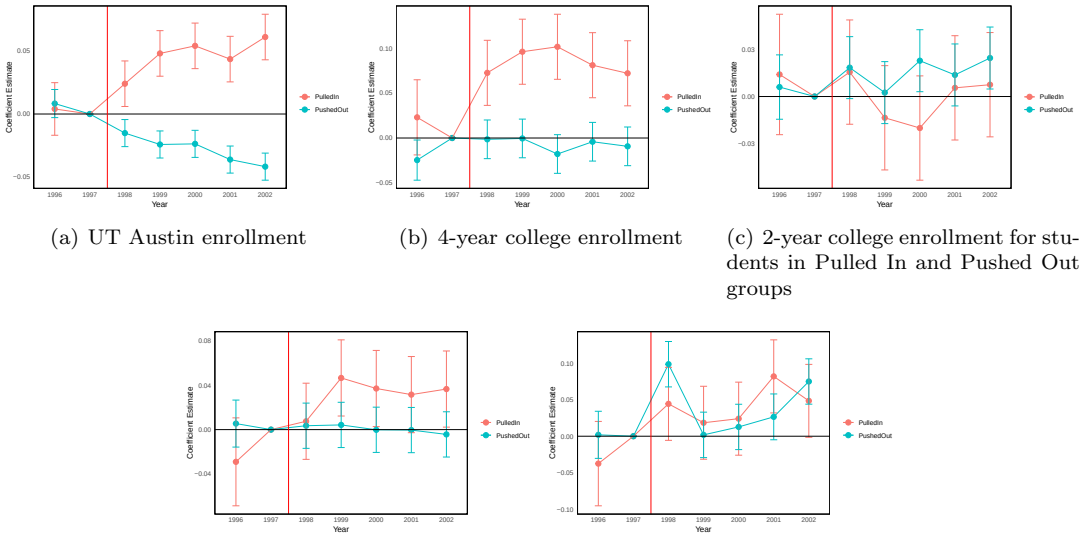
Figure 4 reports the results of estimating equation (11) in an event study design. To study the dynamic effects of the TTP, the Post_{it} indicator is replaced with a set of Year_t indicators, where 1997 is the excluded category. This way, we can estimate β_3 and β_4 for different years. In Section 1.2 we discussed that the identifying assumption is the parallel trends. Notice that in every result in Figure 4 the confidence intervals of estimates for 1996 cover the zero intercept. This lends evidence to the identifying assumption holding, as it suggests that the treated and control groups would have moved together absent the change in the admissions policy. Since in 1996, the TTP had not taken place yet, finding significant effects for the pulled in or pushed out would suggest that the parallel trend assumption does not hold. However, we find evidence it does.

The results suggest a positive effect on the pulled in students in educational and labor market outcomes, while the pulled out were only affected regarding enrollment at UT Austin. According to Figure 4 pulled in students were more likely to enroll both at UT Austin and four-year colleges (see panels (a) and (b)). Furthermore, they were also more likely to have higher earnings 9-11 years after high school completion according to panel (e) in Figure 4. In contrast, the authors find that Pushed Out students were less likely to enroll at UT Austin, meanwhile, the TTP did not have a significant effect on the remaining outcomes.

Table 1 reports the main results of estimating equation (11) for educational and labor market outcomes. According to column (1), we find that Pulled In students are 5.3 percentage points more likely to attend UT Austin due to the TTP, while there is no significant change in enrollment at Texas A&M. Moreover, these students are 6.6 percentage points more likely to attend a public four-year college in Texas. However, they are no more likely to enroll in community college. Finally, Pulled In students are 5.2 percentage points more likely to attend any college. It could have been that some students switched from no college to community college, and some others switched from community college to UT Austin. In any case, the results suggest that Pulled In students in the sample were more likely to enroll in college and more likely to attend UT.

Column (2) in Table 1 reports the results for Pushed Out students. After the TTP, these students were 3.6 percentage points less likely to enroll at UT Austin. However, such a decrease was compensated by a 2.2 percentage point increase in enrollment at other four-year schools, and a 1.1 percentage point increase in enrollment at community colleges. Meanwhile, both the net effect on total four-year enrollment and college enrollment overall is statistically insignificant.

Pulled In students were 3.9 percentage points more likely to graduate with a BA from UT Austin within 6 years after high school graduation relative to control students. Regarding any four-year college in Texas, this number was 3.7 percentage points. Conversely, the Pushed Out students were



(a) UT Austin enrollment

(b) 4-year college enrollment

(c) 2-year college enrollment for students in Pulled In and Pushed Out groups

(d) college completion (BA attainment)

(e) earnings 9-11 years after high school completion

Figure 4: Simulated event study constructed to replicate key results from Black et al. (2023)
Note: Figures show point estimates and 95% confidence intervals for simulated data corresponding to β_3 and β_4 interactions in an event study version of Equation (11) that replaces the Post_{it} indicator with a set of Year_t indicators, where 1997 is the excluded category.

2.1 percentage points less likely to graduate with a BA from UT Austin within 6 years after high school graduation relative to control students. However, the effect on bachelors from any institution was non-significant.

In summary, the authors find that Pulled In students were more likely to attend UT Austin and different four-year universities, were also more likely to graduate from college, and had either equal or higher earnings after the TTP. Furthermore, they find that the Pushed Out students were not substantially harmed. Even though these students were less likely to enroll at UT Austin, this was compensated by enrollment in other four-year schools and community colleges.

Table 1: Baseline Difference-in-Difference Analysis

	Pulled In (1)	Pushed Out (2)
Enrollment outcomes		
UT Austin	0.053	-0.036
	(0.005)	(0.004)
Texas A&M	-0.010	0.008
	(0.007)	(0.004)
Any college	0.052	0.000
	(0.011)	(0.007)
Any 4-year	0.066	-0.005
	(0.011)	(0.008)
Any community college	-0.004	0.011
	(0.008)	(0.005)
Any other 4-year	0.024	0.022
	(0.009)	(0.006)
Degree attainment within 6 years		
Bachelors from UT Austin	0.039	-0.021
	(0.004)	(0.003)
Bachelors from any institution	0.037	-0.001
	(0.010)	(0.006)
Associates or better	0.032	-0.006
	(0.010)	(0.007)
Bachelors with STEM major	-0.007	-0.001
	(0.006)	(0.003)

Note: Results taken from Table 3 in Black et al. (2023).

2 The effect of school-starting age rule on educational outcomes

Evaluating the impact of educational policies, such as school-starting-age (SSA) is not an easy task, especially when we are interested in establishing causality rather than mere correlation. Determining whether starting school at an older age leads to improved academic performance requires a robust method that separates this factor from other potential influences on educational outcomes. In this section, we utilize a regression discontinuity design to estimate the causal effect of school-starting-age rules on educational outcomes for a selection of six OECD countries. For this purpose, we use the PISA dataset for the period 2015-2022. A detailed step-by-step guide to reproduce the results is available in the replication file, available at this Repository.

Regression discontinuity design (RDD) is a quasi-experimental approach that allows researchers to identify causal effects by exploiting cutoffs or thresholds (see Chapter 2). These cutoffs may occur naturally or be enforced by policies, such as age-based enrollment rules. This section offers comprehensive guidelines on applying RDD in the context of educational policy analysis.

The identifying assumption in our framework is that a substantial discontinuity jump in the average enrollment age for a given month and country is exogenous. Consequently, we suggest implementing a fuzzy regression discontinuity design to address the threat of endogeneity caused by omitted variable bias. Such a threat arises here due to unobservable factors, i.e., determinants of school entry decisions that may affect the outcomes among children who start school at different ages, such as red-shirting and strategic birth timing.

The results suggest that individuals who are older when they start school get on average better test scores in both reading and math disciplines. We find either this result or no effect in the two countries considered, with remarkable heterogeneity.

2.1 Motivation: RDD in the Context of Educational Policies

Educational policies often involve eligibility criteria based on arbitrary thresholds, which makes them well-suited for RDD analysis. However, one must be very careful with the caveats and the limitations that the specific context may have. For this, we propose the following framework as an example of application: school starting age policy.

In many countries, there are specific cutoff dates for children to start school. In our setup, August 31st will be the choice. For example, if a child's birthday is just before the cutoff date, they have to wait an extra year to start school. But if their birthday is just after the cutoff, they can start right away. This situation creates a natural experiment where students born just before and after the cutoff are likely similar in observable and unobservable characteristics, except for their school-starting age.

In the proposed application, we evaluate the causal effect of school-starting age (SSA) rules on educational attainment. For this purpose, we use PISA data from 2015 to 2022 for a selection of one OECD country, i.e., Austria, and one non-OECD country, i.e., United Arab Emirates (U.A.E. from now onwards). Based on documented starting age criteria PIRLS (2022) and empirical evidence, we choose the countries with a substantial jump in the average enrollment age for a given month.

In this context, we can estimate the causal effect of starting school at an older age by comparing education outcomes-in our case, standardized test scores- of students born just before and just after the cutoff date (e.g., August vs. September). However, this can be done under certain assumptions that will be discussed later. Therefore, the cutoff date serves as an exogenous factor, unaffected by individual characteristics or choices, since birthdate is random enough. This allows for isolating the effect of school-starting age on outcomes, providing a rigorous identification strategy for policy analysis.

2.2 Data

For our analysis, we use data from the 2015, 2018, and 2022 PISA waves. They contain individual-level information on educational variables, along with individual characteristics. Our main dependent variables are educational outcome variables regarding standardized test scores in mathematics and reading. Furthermore, we also consider being a repeater (variable that takes value 1 if the individual i has repeated a grade at least once, and 0 otherwise) as an outcome variable to evaluate the effect of starting school age in the likelihood of being a repeater. We exploit the role of the socio-demographic

characteristics to mitigate the threat of endogeneity. We have information on age, gender, family structure, migration, whether the student has repeated or not, educational and labor expectations, parents' occupations, home possessions, and index of socioeconomic and cultural status.

The rationale for using the PISA dataset arises from the fact that it intends to evaluate the acquisition of important knowledge and skills required for adult life, rather than the mastery of the school curriculum (Schleicher and Tamassia, 2000). Therefore, their measuring approach for educational attainment is informative about the ability to use the school knowledge in real-life challenges (Fuchs and Wößmann, 2008). As regards the dependent variables, we rely on plausible values for the scores in math and reading to measure performance. The so-called plausible values are generated through multiple imputations based on pupils' answers to the subset of test questions they were randomly assigned and their responses to the background questionnaires. The analysis of such plausible values is a bit more complicated, but for illustration purposes, we take the average score per student on each domain (mathematics and reading) as an outcome variable for the analysis based on the sampling weights available in the data. Notice that the PISA dataset is based on a sample, and not on the whole population of 15-year-olds, therefore we must use the provided sampling weights. Moreover, in the context of PISA, there is no theoretical minimum or maximum score. Instead, the outcomes are standardized to align with approximately normal distributions, with mean scores hovering around 500 points and standard deviations of approximately 100 points.

Our objective is to evaluate the causal effect of SSA rules on educational attainment. Our identification strategy relies on the presence of a discontinuity jump in the average enrollment age for a given month. Consequently, our analysis must be based solely on countries with a clear school-starting rule which depicts a discontinuity jump in the data. Following Oosterbeek et al. (2021),[8] we report in Figure 5 the discontinuity at the cutoff month. It is worth noting that if children were born after this month, they must wait for the next year to start primary school. The cutoff month is August-September. We select among these countries those who have the (greatest) discontinuity jump in the average SSA around the cutoff date. Furthermore, we exploit the fact that countries vary substantially in their educational systems to evaluate how our estimation procedure performs. These countries are Austria, and U.A.E.

In Figure 5, we report both the average for relevant ages and for all ages. The relevant ages are those that might impede a student from starting school. That is, if the cut-off point is six years[9], then the only relevant children for identification purposes are aged either six or seven. Accordingly, calculating the weighted average age including all ages, might introduce noise in the average age. Fortunately, we do not see systematical changes in the selected countries. We document the math

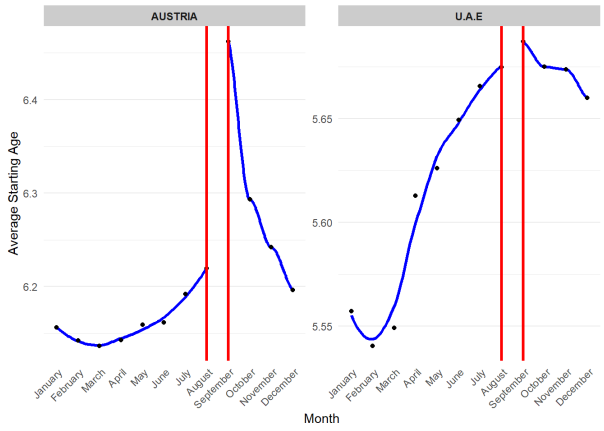


Figure 5: Discontinuity At The Cutoff

and reading scores variation between countries in Figure 6 and within countries in Figure 7 over time. Figure 6 suggests that between-country differences in test scores are not only persistent but are getting

[8]Our analysis is slightly different because we miss the days of birth, which is an important constraint in the data to bear in mind. The constraint is that the running variable will be no longer continuous but discrete. Further analysis will be conducted in the next steps.

[9]For Austria, and U.A.E., it is compulsory to start school from the first day of September following their sixth birthday.

more acute over time. Moreover, such differences are systematic across the student distribution. That is, they are maintained across deciles. This means, that (i) the best students in the over-performing country (e.g., Austria) have better results than the best students in the under-performing countries (e.g., U.A.E.), (ii) the students with the lowest scores in the under-performing countries have worse results than low-achieving ones in over-performing countries. In summary, the results of a random student from an under-performing country such as U.A.E. would be on average worse than a random student from an over-performing country such as Austria if they belong to the same decile, regardless of which. On the other hand, Figure 7, suggests that within-country performance is persistent across countries and worsens over time.

Figure 6: Trend of Test Scores Over Time by Countries

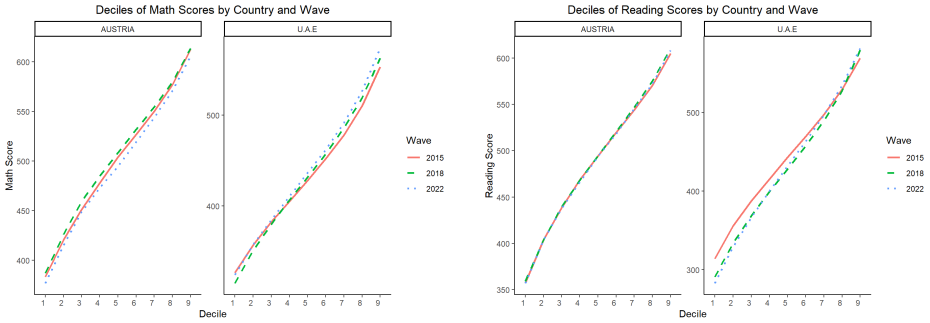


Figure 7: Trend of Test Scores Over Time Within Countries

To assess variation in outcomes within countries, Figure 8 shows the variation in outcomes (test scores) across different regions within countries. As one can see, there exists substantial variation in outcomes displayed by years, with the number of outliers either increasing or decreasing over time. This variation is important for several reasons, first, it highlights regional disparities within countries that might not be apparent when looking at national averages. Understanding these differences can help target interventions where students are most disadvantaged. Second, it provides valuable insight into how different regions or localities respond to the same national education policies, helping policymakers assess the effectiveness of reforms in specific contexts. Third, this variation underscores the need for tailored solutions rather than a one-size-fits-all approach, as educational needs may differ significantly within a single country. However, in this exercise, we do not propose a policy evaluation, but a mere illustration of how starting school at a later age affects academic performance using a methodology that allows us to obtain causal estimates.

To identify the factors that contribute to these variations, Figure 9 shows the differences between children with highly educated parents and those with low-educated parents. As expected, children with highly educated parents tend to perform better on average than children with low-educated parents. This finding, consistent with existing literature (Dubow et al., 2009), suggests that parental education

plays a significant role in shaping educational outcomes. By incorporating this dimension into our analysis, we can better understand how family background contributes to the observed within-country variation.

Analyzing variation in outcomes within countries helps capture the influence of other demographic and socioeconomic factors. Alongside parental education, issues such as regional poverty levels, access to educational resources, and school quality all contribute to differences in outcomes. Regions with lower scores may face these cumulative challenges, which are often hidden when considering national averages alone.

In sum, understanding this variation allows for more precise, data-driven decisions. Regarding

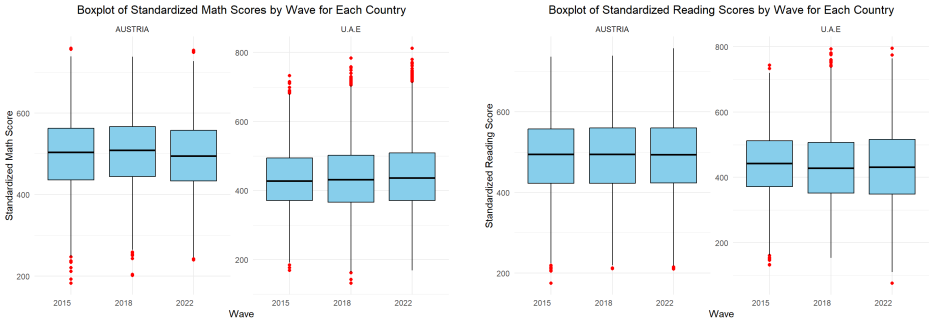


Figure 8: Variation Within Countries

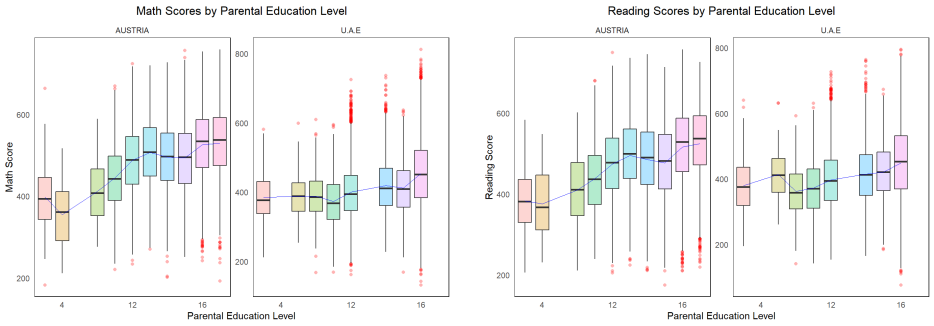


Figure 9: Intergenerational Mobility

the relationship between SSA and test scores, our findings reveal a negative correlation between SSA and test scores, as presented in Figure 10. On the left panel of Figure 10, we find raw test scores, and on the right panel, we find standardized test scores, which will be our outcome variable. Moreover, to enhance the comprehensiveness of our analysis, we considered the characteristics of students to assess the reliability of our correlation. This correlation suggests that students who start their academic journey at a later age tend to exhibit inferior performance in both math and reading. Our findings are in alignment with the existing literature, which emphasizes the significance of personal attributes (Van Bragt et al., 2011; Phelps, 2006; Matthews et al., 2003). However, it is crucial to note that this correlation does not necessarily imply causation.

Moreover, we also provide further insight into the average starting age of schooling across countries and time. Specifically, Table 2 displays the average schooling starting age for each country in each wave. The results confirm our earlier findings and substantiate the average SSA of the country group as 6, making reasonable the choice of the cutoff playing a key role in the choice of the cutoff. Thus, we are comparing children who started school at the age of 6.

2.3 Understanding the Regression Discontinuity Design Framework

As previously introduced, the Regression Discontinuity Design is a quasi-experimental approach that uses cutoff-based rules to identify causal effects. In this context, we have to introduce a key concept: the *"running variable"*. The running variable, denoted by Z, determines who gets the treatment (starting

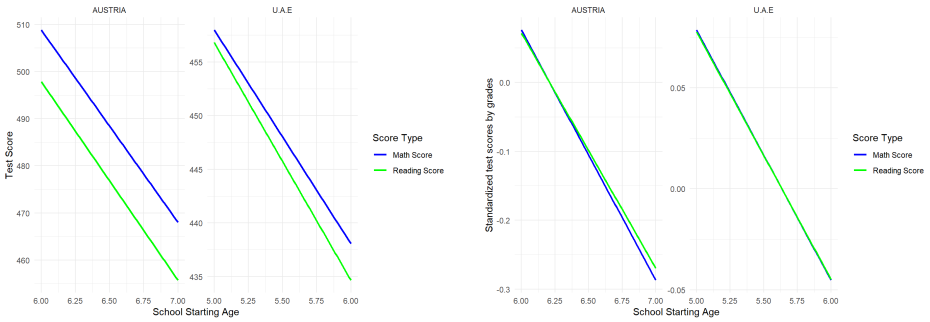


Figure 10: Correlation Between School Starting Age and Test Scores

Country	2015	2018	2022
Austria	6.20	6.23	6.19
U.A.E.	5.59	5.65	5.65

Table 2: Average School Starting Age by Countries and Waves

school) or not based on a cutoff. Usually, this variable is continuous. However, we will present a case in which the running variable is discrete. The idea behind this is to compare individuals just below and just above the cutoff, assuming that these observations are comparable except for their treatment status. In our exercise, individuals who were born in August and in September share the features to be compared, but their treatment statuses are completely different.

Broadly speaking, we can differentiate two types of RDD designs: sharp RDD and fuzzy RDD. In plain words, in a sharp RDD context, the treatment is perfectly determined by whether the running variable exceeds the cutoff. In our context, imagine that a student, Pepe, is born after August 31st. In his case, Pepe must start school the following year. Here, the assignment to treatment (e.g., delayed school entry) is a deterministic function of the running variable (birthdate).

On the other hand, we have the fuzzy RDD. As self-explanatory as it might seem, in this case, the probability of receiving the treatment is no longer perfectly determined by the cutoff (i.e., either 1 or 0), but it changes discontinuously at the cutoff. In our situation, imagine that Pepe's parents decide that Pepe is not starting school this year but the following (e.g., some parents delay their child's school start even if they qualify under the cutoff). In this case, there is partial compliance or exceptions, resulting in not all children who qualify starting school at the designated time.

In our example, our selection rule allows us to exploit such discontinuity jump to identify the causal effect of SSA rules on test scores. We employ a quasi-experimental approach implementing a fuzzy regression discontinuity design (Lee and Lemieux, 2010) to address the threat of endogeneity caused by omitted variable bias. Such a threat arises here due to unobservable factors, i.e., determinants of school entry decisions that may affect the outcomes among children who start school at different ages, such as red-shirting (Dhuey et al., 2019) and strategic birth timing (Buckles and Hungerman, 2013).

In particular, we exploit the exogeneity of the month of birth to elicit causal effects. Specifically, we aim to explore the differences in performance between individuals born in September, who are the eldest in their cohort, and those born in August, who are the youngest. We assume that there exists a linear association between an individual's outcome variable, designated by y_{it}, and their birth month, D_i, within a given cohort t:

$$y_{it} = \alpha_t + \tau D_i + \epsilon_{it}, \mathbb{E}\left[\epsilon_{it} D_i\right] = 0, \tag{14}$$

where ϵ_{it} is the unobserved component determining the test results orthogonal to the birth month, which gives rise to our first assumption.

Assumption 1.

$$\mathbb{E}\left[\epsilon_{it} D_i\right] = 0. \tag{15}$$

Our parameter of interest is captured by τ and α_t denotes birth cohort fixed effects, with cohorts centered around August 31st. In particular, we consider a fuzzy regression discontinuity (RD) strategy

that relies on birth dates[10].

Before proceeding further, we just mentioned that RDD can provide a causal estimate under certain conditions. Let's see what are these and how they apply to our case:

- **Continuity Assumption:** Individuals just below and just above the cutoff are assumed to be similar in both observed and unobserved characteristics. This implies that any discontinuity in the outcome variable at the cutoff can be attributed to the causal effect of the treatment rather than pre-existing differences between the groups. In our case, we compare (within countries) students just below the threshold (being the youngest and therefore born in August) with students just above the threshold (being the oldest and therefore born in September). It seems reasonable to think that any jump in the outcome around that threshold is due to being born in either August or September.

  Assumption 2 ensures that there is a clear discontinuity in the probability of receiving treatment at the threshold.

  Assumption 2.

$$\lim_{z \to z_0^+} Pr(D = 1|Z = z) \neq \lim_{z \to z_0^-} Pr(D = 1|Z = z) \tag{16}$$

$$Pr(D_i = 1|Z_i) = \begin{cases} g_1(Z) & \text{if } Z_i \geq z_0 \\ g_0(Z) & \text{if } Z_i < z_0 \end{cases}$$

  The key assumption is that individuals right above and below z_0 are comparable since only the random variation is the reason why someone is put above z_0 or below, generating differences in treatment. As aforementioned, any difference in the outcome Y right at z_0 is due to treatment.

  Assumption 3 formalizes this requirement by stating that the distribution of potential outcomes should be continuous at the threshold.

  Assumption 3.

$$\lim_{z \to z_0^+} Pr(Y_j \leq r|Z = z) = \lim_{z \to z_0^-} Pr(Y_j \leq r|Z = z) \tag{17}$$

  The RD design entails discontinuity in treatment assignment but continuity in potential outcomes.

- **No Manipulation of the Running Variable:** The running variable (e.g., birthdate) should not be manipulated in anticipation of the treatment. For example, in the context of school enrollment rules, parents should not time births to ensure their child qualifies or does not qualify under specific school-starting rules. If manipulation occurs, it can lead to biased estimates. In this case, the discontinuity arises due to arbitrary selection criteria used for school admissions. Hence the identification of average treatment only requires that there are no manipulation of running variables near the cutoff.

  In this framework, the discontinuity arises due to arbitrary selection criteria used for school admissions. Hence the identification of causal effect only requires that there is no manipulation of the running variable near the cutoff. We confirm no manipulation of the running variable using Figure 11. Nevertheless, one may argue that the histogram is not uniform enough, implying a jump around the cutoff (month 0) for some cases. Since this is the number of students born either in August (month -1) or in September (month 0), this would imply that parents have some sort of strategical timing within a year when they decide to get pregnant, and this is unlikely since the process of getting pregnant is somewhat random. Accordingly, we assume that there is no manipulation of the running variable. However, one should notice that the inclusion of additional variables can cause downstream effects and cause issues in the estimation of the average treatment effect.

- **Local Randomization:** Near the cutoff, the treatment assignment is effectively random. Individuals just below and above the threshold should not differ systematically in other characteristics that might affect the outcome.

[10]The probability of being treated depends on crossing the threshold, albeit it is not the only determinant.

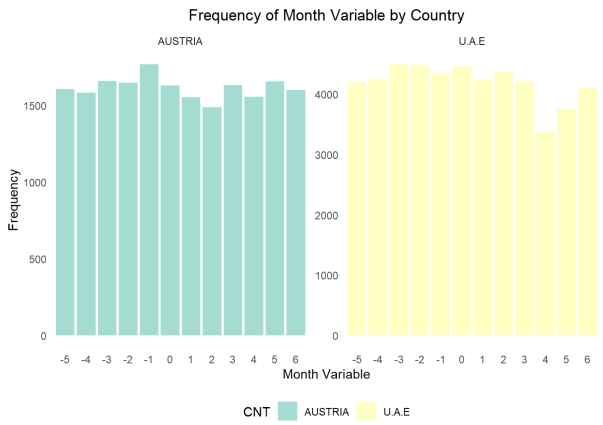


Figure 11: Distribution of Running Variable Near the Cutoff for Selected Countries

The main identification strategy[11] is that individuals in the small neighborhood of the cutoff value will be similar to a randomized experiment at the cutoff point because they essentially have the same value of Z. Thus, a comparison of the average value of Y of those just above and below the cutoff z_0 will produce an estimate of the causal effect. An important limitation is that increasing the interval around the cutoff will bias the estimate of the treatment effect. In this application, individuals in the small neighborhood of the cutoff value (that is, being born close to 31st August) will be similar. Thus, a comparison of the average value of test scores of those just above and below (being born either before or after 31st August and therefore allocated in one course or another) will produce an estimate of the causal effect.

The rule used for students to present the PISA is the same for all countries: PISA sample represents all students between the age of 15 years and 3 months and 16 years and 2 months, and who are enrolled in an educational institution at grade 7 or higher, minus those that are excluded (OECD, 2023). Thus, it does not matter in which academic course students are allocated but their age when taking the exam. However, there are three different effects regarding age and educational performance: *age-at-the-test*, *school-starting-age*, and the *years of schooling at the time of the test.* According to Avvisati and Givord (2021), it is not possible to disentangle these effects separately in a longitudinal study as PISA. Accordingly, we are identifying additively these three.

Remember that our goal is to isolate the impact of the age at which schooling starts. To achieve this, we need to control for the effects of both the age when taking the test and the years of schooling completed at the time of the test. In this regard, regression discontinuity is employed to address the effect of age when taking the test because we are comparing students whose ages on the test are less than one month apart. However, precisely because those in September start schooling when they are one year older than those born in August, they have one less year of schooling when taking the PISA test. Thus, in the RDD that we employ[12], we cannot separate the effect of school-starting age from the effect of one more year of schooling. On this matter, to control for the effect of years of schooling, we standardize the PISA score by the grade in which the student is enrolled at the time of the test. Notice that standardization requires some extra assumptions to take care of the extra year of schooling. That is, we standardize the PISA score by the grade in which the student is enrolled at the time of the test assuming that this standardization approach reflects real-world comparisons, as student outcomes are generally assessed within their cohort. However, if one does not support this assumption, we might inadvertently reintroduce an age effect. On the other hand, the RDD does not need additional assumptions to make a same age-at-test comparison

Another potential threat that prevents us from obtaining the causal effect and therefore affects the identification arises from heterogeneity in the distribution of other characteristics of individuals at the point of discontinuity. One way to check whether this is the case is by visualization. Nevertheless,

[11]The identification strategy allows us to elicit causal effects. In plain words, researchers use it to figure out if a relationship between two things is real or just coincidental. In our context, we figure out if X (age of school-starting) causes Y (better educational outcomes) without being fooled by other factors (like talent).

[12]Other data and/or methodology will compare students with the same years of schooling but with different ages-at-the-test (in addition to different school-starting-ages).

the variables regarding individual characteristics are mostly discrete with many categories. Hence, including the visualization of these variables would be futile.

2.4 Estimation Procedure

2.4.1 Data Processing

Before proceeding to the estimation, it is crucial to conduct proper data processing to ensure reliability and accuracy across countries and waves. The steps involved are as follows:

1. **Selection of Starting Ages**: We limit the analysis to the two most common starting ages for each country, ensuring comparability across regions and time.
2. **Standardization by Country and Wave**:
   - All test scores (e.g., math and reading) are standardized within each country and survey wave by the grade in which the student is enrolled at the time of the test.
   - Importantly, no weights are applied during this standardization process. Although we tested the impact of applying weights, the results indicated that their use did not significantly alter the estimates. Hence, we proceed without them in this step for the sake of simplicity.
3. **Restriction to Boundary Months**: To focus on the discontinuity in the running variable (birthdate), we restrict the dataset to individuals born in the months of August and September. These months represent the cut-off points on either side of the discontinuity, ensuring that the analysis captures the immediate effect of the birthdate threshold.

2.4.2 Estimation Procedure

Once the data is preprocessed, the estimation proceeds as follows:

1. **Country and Wave-specific Estimation**: We conduct separate estimations for each country and each survey wave. This allows for flexibility in capturing country-specific and temporal heterogeneity in the impact of the birthdate cutoff on educational outcomes.
2. **Generation of the Running Variable (Z)**: The running variable, denoted as Z, is discrete in this context and represents the birthdate of the individual. Specifically, Z is a binary variable that indicates whether an individual i was born before August 31st ($Z = 1$) or after ($Z = 0$). This variable serves as the threshold for identifying individuals who are likely to have started school in different cohorts.
3. **Regression Model**: We estimate the effect of the running variable Z on the standardized educational outcomes (e.g., math and reading scores) using a simple linear regression model. The model is specified as follows:
$$Y_i = \alpha + \beta Z_i + \epsilon_i$$
where Y_i represents the standardized score of individual i, Z_i is the binary indicator for birthdate, and ϵ_i is the error term.
   - Robust standard errors are used to account for potential heteroskedasticity in the data.
   - We also include sample weights in this step to correct for any potential survey design or sampling biases that might affect the representativeness of the sample, ensuring that our estimates reflect the underlying population accurately.

2.5 Results

Table 3 suggests that being older when an individual starts school for the first time has a positive and significant effect on test scores for both disciplines (reading and math). Columns (1)-(3) and (5)-(7) report the results from different waves for reading and math, respectively. That is, Columns (1) and (5) depict the reading and math results for wave 2015, Columns (2) and (6) represent the reading and math results for wave 2018, and Columns (3) and (7) show the reading and math results for wave

2022. The results reported in Columns (4) and (8) are the aggregated ones. Moreover, Column (5) in Table 3 indicates that being one year older in Austria increases the math score of a student by 0.10 standard deviations on average.

This gives us a glimpse that the age at which children start school seems to impact their performance later on in several ways. However, our estimates capture the aggregate effect of age-at-test school-starting-age, and years of schooling at the time of the test rather than the effect of a policy. Firstly, older students at the beginning of their primary education will be older at the time of evaluation, giving them an advantage over younger students. Secondly, their readiness for the class, as indicated by factors such as whether they had to repeat a grade, can also influence their performance down the line. However, we find no evidence in the latter case.

The estimates in Table 3 are in line with previous results in the literature (Givord, 2020; Oosterbeek et al., 2021), which, combined with the use of a regression discontinuity design, allows us to give them a causal interpretation. The RDD design, by exploiting the cutoff date for school entry, offers a clear strategy to address endogeneity and selection bias, thereby reinforcing the validity of our findings. However, it is still important to acknowledge several limitations in this analysis.

While the point estimates appear stable across countries, the external validity of these findings warrants attention. The aggregated results may mask important country-specific differences although we have incorporated into our analysis a normalization by country and year. Austria and U.A.E have distinct educational systems, curricula, and cultural expectations. The fact that either positive or no effects at all were found for such countries suggests that factors such as policy changes or cohort effects could influence the results. Concerning this, we again accentuate that this exercise is not aiming to evaluate educational policies themselves but the aggregate of age-at-start and starting-school-age on education outcomes. Thus, while the pooled analysis gives a broad picture, a deeper look into these cross-country variations is necessary to understand how different contexts shape the impact of school starting age.

Additionally, the fact that the results for Austria are significant in only one wave, or even not at all, as is the case for U.A.E. raises questions about the robustness of the age effect. This inconsistency could indicate that the influence of school starting age is not stable over time and may be affected by external factors. A closer investigation of changes in educational policies or assessment methods during the periods where no significant effects were observed could shed light on these variations. However, we remain agnostic about the mechanisms behind our causal estimates.

Another limitation is the lack of analysis of the mechanisms driving the age effect. Although older students may perform better due to their cognitive and emotional maturity, the data does not empirically support these hypotheses. Investigating specific channels, such as cognitive readiness, emotional development, or even peer interactions, could provide a clearer understanding of why being older at school entry positively impacts educational outcomes.

Moreover, cultural norms surrounding school starting age differ significantly between countries. In some countries, delaying school entry might be an accepted practice, while in others, it might be less common. These norms could influence the magnitude and direction of the effect. Exploring within-country regional differences or conducting qualitative analyses could reveal how societal expectations shape the decision to delay school start and its educational outcomes.

Another key limitation is the focus on test scores as the primary outcome, which only provides a snapshot of students' abilities at a specific point in time. Longitudinal data tracking students throughout their educational journey could reveal whether the positive effects of starting school later persist into higher educational attainment, career success, or lifelong learning outcomes.

As an additional exercise, we conduct regressions that take grade repetition into account to determine whether it may influence our findings. This stems from the fact that older students might be more likely to have repeated a grade, which could affect their performance and taint the perceived benefits of being older.

To address this, we propose two approaches. First, we include a dummy variable in the regression equation to capture the differences between students who have repeated a grade and those who have not. Second, we consider being a repeater as an outcome variable to evaluate whether starting school at an older age impacts the likelihood of repeating a grade. However, we must be cautious in interpreting this coefficient, as students who repeat a grade may differ significantly from those who do not.

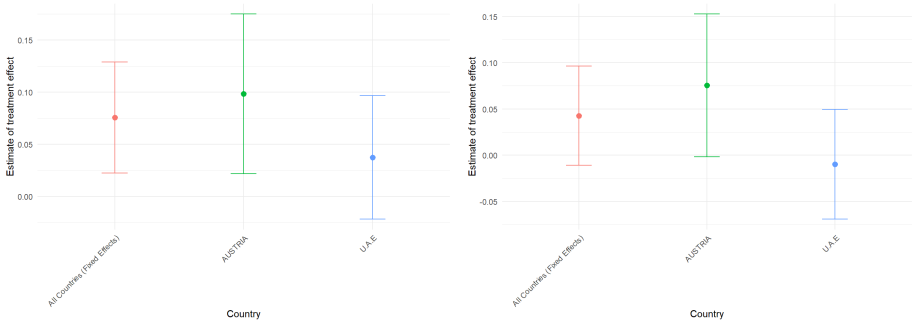
Figure 12 illustrates that controlling for students who have repeated at least once yields consistent results regarding the advantages of starting school later when we pool the data for all waves. Specif-

Table 3: Estimation Results

	Standardized Reading Score				Standardized Math Score			
Country	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Austria	0.032	0.198***	0.073	0.104***	0.099	0.185***	0.095	0.129***
	(0.069)	(0.068)	(0.071)	(0.040)	(0.070)	(0.067)	(0.072)	(0.040)
U.A.E	-0.061	0.004	0.019	-0.009	0.031	0.027	0.050	0.038
	(0.055)	(0.055)	(0.050)	(0.031)	(0.058)	(0.052)	(0.049)	(0.031)
All countries	-0.001	0.128***	0.054	0.064**	0.075	0.127***	0.079	0.097***
	(0.049)	(0.047)	(0.049)	(0.028)	(0.050)	(0.046)	(0.050)	(0.028)

*Note: *p<0.1; **p<0.05; ***p<0.01*

Eicker-Huber-White (EHW) heteroskedasticity-robust standard errors in parenthesis. Columns (1) and (5) report results only considering wave 2015. Columns (2) and (6) report results only considering wave 2018. Columns (3) and (7) report results only considering wave 2022. Columns (4) and (8) report results considering the three waves.



(a) Math (b) Reading

Figure 12: Effect of School-starting age on educational outcomes (controlling for grade repetition)

ically, older students tend to perform better on the PISA assessments compared to younger starters in the case of Austria, but not in the case of the United Arab Emirates. Nevertheless, we cannot determine whether repeating a grade earlier or later affects academic performance, as our analysis only accounts for students who have repeated at least once.

Moreover, we regress a dummy that takes value 1 if the individual repeated at some point or 0 otherwise on being born either in August or September (that is, our running variable). Figure 13 shows that those students who are older are less likely to be repeaters in the case of Austria, and no evidence was found for U.A.E. Again, we show the results pooling all waves. This finding is consistent with our earlier results that older students generally perform better.

Lastly, it is important to consider sample size and statistical power, especially in the puzzling cases where significant results were only observed for one wave (or none). Keeping the window for one month (before and after) gives us less statistical power than using for instance a larger window such as two months. In this regard, smaller sample sizes could limit the power to detect significant effects in certain years. Future studies could benefit from pooling waves for specific countries to assess whether larger sample sizes would reveal more consistent effects.

2.6 Conclusion

In this exercise, we attempt to find the effect of school starting age on academic performance by using a regression discontinuity design framework. We consider data for 2 countries, Austria which is a OECD country, and U.A.E. as a non-OECD using PISA data for waves 2015, 2018, and 2022. The identification assumption in our project is that the arbitrary criteria for enrolling students in primary school create a discontinuity in the age of students enrolled in a class.

We observe positive effects of age on academic performance at the aggregate level and results do

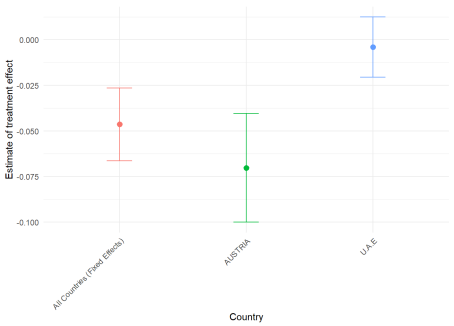


Figure 13: Effect of School-starting age on grade repetition

not change much including country-level fixed effect. However, the effect varies widely across countries. These results are in line with current literature. The procedure had some limitations. The running variable (birth month of the students) is discrete in nature, hence one cannot exploit the information further away from cutoff points. For this reason, we use the mean difference in mathematics and reading scores on either side of the cutoff.

To sum up, the findings indicate that a later school start time positively and significantly influences test scores for Austria, whereas for U.A.E. we do not find any effect. However, to comprehensively understand the underlying mechanisms and ensure the robustness of these effects, further research is necessary. Future studies should explore differences across countries, examine causal relationships, and evaluate the long-term effects of school-starting age on broader educational and life outcomes.

References

Andrews, R. J., S. A. Imberman, and M. F. Lovenheim (2020). Recruiting and supporting low-income, high-achieving students at flagship universities. *Economics of Education Review 74*, 101923.

Avvisati, F. and P. Givord (2021). How much do 15-year-olds learn over one year of schooling? an international comparison based on pisa.

Black, S. E., J. T. Denning, and J. Rothstein (2023). Winners and losers? the effect of gaining and losing access to selective colleges on education and labor market outcomes. *American Economic Journal: Applied Economics 15*(1), 26–67.

Buckles, K. S. and D. M. Hungerman (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics 95*(3), 711–724.

Dhuey, E., D. Figlio, K. Karbownik, and J. Roth (2019). School starting age and cognitive development. *Journal of Policy Analysis and Management 38*(3), 538–578.

Dubow, E. F., P. Boxer, and L. R. Huesmann (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer quarterly (Wayne State University. Press) 55*(3), 224.

Fuchs, T. and L. Wößmann (2008). *What accounts for international differences in student prformance? A re-examination using PISA data*. Springer.

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

Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspectives in clinical research 2*(3), 109–112.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences 44*(1), 1–12.

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The review of economics and statistics 91*(4), 717–724.

Kozakowski, W. (2020). Are four-year public colleges engines for mobility? evidence from statewide admissions thresholds. In *2020 APPAM Fall Research Conference*. APPAM.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of economic literature 48*(2), 281–355.

Long, M. C. and M. Tienda (2008). Winners and losers: Changes in texas university admissions post-hopwood. *Educational evaluation and policy analysis 30*(3), 255–280.

Matthews, G., I. J. Deary, and M. C. Whiteman (2003). *Personality traits*. Cambridge University Press.

OECD (2023). *PISA 2022 Results (Volume I)*.

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

Phelps, R. P. (2006). Characteristics of an effective student testing system. *educational HORIZONS 85*(1), 19–29.

PIRLS (2022). National policies on age of school entry and promotion.

Schleicher, A. and C. Tamassia (2000). *Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy. Education and Skills.* ERIC.

Van Bragt, C. A., A. W. Bakx, T. C. Bergen, and M. A. Croon (2011). Looking for students' personal characteristics predicting study outcome. *Higher Education 61*, 59–75.

Zimmerman, S. D. (2014). The returns to college admission for academically marginal students. *Journal of Labor Economics 32*(4), 711–754.