Econometric Game 2024

Closing Doors, Widening Gaps: Exploring the Impact of COVID-19 School Closures on Students from Low Socio-Economic Backgrounds

Cil Bemelmans **Tobias Kers** **Stan Koobs** **Sam van Meer**

Abstract

This paper investigates the causal effect of not attending school on educational outcomes in math and reading, using Pisa data. We do this by investigating the effects of school closures due to the COVID-19 pandemic. We analyse Canadian highschools due to their large regional heterogeneity in lockdown policy. Through a matching algorithm and regional difference-in-difference analysis, we find that school closures had a large negative effect on educational performance. This effect was particularly strong for students from lower socio-economic backgrounds. Our results underline the importance of accessibility to education for these subpopulations.

Keywords: returns to education, school closures, difference-in-differences, asymmetric panel data

1 Introduction

The benefits of education are obvious, both for individuals and society as a whole. This becomes especially obvious when studying children who do not complete school: they suffer more from substance abuse and are arrested more often, even when controlling for factors such as age and income (Maynard et al., 2015; Thornberry et al., 1985).

Considering how crucial it is for children to go to school, the school closures caused by the COVID-19 pandemic were potentially very dangerous: suddenly all children were not allowed to go to school anymore. From the beginning of the pandemic deciding whether or not to close school has been a topic of fervent debate Wen (2020). Still, education did not come to a complete halt for students as schools switched to remote learning, but the longer-term effects of this change still remain to be seen, as especially for primary and secondary schools, this was an unprecedented shift.

Early signs are worrying. PISA data, collected by the OECD, show that in 2022 students' scored on average worse in reading, math and science compared to 2018. Only three countries in their set of 80 performed better in all three subjects, meaning a stark reversal of the trend of gradually increasing scores before 2018 OECD (2022). The pandemic seems the obvious factor, but the exact impact of school closures still remains to be investigated, as so many parts of society were upended by the pandemic.

Crucially, not every student was impacted by school closures equally. For one, not every school was closed for the same amount of time. Different countries made different decision regarding whether to close schools or not, both based on infection level and political considerations. Even within countries, there was strong variations, as countries like Australia left COVID policy mostly up to local governments. Furthermore, students from lower socio-economic backgrounds were potentially impacted more by school closures, with early evidence showing inequality in education outcomes to have increased (Haeck and Lefebvre, 2020; Maldonado and De Witte, 2022).

To investigate the causal effect of not attending school on educational performance and inequality, we use a novel Difference in Differences effect estimation procedure that allows for different schools in both time periods and a continuous treatment effect which will be the number of days the school was closed. To do this, we introduce an asymmetric panel data model. We provide several conditions which guarantee consistency of our estimator and focus on of these which is based on the idea that we match schools which have similar unobserved characteristics. To do this, we employ a novel matching algorithm coming from the Operations Research literature. Subsequently, we use the framework developed by Callaway et al. (2024) to allow for a continuous treatment variable. Lastly, we employ machine learning methods to detect possible interactions between the school closing days and other variables in our dataset.

In our analysis, we want to cluster the standard errors per region in our country. This leads us to the decision to investigate Canada in depth. For one, whether to close schools or not was decided on a province-level, leading to strong heterogeneity Paling (2021). Secondly, Canada is one of the few countries in the PISA dataset that not only contains data on what region a school is in, but also has 10 or more distinct regions.

As the OECD selects the schools for its survey randomly every testing wave, it is not possible to compare performances in the same school in 2018 and 2022. To combat this, we apply a matching algorithm, where we match schools in the 2022 dataset to the 2018 dataset based on several key school characteristics.

Using our estimation strategy, we find that longer school closure has a negative effect on educational outcomes. This effect is mostly driven by students coming from a poor background or having low-educated parents. This shows that when evaluating educational policies, it is crucial to consider different subgroups in societies.

Section 2 deals with the data, provides descriptive statistics, both on world-level and specifically for Canada. Then Section 3 discusses our methodology, first to find the matches and then to estimate the causal effect of not attending school on educational outcomes. Section 4 gives our results and then Section 5 rounds off this paper by providing the conclusion and discussion and giving policy recommendations.

2 Data

The OECD perform a wide range of surveys every testing wave. It is most known for assessing students' academic ability in three subjects: reading, math and science. However, it also surveys students about their well-being. Furthermore, it also surveys school principals.

By also surveying school principals the OECD gains new insights about the schools at which students are enrolled. School principals are asked both about school characteristics and for 2022 were specifically asked a swathe of questions related to the pandemic, asking about the amount of school closures and how the school tackled remote learning. Data on students also contains their school ID, which allows us to link their results, to the answers input by their school principal.

Figure 2 shows that there is very large variation in the number school closure days (since the start of the pandemic) between rich and poor countries with much more closure days caused by COVID-19 for richer countries. Furthermore, the large differences between countries in terms of education standards and wealth means there is a large violation of parallel trends when using global data to estimate the effect of school closure days on student performance. The parallel trend assumptions is much stronger across schools within the same country, so we choose to analyse the relation instead.

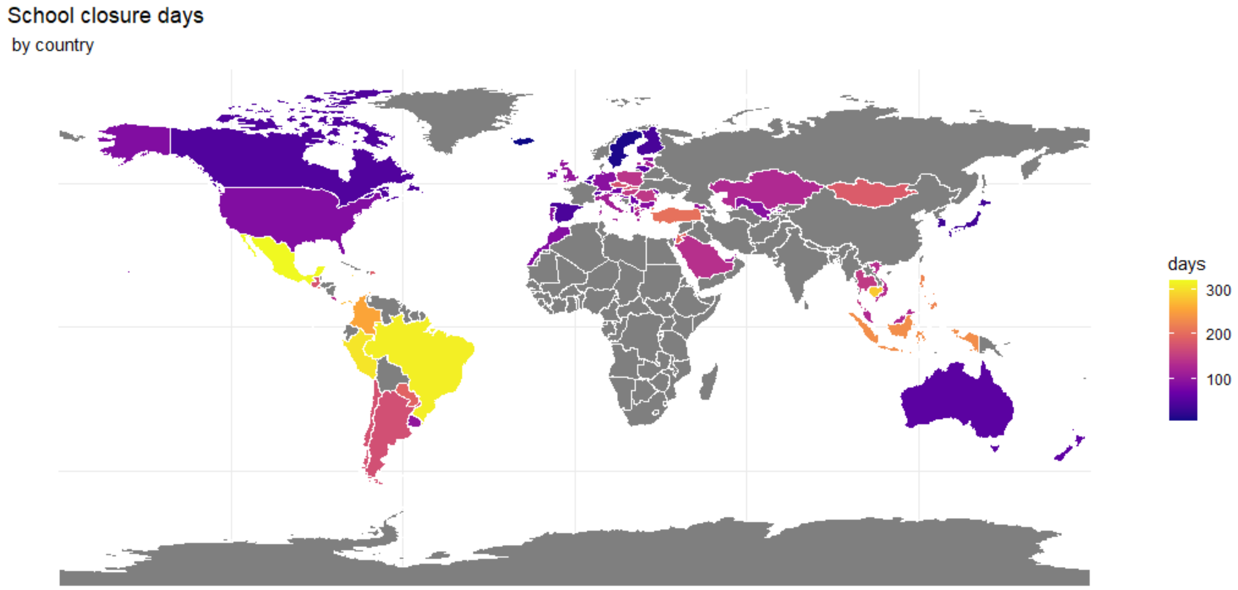

Figure 2: Days school closed due to COVID-19 across regions across the globe.

In our analysis, we want to cluster the standard errors per region in the country of interest, so the country should be split into different regions in the PISA data. There should also be enough heterogeneity in school closure days both across regions and within regions to ensure there is enough variation to estimate the effect of closure days on performance accurately. Canada meets these criteria as the PISA survey splits the country up into its ten provinces denoted by codes ranging from 12400 to 12410. Figure 3 shows that there is both large heterogeneity in the mean and variance in closure days across regions as well as large variation between schools within regions. The situation exists because whether to close schools or not was decided on a province-level in Canada Paling (2021). Table 2 in the appendix contains information on which code refers to which Canadian province.

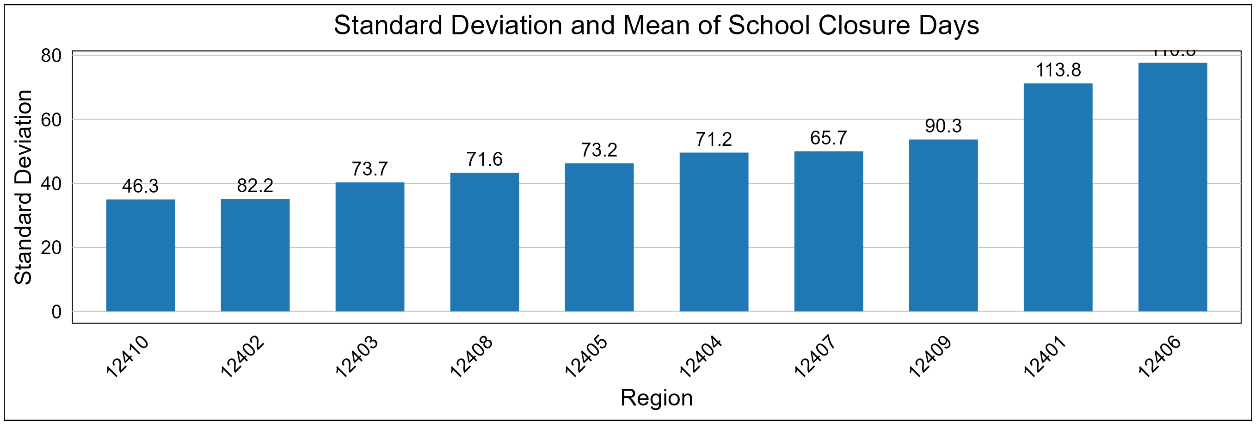

Figure 3: Standard deviation in days school closed due to COVID-19 across Canadian regions. The numbers above the bars are the mean closure days in a region.

Continuing the analysis of Canada, we shift our focus to the trend in test scores before and after the temporary closing of schools. between 2018 and 2022 Figure 4 shows both math and reading scores decreased substantially between 2018 and 2022, suggesting that the closure days might have had an adverse effect on the performance of students. The figure also shows large variation in average test scores across regions, though there was no large increase in the magnitude of disparities variation in 2022 relative to 2018. Finally, there are large differences between top and bottom students across all regions in both reading and math for both 2018 and 2022, with no clear increase or decrease in the differences between the two waves.

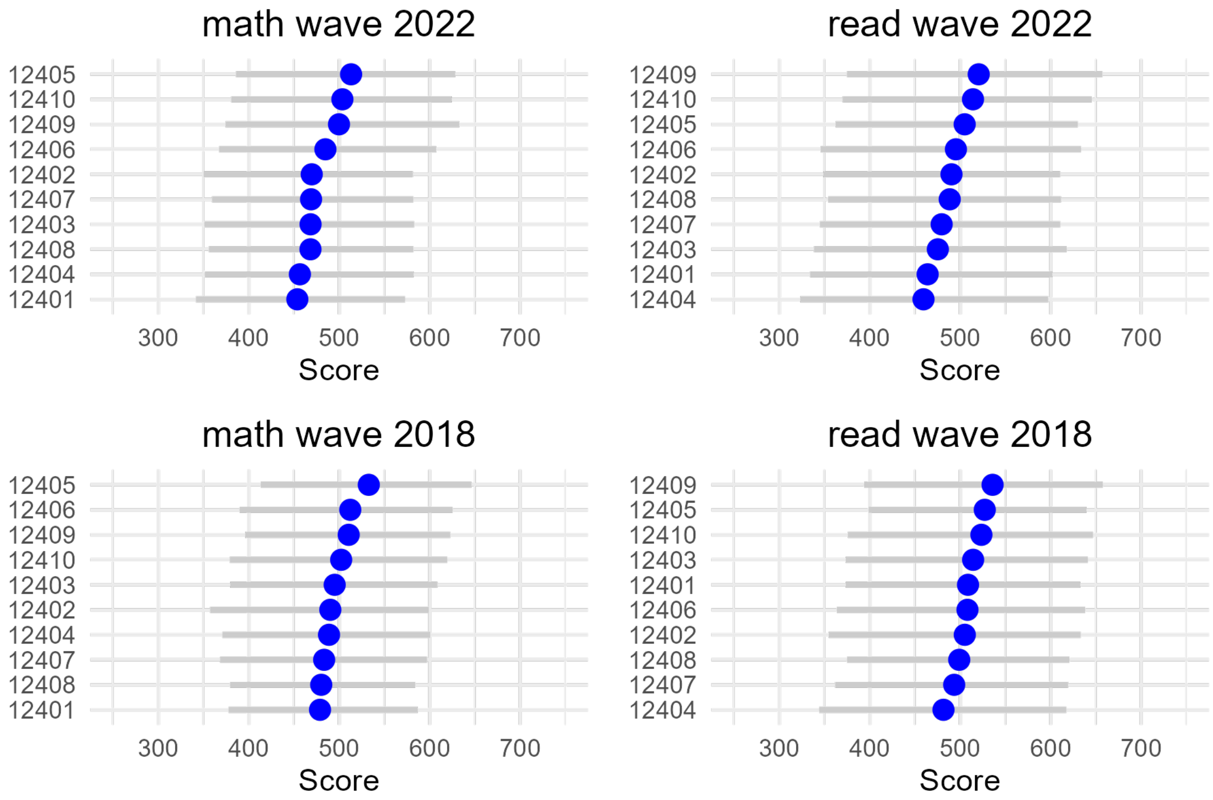

Figure 4: Reading and math scores in 2018 and 2022. The grey line denotes the 10th to 90th quantile, while the blue dot is the 50th quantile.

To analyse the effect of school closure days on student performance we divide the schools into three bins. These bins are based on the number of days the school was closed due to Covid since the pandemic and the bins are 0 to 50 days, 50 to 150 days and 150 days or more. We choose these bins such that all have enough schools.

In figure 5 we see that students who spent the least amount of time at home performed the best with average scores of 489 and 494 in math and reading respectively. Schools in the medium group with 75 and 150 score averages of 484 and 491, while the group with the most closure days scores the lowest of all, with average scores of 484 and 490. This means the difference in the reading scores is larger than the difference in math scores.

Figure 5 also shows first and ninth quantile. We see that the gap between these two is large and remains relatively constant for both subjects and bins.

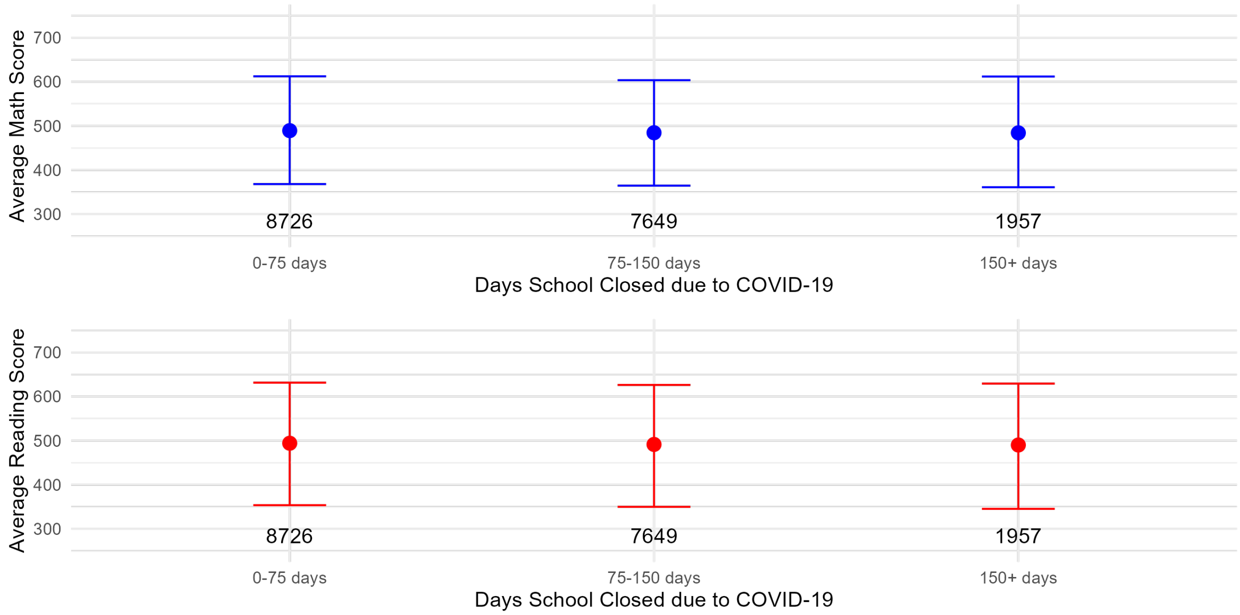

Figure 5: Days school closed due to COVID-19 and reading and math scores in 2022. The lines denotes the 10th to 90th quantile, while the dot is the 50th quantile. The numbers below the intervals are the number of students in the bins for 2022

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, denoted in the graph by *paredint*. First of all, we see a positive relation between the education level of the parent and the average score of the student. When considering the school closure days, we see that in the lowest quintile students with less closure days score much better than the other groups, both in reading and math. The scores start to converge as we move to a higher quintile. For the highest quintile, the scores are even almost equal for the different bins.

Next, we study intergenerational mobility by considering the income of parents instead of their education. PISA does not ask students the exact income of their parents, but instead asks students how many different possessions their family owns. On the basis of that number, an assessment can be made into the income of the parents. Like with education, we see that students from wealthier families score better. Furthermore, the group with the most closing days scores the worst, while the group with the least closing days scores the best. Contrary to education, this difference is scores seems to persist as we move to higher quantiles. Furthermore, it appear to remain relatively constant.

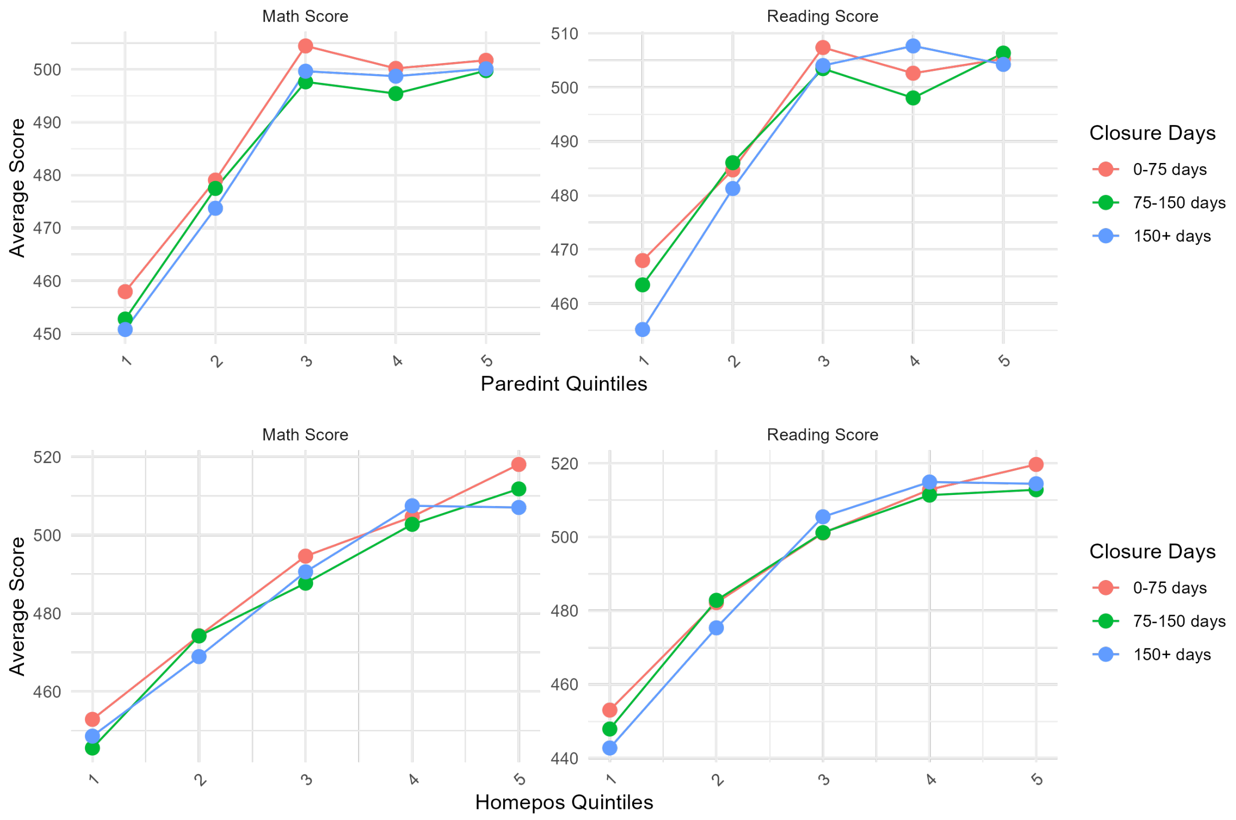

Figure 6: Intergenerational Mobility and Average Performance in Reading and Math. PAREDINT indicates the years of education received by the most educated parent. HOMEPOS indicates the income decile using home possessions as proxy.

3 Methodology

In this section, we introduce our novel Difference in Differences effect estimation procedure that allows for different units in both time periods and a continuous treatment effect. We first introduce an asymmetric panel data model that allows for different units in different time periods and we formulate conditions that guarantee consistency of the estimator. Furthermore, we elaborate on how to symmetrize the panel data set by employing an innovative matching algorithm. Because our treatment variable is not binary as it indicates the number of closing days because of COVID, we subsequently discuss a Difference in Differences procedure that allows for continuous treatment. Finally, we discuss how these elements can used to estimate the causal effect of school closing days and how complex interactions can be detected by using machine learning methods.

3.1 Asymmetric panel data model

We will first introduce our general model. In particular, we focus on two time-period panel data models where we allow for different units in the two time-periods. We refer to such a panel data set as an *asymmetric* panel data models. This is related to what in the literature is referred to as asymmetric fixed-effects (Allison, 2019). However, the difference is that these models only allow for different fixed effects whereas we consider different units. This means that we also allow for different assumptions on the corresponding error terms. Up to our knowledge, this model is novel in the literature of panel data modelling.

Let us define our framework mathematically. Let Y_{i1} be the outcome of unit i in time period 1 and let Y_{j2} be the outcome of unit j in time period 2. Moreover, let x_{i1} and x_{j2} be the corresponding vectors of regressors. Here, we assume that both $i, j \in \{1, \ldots, N\}$. However, note that we assume that i and j can be different units, which is different from the classical panel data literature (Hsiao, 2022; Baltagi, 2008) and therefore we coin this model an asymmetric panel data set. Moreover, for simplicity, in this study we only focus on two time periods. We assume the following linear panel data model

$$\begin{aligned} Y_{i1} &= \alpha_i + x_{i1}'\beta + \epsilon_{i1} \\ Y_{j2} &= \alpha_j + x_{j2}'\beta + \epsilon_{j2}, \end{aligned}$$

where α_i and α_j are the fixed effects of units i and j, respectively. Next, ϵ_{i1} and ϵ_{j2} denote the corresponding error terms and β is the vector of parameters of interest. When we employ a first-difference (FD) transformation, we obtain the following equation

$$Y_{j2} - Y_{i1} = \alpha_j - \alpha_i + (x_{j2} - x_{i1})\beta + \epsilon_{j2} - \epsilon_{i1}.$$

When we denote $v_{i1} = \alpha_i + \epsilon_{i1}$ and $v_{j2} = \alpha_j + \epsilon_{j2}$, we can also rewrite this to the following form

$$Y_{j2} - Y_{i1} = (x_{j2} - x_{i1})\beta + v_{j2} - v_{i1}. \tag{1}$$

3.2 Consistent estimation

In this subsection, we will provide sufficient conditions for consistent estimation of the parameters in the asymmetric panel data model. Firstly, we assume that all standard assumptions from linear regression hold. In particular, we assume that all regressors are orthogonal to the corresponding error terms. However, for consistent estimation of (1), is not sufficient. This is because we need to have some assumption on the correlation between the fixed effects and the regressors. To be precise, for consistent estimation of (1), we require,

$$\mathbb{E}((x_{j2} - x_{i1})(v_{j2} - v_{i1})) = 0.$$

Under the assumptions that all regressors are orthogonal to all error terms, this boils down to the requirement that

$$\mathbb{E}((x_{j2} - x_{j1})(\alpha_j - \alpha_i)) = 0. \tag{2}$$

We now provide three conditions which all sufficiently under which (2) is satisfied.

Condition 1: *(Equal units)*
When $i = j$, we have that $\alpha_j - \alpha_i = 0$ and (2) holds. Under this condition, our model boils to a classical linear panel data model where we apply first-differencing.

This first condition is the trivial case and not of particular interest but it demonstrates the link between our framework and the classical panel data framework.

Condition 2: *(Random effects)*
Under the standard assumptions of random effects (RE) models and the assumption that $\mathbb{E}(x_{j2}\alpha_i) = 0$ *and* $\mathbb{E}(x_{i1}\alpha_j) = 0$*, it follows that* (2) *is satisfied.*

It is trivial to see that the above assumptions guarantee sufficiency. However, in most applications, these assumptions do not hold. In practice, one is usually interested allowing for correlation between the unobserved heterogeneity and the regressors. We allow for this in the last condition we coin the *symmetry* condition.

Condition 3: *(Symmetry)*
Assume that $\alpha_j - \alpha_i = u_{ij}$ *where* $u_{ij} \perp\!\!\!\perp (x_{j2} - x_{i1})$ *and* $\mathbb{E}(u_{ij}) = 0$ *for all* $i, j \in \{1, \ldots, N\}$*. This guarantees that* (2) *holds.*

The last of these three conditions is novel and will provide the basis for the rest of our methodology. We will provide further intuition in the next subsection.

3.3 Symmetrization by matching

In most applications, the symmetry condition is likely not to hold when units i and j are not the same. This is because when units i and j are different, they are also different in terms of unobserved characteristics. However, our approach will be to provide an innovative matching procedure that allows us to match units based on observed characteristics. When the observed characteristics are highly similar, we assume that the difference between the unobserved characteristics of the two units only amounts to a random noise term u_{ij} as in the symmetry condition. Thus, the key for this condition to hold is that we need to find units that are very similar. This is related to the ideas developed in the field of Propensity Score Matching (Caliendo and Kopeinig, 2008) where units are matched using their propensity to be treated. This is not suited for our framework as we also want to allow for continuous treatment effects.

3.4 Matching algorithm

As the OECD selects schools randomly, it is not possible to compare performance of the same schools between 2018 and 2022. Therefore we implement a matching algorithm. The goal of this algorithm is to match schools from the 2018 dataset to school in the 2022 dataset.

3.4.1 Selection criteria

This matching is done on the basis of several criteria. Firstly, considering the importance of regional heterogeneity in later parts of our analysis, we only match schools that are in the same region. Then, we consider the following variables in our selection procedure:

- Size of the community where the school is located, which is a categorical variable ranging from 1 (less than 3000 inhabitants) to 6 (more than 10 million). (*SC001Q01TA*)

- The type of funding a school receives, either public or private, with corresponding variable 1 or 0 respectively (*SC013Q01TA*).
- The amount of schools in the region that compete for students, equal to either 0,1,2 or 3 (*SC011Q01TA*).
- Whether the schools groups students by ability, 1 if yes, 0 if not (*SC042Q01TA*).

Then, for every pair of school i from 2018 and school j from 2022 we calculate the matching scores s_{ij} by considering the Euclidean distance. Here, k denotes the index of the variables, which ranges from 1 to 4 as we use 4 different variables in our selection procedure.

$$s_{ij} = \sum_{k=1}^{4} (x_i^k - x_j^k)^2 \tag{3}$$

Our goal is to find matches where the matching score is as low as possible, which indicates that schools' characteristics match closely. We want to ensure that every school, either from 2018 or 2022 is matched to exactly one other school. However, it is not possible to match every school, since both datasets have different sizes. Furthermore, every region has a different number of observations in 2018 and 2022 and since matches must be in the same region the maximum number of matches is limited.

3.4.2 Binary programming formulation

We can use a binary programming formulation to display how to find the optimal matches. We first must define the variables y_{ij}, which is a binary variable and equal to 1 if school i (from 2018) is matched to school j from 2022. Lastly, n_1 and n_2 denote the number of schools in 2018 and 2022 respectively.

$$\begin{aligned}
&\min \sum_{i}^{n_1} \sum_{j}^{n_2} s_{ij} y_{ij} && (1)\\
\text{Subject to: } &\sum_{j}^{n_2} y_{ij} \leq 1, \forall i = 1, 2, .., n_1 && (2)\\
&\sum_{i}^{n_1} y_{ij} \leq 1, \forall j = 1, 2, ..., n_2 && (3)\\
&\sum_{i}^{n_1} \sum_{j}^{n_2} y_{ij} = c && (4)\\
&y_{ij} \in \mathbb{Z}_{\not=} && (5)\\
& && (6)
\end{aligned}$$

In this formulation, constraints 2 and 3 make sure that every school is matched to at most one other school. They are inequality constraints instead of equality constraints, because not every school will be matched. Constraint 4 makes sure the total number of matches is equal to c. The last set of constraints make sure that all y_{ij} variables, which indicate matchings, are equal to either 1 or 0, meaning that a match must either be fully included or not at all.

However, the total number of potential matches is very high, disregarding the regions this number is around 800^2 as every dataset contains around 800 schools. This implies that there are so many variables in our formulation that it is not computationally feasible to solve this problem to optimality, as the running time would be too high. We must therefore explore heuristic solutions, that can find a set of matching within an acceptable time.

3.4.3 Heuristic

To come up with a good matching, we apply a greedy heuristic. This means that for every school in 2022 we find which schools from 2018 are in the same region. We calculate the matching score according to equation (3). We add the matching with whichever school from 2018 the school from 2022 has minimal score. This does mean that schools from 2018 can be matched to multiple schools from 2022, but every school from 2022 can only be matched to one school from 2018. This is not ideal, but we can deal with this by clustering observations later on.

3.5 Difference in differences with a continuous treatment

We first provide a brief introduction into the classical Difference in Differences (DiD) framework which assumes a binary treatment. Afterwards, we elaborate on using DiD for continuous treatments.

3.5.1 Difference in Differences

Difference in differences (DiD) is a highly popular technique used in many domains of science to study causal effects in observational studies (Card and Krueger, 1993; Bertrand et al., 2004). It compares the impact of a treatment on a designated treatment group against a control group within the context of a natural experiment. The effect of the treatment is measured by assessing the average change over time in both the treatment and control groups. By taking the difference between their differences over time, we obtain the DiD estimator. Similarly, we can the obtain the DiD by applying regression on the difference in outcomes and differences in regressors and the treatment.

For simplicity of notation, we now only consider $i \in \{1, \dots, N\}$ as we assume that we have matched i and j. Then given the model

$$\begin{aligned} Y_{i1} &= \alpha_i + x'_{i1}\beta + D_{i1}\tau + \epsilon_{i1} \\ Y_{i2} &= \alpha_i + x'_{i2}\beta + D_{i2}\tau + \epsilon_{i2}, \end{aligned}$$

where D_{it} denotes the treatment for unit i in period t and τ denotes the treatment effect. Taking first differences subsequently gives

$$\Delta Y_{i2} = \Delta x'_{i2}\beta + \Delta D_{i2}\tau + \Delta\epsilon_{i2}.$$

Note that for binary treatment we have that $\Delta D_{i2} = 1$ for the treatment group and $\Delta D_{i2} = 0$ for the control group. By applying standard least-squares regression techniques we then obtain the DiD estimator. Note that this is similar taking first-differences in a two-period panel data model which shows that we can apply the results from Section 3.2.

3.5.2 Continuous treatment

We now focus on the DiD for continuous treatment. Here, we follow the methodology from the recent paper by Callaway et al. (2024). A key difference with binary treatment is that there is not a single treatment parameters anymore but there are more potential parameters of interest. First, let us introduce some more notation. Let Y_i now denote the observed outcome for unit i in period 2. Formulating our framework in terms of the Rubin Causal Model (RCM) (Rubin, 1974), let $Y_i(0)$ and $Y_i(d)$ denote the pair of potential outcomes for unit i, where $Y_i(0)$ is the outcome without the treatment and $Y_i(d)$ is the outcome with $D = d$. The main issue in causal inference is that we never observe both of these.

We now define the first parameters of interest. That is, we define

$$ATE(d) = \mathbb{E}(Y_i(d) - Y_i(0)) \quad \text{and} \quad ACR(d) = \frac{\partial ATE(d)}{\partial d}.$$

Here, $ATE(d)$ denotes the average causal effect of receiving dose d of the treatment and $ACR(d)$ denotes the causal response of a marginal change in the dose of treatment d.

To identify the above parameters, the key identification restriction that Callaway et al. (2024) provide is the *strong parallel trends* (SPT) assumption which says that

Condition C1: *(Strong Parallel Trends Assumptions)*
For all possible values of d,

$$\mathbb{E}(Y_{i2}(d) - Y_{i1}(0)) = \mathbb{E}(Y_{i2}(d) - Y_{i1}(0)|D_i = d).$$

This is a relatively strong assumption. It says that the average evolution of outcomes for the whole population if all get treated by dose d (left-hand side of equation) is equal to the path of outcomes that dose group d actually experienced. In terms of our study, this says that the average of effect all school closing on the same number of days in our data, is the same as the average effect on the schools that in fact closed on this number of days. Thus, this path of outcomes that closed on these days, is representative for the whole population. This is a strong assumption but it does not seem highly unrealistic for our study and we therefore employ it.

Under this condition and some other standard conditions in the paper, it can be shown that

$$ATE(d) = \mathbb{E}(\Delta Y_{i2}|D_i = d) - \mathbb{E}(\Delta Y_{i2}|D_i = 0).$$

The right-hand side of the above equation is something we can estimate as Y_{i2} and Y_{i1} are observed. In particular, the original paper employs the following regression model

$$\Delta Y_{i2} = \beta_0 + \sum_{k=1}^{K} 1_{D_i=d_k}\beta_k + \Delta\epsilon_{i2}. \tag{7}$$

It has been establisehd that under the SPT assumption, each $\widehat{\beta}_k$ is a consistent estimator for $ATE(d_k)$ and $\widehat{\beta}_k - \widehat{\beta}_{k-1}$ is a consistent estimator for $ACR(d_k)$. Furthermore, they mention that this generally works well when the dose groups are large (i.e. the treatment is not continuous but discrete and only take few values). When the treatment is continuous or almost continuous (discrete but taking many different values) this does not work well due to the high number of parameters. To accommodate this scenario, they provide an alternative nonparametric estimation strategy which is based on approximating the functional form using splines. However, we will now only focus on a treatment policy for now that takes a small number of discrete values. How this can be applied to the topic of this particular paper will be elaborated upon in the next subsection.

3.6 Measuring the causal effect of school closing days

We are now ready to discuss the exact model applied in this paper. As we are dealing with different schools per wave of the PISA data, we need to ensure that some of the assumptions from Section 3.2 holds. For this, we follow the symmetrization procedure described in Section 3.3 and the algorithm from Section 3.4. This leaves us with a smaller set of N units that we observe in both time periods and have similar unobserved characteristics.

Let Y_{it} denote the average test outcome (either math or reading) from students at school i at period t. Moreover, let D_{it} denote the bin number of the corresponding number of school closing days of school i at time period t. For example, $D_{i,2022} = 1$ means that school i in 2022 was closed for 0-75 days. Moreover, let x_{it} denote a vector of other regressors for school i and period t. We then employ the following regression model

$$\Delta Y_{i,2022} = \sum_{k=1}^{3} 1_{D_{i,2022}=k}\beta_k + \Delta x'_{i,2022}\theta_1 + \sum_{k=1}^{3} 1_{D_{i,2022}=k}\Delta x'_{i,2022}\theta_2 + \Delta\epsilon_{i,2022}, \tag{8}$$

where θ_1 and θ_2 denote vectors of parameters. Also note that we do not include an intercept in this model as all schools fall into any of the bins and we thus cannot have an intercept due to identification restrictions.

3.7 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 using model (8) where we use variables such as home possession and parent education as other regressors. 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.

4 Results

In this section, we first look at the causal effects of school closing days using the DiD design described above. Furthermore, we consider the interactions of the school closing days with other different explanatory variables. We select these explanatory variables using Shapley values resulting from the XGBoost model.

Firstly, we are interested in the direction and strength of the effect of the school closing on the later educational outcomes. To study this, we first employ the estimation procedure discussed in Section 3.6, including home possession and parent education as possible regressors and interactions. The results of this can be found in Table 1. In the first column, we have regressed the treatment dummies of school closures without any other regressors. We note that all of these dummies have a negative effect and that the effect is the strongest for high closure. This means that as the number of days increases in which the school is closed, the reading scores of the students go down. This is line with our hypothesis: sitting at home or schooling which is done online is worse for the students than having education at school.

Table 1: Regression Results

	Dependent variable: Change in reading scores			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Low closure	-13.955***	-25.864***	-194.283***	-98.182*
	(3.072)	(3.665)	(51.495)	(54.752)
Medium closure	-16.354***	-28.061***	-196.464***	-100.186*
	(3.393)	(3.829)	(51.332)	(54.617)
High closure	-20.425***	-44.537***	-824.435***	-654.490***
	(6.977)	(9.655)	(191.693)	(225.662)
HOMEPOS		35.576***		31.853***
		(6.259)		(6.833)
HOMEPOS×high		36.492*		2.033
		(21.244)		(25.571)
PAREDINT			12.137***	4.955
			(3.464)	(3.743)
PAREDINT×high			41.903***	36.899**
			(13.333)	(15.917)
Observations	727	709	709	709
R^2	0.068	0.127	0.109	0.138
Adjusted R^2	0.064	0.120	0.102	0.129
Note:			*p<0.1; **p<0.05; ***p<0.01	

Moreover, it is interesting to see whether these effects are stronger or weaker for particular subgroups of the population. To see which particular subgroups likely have the strongest interaction effect with the closure variable, we have run an XGBoost models using several variable. From ordering the absolute Shapley values we found that the largest interaction effects came from the variables HOMEPOS and PAREDINT. Therefore we solely focus on these variables in the rest of the results.

The results of the models including interaction effects can be found in Table 1. For model 2, we note that the coefficient of the interaction between HOMEPOS and high closure is positive and significant. This means that the effect of school closing on education outcomes of students from wealthy backgrounds is less severe. The negative effect in the first column seems mostly driven by the strongly negative effect for students from poor backgrounds. Next to this, model 3 includes PAREDINT as an extra regressor instead of home possession. Here, we also observe that students who have parents with a strong educational background, experience a smaller effect from school closures on their subsequent educational performance. Lastly, model 4 includes both variables as interactions which shows that HOMEPOS becomes insignificant while PAREDINT is still significant. It is likely that there is a strong correlation between how wealthy a family is and their educational background. However, for later educational outcomes of the children, it seems like the parental education is the strongest driving factor here and when controlling for this, the home possession becomes insignificant.

5 Discussion and recommendations

In this paper, we investigated the causal effect of not attending school on subsequent educational outcomes. In particular, we used the PISA dataset which consists of test outcomes for several subjects for 15 year old students.

To test this, we made use of the large number of school closures due to the COVID-19 pandemic. As not every school closed for the same amount of time, this allowed us to find the effect of not attending school by comparing groups of students that either barely missed school or missed a lot of school.

The pandemic affected every country in unique ways, which makes comparisons between countries harder. Therefore, we opted to just look at one specific country Canada and then consider its regions. We chose Canada as the PISA dataset contains information about the region and there was strong variation in the amount of school closure days both between and within regions.

We first match schools in 2022 with schools in 2018 through a novel matching algorithm, which allows for unobserved heterogeneity and allows us to perform a difference-in-difference analysis. We find that more school closure days has a significantly negative effect on reading scores. Furthermore, this effect is especially strong for students with lower socio-economic status.

On the basis of these results, we recommend that governments must do as much as possible to prevent school closures, which can also happen due to labor strikes. However, were something similar to a pandemic to happen again, it must pay special attention to disadvantaged students, to make sure educations remains accessible for these students. Countries could consider opening schools in smaller classes to accomodate these more vulnerable students.

In terms of methodology, one limitation of our work is that we aggregate the number of closing days of school caused by COVID and sort them into bins. This leads to a loss of information. It would be interesting what the results would be when we do not aggregate and apply the nonparametric estimation strategy from Callaway et al. (2024) which is based on approximating the functional form using splines. Another methodological drawback of our work is that we need to impose the strong parallel trends assumption. For this application, it is dubious whether this assumption holds. More work would be needed to study the validity of this assumption. Lastly, in our current research we have just looked at the effects in Canada. More research should be done to study to what extent our findings can be generalised.

References

Allison, Paul D (2019). Asymmetric fixed-effects models for panel data. *Socius 5*, 2378023119826441.

Baltagi, Badi (2008). *Econometric analysis of panel data*, Volume 4. Springer.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics 119*(1), 249–275.

Caliendo, Marco and Sabine Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys 22*(1), 31–72.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2024). Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research.

Card, David and Alan B Krueger (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

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.

Haeck, Catherine and Pierre Lefebvre (2020). Pandemic school closures may increase inequality in test scores. *Canadian Public Policy 46*(S1), S82–S87.

Hsiao, Cheng (2022). *Analysis of panel data.* Number 64. Cambridge university press.

Maldonado, Joana Elisa and Kristof De Witte (2022). The effect of school closures on standardised student test outcomes. *British Educational Research Journal 48*(1), 49–94.

Maynard, Brandy R, Christopher P Salas-Wright, and Michael G Vaughn (2015). High school dropouts in emerging adulthood: Substance use, mental health problems, and crime. *Community mental health journal 51*, 289–299.

OECD (2022, Nov). Pisa 2022 results the state of learning and equity in education.

Paling, Emma (2021, Jun). Who decided ontario schools should stay closed? doctors call on ford to reconsider — cbc news.

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.

Thornberry, Terence P, Melanie Moore, and RL Christenson (1985). The effect of dropping out of high school on subsequent criminal behavior. *Criminology 23*(1), 3–18.

Wen, Leana (2020, Nov). Opinion — most schools should close and stay closed through winter - the washington post.

Appendix

A Tables

Table 2: PISA region codes and the associated province

PISA region code	Province Name
12401	Newfoundland & Labrador
12402	Prince Edward Island
12403	Nova Scotia
12404	New Brunswick
12405	Quebec
12406	Ontario
12407	British Columbia
12408	Alberta
12409	Saskatchewan
12410	Manitoba