School Choice and Class-Size Effects : Unintended Consequences of a Targeted Voucher Program

Olivier de Groote, Ana Gazmuri

WORKING PAPERS

N° 1526

April 2024

"School Choice and Class-Size Effects: Unintended Consequences of a Targeted Voucher Program"

Olivier De Groote and Ana Gazmuri

Toulouse School of Economics

School Choice and Class-Size Effects: Unintended Consequences of a Targeted Voucher Program

Olivier De Groote* Ana Gazmuri†

April 23, 2024

Abstract

We propose a novel method to estimate education production functions on observational data in a context of school choice. We exploit panel data of schools and estimate heterogeneous effects, while allowing for unobserved school, student, and teacher characteristics to be correlated with observed inputs. We then use this model to study the channels behind changes in observed test scores following a voucher reform in Chile. After the reform, many students left public schools, leading to a passive decrease in class size. We show that this can explain part of the policy effects as we find large class size effects for several schools, especially those that saw a decrease after the policy change.

*Toulouse School of Economics, University of Toulouse Capitole, olivier.de-groote@tse-fr.eu.

†Toulouse School of Economics, University of Toulouse Capitole, ana.gazmuri@tse-fr.eu. This paper benefited from comments at TSE, IAAE 2023 and NLS-E 2024. We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010 and grant MATCHINEQ - ANR-22-CE26-0005-01.

1 Introduction

Many countries have implemented voucher programs in education with the idea that competition will pressure public schools to improve quality. Yet, improving educational quality has proven difficult, especially in public schools that often fail to respond to financial incentives. Furthermore, voucher programs have been associated with increased inequality and socioeconomic segregation. Targeted vouchers have been suggested as a possible way to mitigate these issues. In these types of policies, schools enrolling a larger share of students with a low socioeconomic status (SES) receive more financial resources.

An understudied question in the context of targeted educational vouchers (and in voucher policies more generally) is the channels through which school quality is affected by these additional resources. Changing the value of the voucher for certain types of students will necessarily result in changes not just in the resources available, but also in other school characteristics that change endogenously in response to changes in student sorting (for example, class size and peer composition). Therefore, when evaluating quality changes in schools, it is important to consider the policy's indirect effects through these other changes.

In this paper, we propose a novel estimation strategy for an education production function to identify the effects of school characteristics in a school choice environment. Studying the effects of schooling inputs is challenging in a context with a significant level of choice. When parents are free to choose the school they want their children to attend and schools have significant leeway in input choices, most quasi-experimental methods would not be applicable, as schools with different choices of input would also differ in unobserved characteristics both at the student and school level. In our model, we directly take sorting into account by incorporating student and school unobserved heterogeneity in the estimation. We apply the insights of the dynamic panel data literature (Arellano and Bond, 1991; Blundell and Bond, 2000) to allow for multiple inputs to be correlated with both persistent and time-varying unobserved heterogeneity at the student and school level. We then apply a k-means clustering algorithm as in (Bonhomme and Manresa, 2015) to allow for unobserved heterogeneity in treatment effects. Our method is widely applicable in large observational datasets when researchers have access to three or more periods of the same schools, without requiring lagged test scores of students.[1]

We apply our methodology to study the channels behind changes in school quality following the implementation of a targeted voucher reform in Chile in 2008. In particular, we study the effects of class size, teacher experience, and financial resources which all show important changes after the reform. The reform substantially increased resources for schools that serve low socioeconomic status students. Several papers have studied the effects of this reform, showing significant improvement in academic outcomes (Neilson, 2017; Murnane et al., 2017; Correa et al., 2014). Furthermore,

[1] Andrabi et al. (2011) and Ding and Lehrer (2014) use dynamic panel methods to deal with endogeneity concerns in value-added estimators.

Gazmuri (2018) shows important changes in the sorting of students, where private schools respond to the new program by changing admissions policies. After enrolling in the program, they started admitting more lower-income students, aligned with the incentives from the program. We take advantage of the rich individual data of students and panel data on schools, including information about their class sizes, teachers, and financial resources to study the different potential channels through which school quality may be affected.

We first show that most of the improvements in quality surprisingly come from public schools, despite having a decrease in the average socioeconomic level of their students. Additionally, private schools respond strongly to the incentives of the new program, increasing enrollment of low-SES students. This increase in competition for low-SES students is significantly affecting schools' inputs in the production function, in particular decreasing class sizes in public schools and lowering teacher experience.

Studying the effects of class size and teacher characteristics is challenging for two reasons. First, they are endogenous. One reason is that good schools attract more students, potentially leading to larger classes. Another is that class size itself can be valued by students in ways that depend on their unobserved skills. Teachers are also sorting based on schools' unobserved quality. Good teachers might also be in a position to choose schools with their preferred class size. Second, effects might be heterogeneous. Schools could differ in the unobserved ability to manage large classes and use that information to decide on their optimal class sizes.

In our model, we assume math and language test scores in fourth grade are measures of a child's human capital, up to some additive measurement error. Human capital is the outcome of an education production function which is allowed to differ by type of school (public and private) and by subject (math and language). The inputs of the production function include student skills, school productivity, and the human and financial resources available to the child at school. Financial resources are observed for each school in the data. Human resources are assumed to be determined by class size (quantity), teacher experience (observed quality), and unobserved teacher quality. Student skills are captured by a set of observable and unobservable characteristics at the school-year and individual level. School productivity remains unobserved.

To estimate the model, we proceed in two steps. We use data from a panel of all schools over 14 years, where we observe all students enrolled in 4th grade. In the first stage, we obtain a vector of school-year-specific quality, net of student individual effects. We do this by regressing individual test scores on individual observed characteristics and school-year fixed-effects. In this stage, we also estimate heterogeneity in class size effects through observable student characteristics by estimating the interaction between class size and student characteristics. In a second step, we decompose the school-year fixed effects into class-size effects, teacher quality, and financial resources, plus an error term. In line with the model, this error term contains unobserved school productivity, as well as unobserved student and teacher quality. It is therefore likely correlated with observed inputs. We use a first-differenced GMM estimator to estimate a production function as in Blundell and

Bond (2000) by using dynamic panel data techniques (Arellano and Bond, 1991). This allows for persistent school fixed effects, as well as a time-varying AR(1) shock. We take advantage of our long panel to define past observable inputs and fixed effects as instruments. We carefully choose long enough lags to make sure these are valid instruments, according with the timing assumptions in the model. Given that observable characteristics are moderately persistent in the data, their past levels are correlated with current period differences, providing strength to the instruments. By using sufficiently deep lags, we can exclude that they are correlated with the current period unobserved quality of schools or inputs, even if teachers and students sort on observed inputs through a first-order Markov process. Sensitivity checks with deeper lags show that we could also allow for higher-order processes. We then extend the model to allow for unobserved heterogeneity in class-size effects. Following Bonhomme and Manresa (2015) we assign schools to a finite number of types using a k-means clustering algorithm and estimate different class size effects for each.

In a basic model with heterogeneous effects through observable student characteristics only, we find that an additional 10 students in class decrease math test scores by 4.8% of a standard deviation for public schools and 15.4% for private schools for the baseline group of male students with low parental education and with average income. Reading scores are decreased by respectively 9.8% and 14.6%. A standard fixed effects estimator finds the same sign of effects, but effect sizes are three to nine times smaller than what we find, suggesting an important role for time-varying unobserved heterogeneity. For both scores, we find stronger effects for female students, but only in public schools (2 % points). In all schools, children of highly educated parents and students from low income households experience more negative effects.

In our more flexible model, we allow for eight unobserved school types, four for public and four for private, that are allowed to have different class-size effects. We find significant heterogeneity, with some types in public schools having negative effects of up to 14% and in private schools of up to 30%, while others experience almost no effect. Combining the student and school heterogeneity of our most flexible specification, we find average class size effects in math to be -5.1% in public schools and -12.0% in private schools. For reading, we find respectively -5.8% and -4.9%. In a decomposition exercise, we find that this can explain an important part of the catch-up of test scores in public schools, particularly in those with low test scores that do not handle large classes well.

We also obtain results for teacher experience and financial resources. For teacher experience, we find positive effects but most are small and insignificant. The effect of funding per student is positive everywhere, but not precisely estimated for public schools. We do find significant effects in private schools: a 10% increase in funding per student raises scores by 1.5% of a standard deviation in math and 1.4% in reading. For these reasons, we focus most of our discussion on the results for class size as these are the coefficients we can estimate more precisely and get robust results.

The first contribution of this paper is to the literature on education and firm production functions. The education production function literature studies the relationship between schooling

inputs and academic achievement in non-experimental settings. Most of the methods normally used to estimate input effects, implicitly impose strong assumptions on the production function technology, like for example, independence of inputs with unobserved school productivity or constant effects of inputs (Todd and Wolpin, 2003). In our model, we explicitly account for unobserved heterogeneity at the student and school level in determining education inputs. For this, we borrow from the literature on firm production function estimation as it generally deals with multiple endogenous inputs in rich panel data at the firm level (see De Loecker and Syverson (2021) for a recent review). Given the richness in our data, we can allow for more flexibility to handle our setting. First, we difference out a time-invariant school fixed effect, before dealing with the remaining unobserved heterogeneity. Sensitivity analysis shows that allowing for fixed effects is crucial to detect class size effects. While this has been already suggested in the dynamic panel approach of Blundell and Bond (2000), it is uncommon to do this for firms because of the lack of within-variation in inputs. We have this variation because of the recent change in the voucher policy. Second, unobservables are usually interpreted as productivity. We extend this to also capture unobserved quality of students and teachers. This is another reason to use the dynamic panel approach, as opposed to a control function approach in Ackerberg et al. (2015) as this does not require an input that is monotonic in this unobservable. We show that this yields the same GMM estimator, but requires Markovian assumptions on the information that students and teachers use when deciding their school. This impacts the choice of lag length when choosing the set of instruments and we show sensitivity of our results to different lag choices. Third, we allow for heterogeneous effects through unobservables by using a k-means clustering algorithm (Bonhomme and Manresa, 2015) to group schools and interact it with the class size input.

The second contribution of this paper is to show the indirect effects of targeted vouchers on school quality through endogenous student sorting. While the existing literature shows significant improvement in average school quality following the voucher reform in Chile, it is unclear which mechanisms explain the improvement and whether public and private schools were affected in a similar way (Neilson, 2017; Murnane et al., 2017; Correa et al., 2014). Some studies attribute the improvement in public schools to competition effects (Navarro-Palau, 2017). This is hard to reconcile with the fact that public schools have been facing high competition for decades without much effect on quality. We show that indirect effects associated with the changes introduced by the SEP program can explain part of these effects without implying an active effort from public schools. As a result of the shift of students from public schools to private schools, public schools have decreased average class size, something that is not observed in private schools.

Finally, this paper contributes to the estimation of class-size effects. Class size is often at the heart of the debate on school quality and the allocation of school resources. Much of the research on the relationship between class size and achievement is for developed countries using experimental or quasi-experimental methods. A famous example using quasi-experimental methods is Angrist and Lavy (1999), who exploits a ceiling rule on class size in Israel to estimate class size effects using

a regression discontinuity (RD) approach. Generally, the evidence points to small average effects. In school choice settings, it is challenging to recover class-size effects using such quasi-experimental methods. Although Chile has a cap on class size at 45 students, the assumptions for an RD design would typically be violated in a school choice setting where parents have a substantial degree of school choice and schools are free to allocate their resources and decide on their preferred class size (Urquiola and Verhoogen, 2009). Alternatively, randomized control trials can be used. Results from Tennessee and Ontario have also shown moderate benefits from class-size reduction (Word et al., 1990; Krueger and Whitmore, 2001; Finn and Achilles, 1990). More recently, Adusumilli et al. (2024) re-evaluate the results in Tennessee by accounting for heterogeneous compliance responses and show that negative average effects are driven by a minority of schools experiencing very large effects. Similarly, we show that class size effects can be large in a subset of schools. This heterogeneity shows that (quasi-)experimental variation coming from specific groups is difficult to extrapolate to the entire population of students. We contribute by developing a general method that can be used with rich administrative data to study the entire distribution of class size effects. For Chile, we show that they are large on average, and heterogeneous both within public and private schools.

The remainder of the paper is organized as follows. Section 2 describes the institutional context, Section 3 presents the data and shows some descriptive statistics, Section 4 describes our model and estimation. Section 5 presents the estimation results, and Section 6 concludes.

2 Institutional context

Since 1981, Chile has a nationwide school voucher program under which students may choose among both public and private schools. Private voucher schools receive the same per-student voucher from the government as public schools. The main differences between the two types of schools are in the way they are managed and the fact that private voucher schools were allowed to charge a small top-up tuition. Public schools depend on local municipalities, while private voucher schools are managed privately by different types of private organizations (for-profit, non-profit, religious, secular, etc). There is a similar share of students in private subsidized schools and public schools (around 45-50%). Additionally, there is a small number of tuition-charging private schools that operate without public funding and account for less than 8% of enrollment. We do not observe resources at these schools and therefore are excluded from the analysis. Generally, schools are free to choose the number of students they enroll subject to a 45 student class-size cap. Most schools have class sizes significantly below this cap. In this paper, we focus on elementary schools and use 4th grade outcomes. In 4th grade, all students in the country take a standardized test (SIMCE test) which allows us to use an outcome that is comparable across schools and regions.

School applications and the admission process are decentralized. Private voucher schools have discretion on how many and which students they admit. There is extensive literature on the

Chilean voucher program that shows significant sorting, unequal academic results, and evidence of selectivity from private subsidized schools based on student socioeconomic characteristics(Hsieh and Urquiola, 2006; Contreras et al., 2010; Gazmuri, 2018)[2].

In response to these critics regarding the old voucher system, in February 2008, Chile adopted a new policy creating a targeted schooling subsidy for the most vulnerable students (SEP law, for 'Subvencion Escolar Preferencial'). It modified the existing flat subsidy per student by introducing a two-tier voucher, with a higher subsidy for low-SES students with the objective of 'promoting equal opportunity and improving the quality of education' (Weinstein et al., 2010; Mineduc, 2015)[3]. Participating schools (both public and private subsidized) received approximately 40% higher vouchers for students defined as priority by the SEP program compared to the baseline voucher. Schools have to opt in to participate in the SEP program. Only participating schools receive the additional benefits, otherwise they keep receiving just the baseline voucher for all students, even if priority students are enrolled. In terms of participation, virtually all public schools and slightly over 65% of private subsidized schools registered in the SEP program. The main reasons not to participate are related to additional requirements. Participating schools are required to design and implement an educational improvement plan and these schools were required to accept the value of the voucher as full payment of tuition for preferential students, eliminating extra tuition and other fees for eligible students.

Student eligibility is determined annually by the Ministry of Education according to several criteria associated with socioeconomic vulnerability. By 2012, 44% of elementary school students were classified as eligible for the SEP benefits.

In the analysis, we include all public and private voucher schools, both SEP and non-SEP schools. On the student side, we include all students enrolled in any of these schools, both eligible or non-eligible for the additional SEP benefits.

3 Data and Descriptives

Datasets The empirical analysis in this paper relies on four main datasets from the Chilean Ministry of Education. The first one is a comprehensive dataset of yearly school and student-level data from 2005 to 2018. It contains the universe of students and the schools where they are enrolled, along with school characteristics. It reports the type of school, whether the school is registered in the SEP program, and the exact location of each school. We exclude rural schools from the analysis because these schools tend to group cohorts together in one class and have very low enrollment.

The second one contains SIMCE test results of all 4th grade students from 2005 to 2018. The

[2]In 2016, a new reform started to centralize admissions and eliminated schools top-up tuition, but this is outside of our time frame. The data in this paper covers the period from 2005-2015.

[3]For a more detailed discussion of the 2008 reform, see Gazmuri (2018).

Table 1: Summary statistics

Variable	Mean	SD	Min	Max
Mother educ mid	0.539	0.498	0.000	1.000
Mother educ high	0.267	0.442	0.000	1.000
Log(income)	6.025	0.823	4.173	8.183
Female	0.504	0.500	0.000	1.000
Class size / 10	3.456	0.189	1.382	3.682
Teacher experience / 10	1.603	0.360	1.060	2.400
Log(money/student)	13.410	0.452	11.558	14.281
Math score	-0.050	0.982	-3.462	2.491
Reading score	-0.038	0.987	-3.163	2.244

Summary statistics of 1,885,344 observations in years 2005-2018.

SIMCE is a standardized test taken by all 4th graders in the country, so we use it to measure student achievement.

The third dataset has all teachers' contracts in public and private subsidized schools. This data includes details about the teachers in each school, including hours in each contract, specific tasks (teaching, administration, etc), and teacher characteristics like gender, age and experience. We use this to construct pupil-teacher ratios and teacher experience.

Additionally, we use student demographic characteristics like family income and education of the mother. This information is included in a questionnaire sent to the families of students taking the SIMCE test.

Summary statistics Table 1 summarizes the data. Observed student characteristics comprise math and language test scores, gender, mother's educational background, and household income. We use these last two to capture the socio-economic environment of children. We observe the years of education of the mother and define 3 categories, corresponding to finishing elementary school, high school and continuing education afterward. We normalize the original test scores to have mean 0 and standard deviation equal to 1. Observed school characteristics are class size, years of teacher experience, and the funding in money per student, which are all accounted for in fourth grade to match the year in which we observe test scores. We drop observations with missing data, but we see that the mean in test scores changes only marginally.

Evolution over time Figure 1 summarizes the characteristics by type of school and their evolution over time. After the 2008 policy change, we see a large drop in both class size and teacher experience in public schools, while it remained relatively constant in private schools. For class size, both types of schools started with classes of around 36 students, but this dropped to less than 32 in public schools. Teacher experience was instead very diverse at the start (23 years in public schools and 14 in private) but in public schools it dropped to almost the same level as private (15 years). Funding per student went up a lot in both. In 2018, the level of funding was four times the

amount schools received in 2005 (consumer prices were only 50% higher). It increased a bit more in public schools, consistent with the focus of the policy on low SES students, who are more often in public schools. Finally, test scores improved in all schools but to a different extent. Math scores saw a steep increase in public schools until around 2012, after a temporary drop it stabilized at almost the maximum level. Private schools were already at a much higher level, and experienced smaller changes. Reading scores show similar patterns, but with two noticeable differences. First, the evolution in private and public schools before 2012 is more similar with private schools also experiencing steep gains in the beginning. Second, scores have not yet stabilized, particularly in public schools which are experiencing a steeper increase in recent years.

Figure 2 shows average math scores for students of different levels of mother's education. We plot average scores for 2007, 2011, 2015, and 2018. The figure shows that the largest gains are concentrated among students with mothers with less than high-school education. This is consistent with the trends in school characteristics as low-SES students are much more likely to be enrolled in public schools. Therefore, these students are likely benefiting from the class size reductions and increase in financial resources.

These figures suggest that the decrease in class size might be partly responsible for the large gains in public schools. However, these schools also experience a simultaneous drop in teacher experience and an increase in their funding. Furthermore, we cannot rule out other changes over time such as teacher quality and, in this context of school choice, student quality. In the next section we will provide an estimation approach to uncover the causal effects of these characteristics and then use this to analyze how much class size changes contributed to the change in test scores.

4 Model and estimation

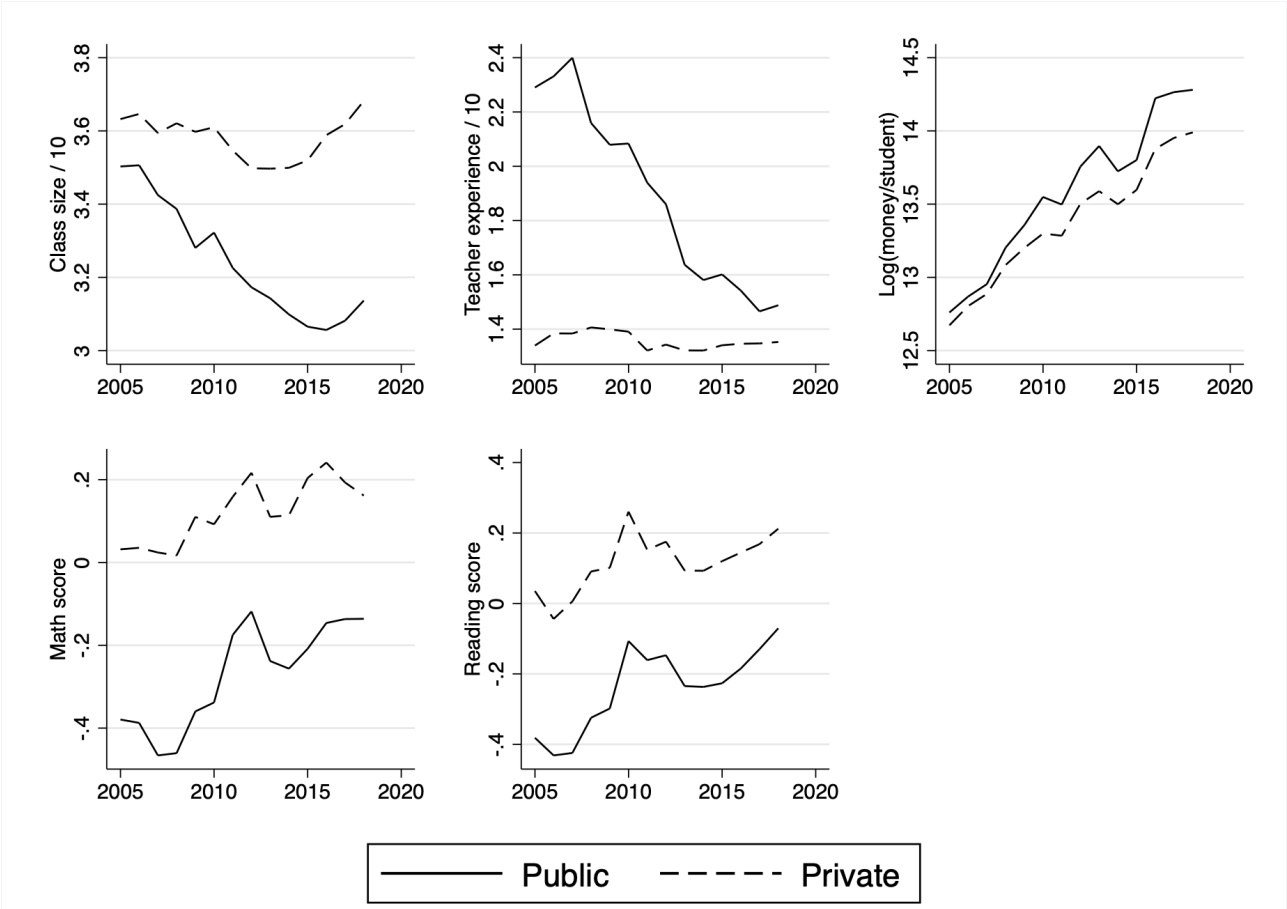
This section proposes a general estimation method for educational production functions in the context of school choice. We first propose a production function in which treatment effects are heterogeneous, only through observable student characteristics. Subsequently, we discuss estimation of this model. Finally, we extend the model to allow for unobserved treatment effect heterogeneity at the school level.

4.1 Education production function

Let the human capital H_i of student i after attending school j in period t be given by the following Cobb-Douglas production function:

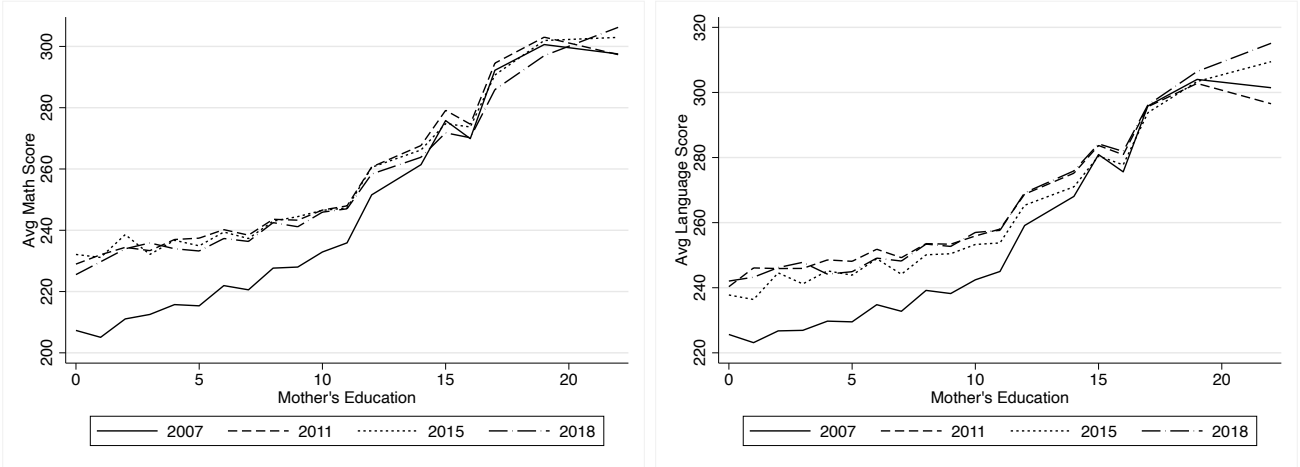
$$H_i = \Omega_{jt} S_i^{\alpha^s} L_{jt}^{\alpha^l} F_{jt}^{\alpha^f}$$

Figure 1: Evolution of characteristics and scores by type of school



Note: Averages weighted by enrollment numbers of students. Math and reading score normalized to be mean zero and standard deviation 1 in the sample.

Figure 2: Evolution of test scores by mother's education



Note: Average math and reading scores for students of different mother's education. Mother's education is expressed in years (12 years is equivalent to finishing high-school).

where Ω_{jt} represents school productivity, S_i initial student skills, L_{jt} human resources available to the school and F_{jt} financial resources available to the school. Note that we do not distinguish

between class and school and assume each school only contains one class. This is the case for 68% of our sample, in other cases we use school-level averages.[4]

We take logs to obtain a linear equation (represented by lower-case letters):

$$h_i = \omega_{jt} + \alpha^s s_i + \alpha^l l_{jt} + \alpha^f f_{jt}$$

In this equation, we only observe f_{jt}. We further decompose the initial student skills and human resources in observed and unobserved characteristics:

$$\begin{aligned} s_i &= X_i'\beta^s + \epsilon_i \\ l_{jt} &= Z_{jt}'\beta^l + \beta^c C_{jt} + \eta_{jt} \end{aligned}$$

where X_i are observed student characteristics, ϵ_i unobserved student skills, Z_{jt} observed teacher characteristics, C_{jt} class size and η_{jt} unobserved teacher quality. By letting class size enter through the human resources available to the school, we explicitly account for the fact that larger classes reduce the proportion of human resources available to the student.

We also assume that we observe h_i through a test score g_i, up to an additive measurement error m_i. This results in the following equation to bring to the data:

$$g_i = X_i'\alpha^s\beta^s + Z_{jt}'\alpha^l\beta^l + \alpha^l\beta^c C_{jt} + \alpha^f f_{jt} + \omega_{jt} + \alpha^l\eta_{jt} + \alpha^s\epsilon_i + m_i$$

4.2 Estimation and identification

School-year fixed effects and individual heterogeneity In a first step, we estimate the relative impact of individual characteristics on human capital $a^s \equiv \alpha^s\beta^s$ using a fixed effects approach. To see this, we first define the individual deviation of the unobserved skills of student i from the class mean $\overline{\epsilon}_{jt}$:

$$\tilde{\epsilon_i} = \epsilon_i - \overline{\epsilon}_{jt}$$

We can consistently estimate a_s by regressing g_i on X_i and (j,t) fixed effects if we assume X_i is uncorrelated with $\tilde{\epsilon_i} + m_i$. Moreover, school-year fixed effects δ_{jt} can be decomposed in the following way:

$$\delta_{jt} = Z_{jt}'\alpha^l\beta^l + \alpha^l\beta^c c_{jt} + \alpha^f f_{jt} + \omega_{jt} + \alpha^l\eta_{jt} + \alpha^s\overline{\epsilon}_{jt} + \overline{m}_{jt}$$

with $\overline{m}_{jt}$ the mean of the measurement error at the school-year level.

Regression of fixed effects on school and teacher characteristics We rewrite the equation for δ_{jt} as a linear regression with parameters $a^l \equiv \alpha^l\beta^l$, $a^c \equiv \alpha^l\beta^c$, $a^f \equiv \alpha^f$ and error term

[4]For the schools that do have multiple classrooms per cohort, differences in class size and other inputs within school-year are minimal.

$\xi_{jt} \equiv \omega_{jt} + \alpha^l \eta_{jt} + \alpha^s \overline{\epsilon}_{jt} + \overline{m}_{jt}$:

$$\delta_{jt} = Z'_{jt} a^l + a^c C_{jt} + a^f f_{jt} + \xi_{jt}. \tag{1}$$

We cannot estimate these parameters using OLS as we expect school and teacher characteristics to be correlated to each element in the error term. For example, more productive schools (ω_{jt}) are likely to hire teachers with better observed characteristics (Z_{jt}). Teachers of better unobserved quality (η_{jt}) could have more school options to choose from and might be attracted by smaller class sizes (C_{jt}). Similarly, parents of higher-skilled students ($\overline{\epsilon}_{jt}$) might devote more attention to favorable observed school characteristics when choosing a school.

Error structure To estimate this model, we proceed in a way that is similar to Blundell and Bond (2000). First, we decompose the error term to allow for school fixed effects, time fixed effects and an AR(1) structure on the residual:

$$\begin{aligned} &\xi_{jt} = \xi_j + \xi_t + \tilde{\xi}_{jt} \\ &\text{with } \tilde{\xi_{jt}} = \rho \tilde{\xi}_{jt-1} + e_{jt} \end{aligned} \tag{2}$$

This error structure imposes restrictions, but these are natural in our context. ξ_j captures unobserved school productivity, teacher quality and average student skills that are present in every period for a given school. Similarly, common shocks across schools are captured by ξ_t. Importantly, time-varying shocks at the school level are allowed to be persistent. ρ measures the persistence of a school productivity shock. The persistence of a shock in teacher quality (average student skills) is given by $\rho\alpha^l$ ($\rho\alpha^s$). This implies they are weighted by their impact on human capital formation, relative to the impact of school productivity. If teachers (students) have a large impact in the education production function, the impact of a shock in t will not only be more important in t but also in future periods.

GMM Estimator To build an estimator, note first that

$$\tilde{\xi}_{jt-1} = \delta_{jt-1} - Z'_{jt-1} a^l - a^c C_{jt-1} - a^f f_{jt-1} - \xi_j - \xi_t \tag{3}$$

Substituting (3) in (2) in (1):

$$\delta_{jt} = \rho\delta_{jt-1} + (Z'_{jt} - \rho Z'_{jt-1})a^l + a^c(C_{jt} - \rho C_{jt-1}) + a^f(f_{jt} - \rho f_{jt-1}) + (1-\rho)\xi_j + \xi_t - \rho\xi_{t-1} + e_{jt}$$

Take first differences to remove the school fixed effect:

$$\triangle\delta_{jt} = \rho\triangle\delta_{jt-1} + (\triangle Z'_{jt} - \rho\triangle Z'_{jt-1})a^l + a^c(\triangle C_{jt} - \rho\triangle C_{jt-1}) + a^f(\triangle f_{jt} - \rho\triangle f_{jt-1}) + \triangle\xi_t - \rho\triangle\xi_{t-1} + \triangle e_{jt}$$

We can now estimate ρ, a^l, a^f, a^c and $\tilde{\xi}_t \equiv \triangle\xi_t - \rho \triangle \xi_{t-1}$ using GMM by using moments conditions that interact the quality shocks $\triangle e_{jt}$ with a set of instruments. First, we include year dummy variables to capture $\tilde{\xi}_t$. For the lagged dependent variable and the school inputs (financial resources, class size and teacher experience), we proceed by imposing the following moment conditions:

$$E[x_{jt-s} \triangle e_{jt}] = 0 \text{ for } x_{jt} = (Z_{jt}, C_{jt}, \delta_{jt}) \text{ and } s \geq 3.$$

Choice of lag length The lag length is chosen to be at least 3, which given our assumptions, should be sufficiently long to satisfy the moment condition. To see this, note that a sufficient condition is $E[x_{jt-s}e_{jt}] = E[x_{jt-s}e_{jt-1}] = 0$, with $E[x_{jt-s}e_{jt-1}] = 0$ being the hardest to satisfy. Clearly $s = 1$ does not work because δ_{jt-1} depends on its shock e_{jt-1}, making $E[x_{jt-1}e_{jt-1}] \neq 0$ for $x_{jt} = \delta_{jt}$. The same problem arises for characteristics (Z_{jt}, C_{jt}) that are correlated to contemporaneous shocks. This arises for example when good teachers or strong students sort on observed characteristics. $s = 2$ only works under strong assumptions on sorting behavior. If students and teachers in $t-1$ sort not (only) on x_{jt-1}, but on x_{jt-2}, it implies $E[x_{jt-2}e_{jt-1}] \neq 0$. Instead, we assume a first-order Markov assumption for all actors in the model: they can act on current and last-period information, but not earlier. This then allows us to use $s \geq 3$. We conduct some sensitivity analysis regarding the choice of lag length in section 5.2 showing that our results are robust to this assumption and we can even be more flexible restricting the instruments to longer lags.

Identification As we express in equation 4.2, we achieve identification in this model by exploiting the large panel dimension at the school level. Standard fixed effects assumptions would account for time-invariant heterogeneity, addressing some of the endogeneity concerns as productive schools might attract more students, which could lead to an increase in class size, or more experienced teachers. The added complexity in the model comes from the time-varying component of unobserved school, teacher, and student quality which is allowed to be correlated with observed inputs.

Two assumptions are crucial to achieving identification. First, we assume an AR(1) process on the evolution of the unobserved component of the production function (see 2), which allows temporary shocks to be persistent in a specific, yet flexible way. For example, if scores increase because a new hire causes a shock in the unobserved teacher quality, that shock can persist in the future. A higher estimate of ρ denotes higher persistency of the shocks, while the relative persistency of each component of the shock (unobserved student and teacher characteristics) is weighted by their respective impact on the production function.

A second assumption relates to the timing of sorting by agents. We allow school administrators, teachers and students to observe the current and previous period inputs of the school when

searching for a match. For example, if good teachers or students (in unobserved ways) were attracted by a school with a small class size the previous year, it would cause a correlation between e_{jt} and C_{jt-1}. It is also likely that current inputs are correlated with e_{jt}. A school could allow for more children per classroom, if it knows it can hire a teacher with great unobserved skills this year. Our identifying power comes from the restriction that good teachers or students should not respond directly to observed inputs from two periods ago or earlier. This allows us to exploit sufficiently deep lags as instruments to tell us about the causal effect of these inputs on human capital. As long as there is sufficient autocorrelation in observed inputs, the instruments should satisfy the relevancy requirement for a good instrument. An increase in an input at time t would cause changes in all future periods to be negatively correlated with past levels. This is reasonable for many inputs as schools could be deviating from their preferred class size or teacher experience, but physical and budgetary constraints likely force them to make such shocks not last too long. This assumption is testable by running first-stage regressions and is confirmed in the data. [5]

With sufficiently rich data, we can provide a more robust approach by using deeper lags only, which comes at the cost of efficiency. The optimal lag length should be chosen such that: (1) it is sufficiently recent to ensure that a past change in inputs is still affecting the adjustments to that same input today, and (2) it should be sufficiently far in the past such that a past change in one input is no longer directly affecting a change in unobserved quality of schools and inputs today. In practice, we show our results are robust for using $s \geq 7$ instead of $s \geq 3$ in section 5.2.

4.3 Heterogeneous class size effects

We extend the model to allow for unobserved heterogeneity at the school level in the class size coefficient. Let the log production function be given by:

$$
\begin{aligned}
&h_i = \omega_{jt} + \alpha^s s_i + \alpha^l l_{ijt} + \alpha^f f_{jt} \\
&\text{with } l_{ijt} = Z'_{jt}\beta^l + \beta^c_{ij} C_{jt} + \eta_{jt} \\
&\text{and } \beta^c_{ij} = X'_i \beta^c + \kappa_j
\end{aligned}
$$

with κ_j capturing unobserved heterogeneity in class size effects by school. We allow κ_j to differ by schools' unobserved types and identify these types in a first step using the k-means clustering algorithm as in Bonhomme and Manresa (2015). This procedure classifies schools into types that have similar unobserved characteristics that drive important variation in the data.[6] This allows for grouped fixed-effects, which are allowed to be arbitrarily correlated to other characteristics of

[5] In Appendix Table 9 we show the strong correlation of our differenced outcomes and inputs on their third lag. Furthermore, we show how deeper lags often also contain useful information. Note that in such regressions, we lose many observations when adding deeper lags. As in Arellano and Bond (1991), we can avoid this issue in a GMM estimator by using time-specific instruments.

[6] These types should be interpreted in a similar way as ξ_j: they can capture unobserved heterogeneity of the school, but also of the average unobserved quality of their students or teachers.

the school.

The k-means clustering algorithm classifies schools by minimizing the within-group distance from the group average of the specific moments. We do this separately for public and private schools and use four variables: the estimated fixed effects of math scores and reading scores, as well as teacher experience and class size.[7]

5 Results

We first discuss the estimation results and show the distribution of class size effects. We then show how they depend on different modeling choices. Finally, we use the estimates to decompose the changes in test scores after the policy change.[8]

5.1 Estimation results

First step In a first step, we use the individual data of students between 2005-2018 to estimate school-year fixed effects, as well as the individual heterogeneity in class size effects using an OLS regression. Results can be found in Table 2.

We find strong effects of initial characteristics on test scores. For example, children who have a parent who graduated secondary education have math scores that are 32% of a standard deviation (SD) higher in public schools and 22% in private schools, compared to children of parents without primary or secondary degrees. We also see lower test scores for female students. Results for reading scores are almost identical, except for gender, with girls now obtaining higher scores.

In this stage, we also capture the observed, individual heterogeneity in class size effects. The heterogeneity is precisely estimated, but rather small. When class size increases by 10, children of highly educated parents experience an additional decrease (or reduced increase) in test scores of about 2% of a SD in both types of schools, compared to children of parents with a low educational level. We find the same result for female students, but only in public schools. Higher parental income instead reduces any negative effect of class size effects by 0.05 to 0.09% of a SD for a 10% increase. Again, the results for reading scores are almost identical.

Second step In the next step, we regress the estimated school-year fixed effects on teacher and school characteristics. The results for public schools can be found in Table 3 (math) and Table 5

[7]To apply the clustering algorithm, we need to define first the number of groups. Before applying the algorithm, we take the population-weighted average over time of these variables and normalize each to be mean zero and standard deviation 1. We then use 1000 starting values and run the algorithm using different number of groups, selecting the solution with the smallest within-group distances. We use a specification of four types as more types did not allow us to see the same solution reappearing with different starting values, putting into question the number of starting values needed to obtain the global optimum.

[8]We show standard errors robust for clustering at the school-level, however we do not currently correct for potential estimation noise in the categorization of types or in the school-year fixed effects recovered from the first step.

Table 2: First stage results

		Math				Reading			
		Public		Private		Public		Private	
Variable		est	se	est	se	est	se	est	se
Mother educ mid		0.177	(0.003)	0.130	(0.003)	0.164	(0.003)	0.125	(0.003)
	x class size	-0.004	(0.003)	-0.007	(0.004)	-0.004	(0.003)	-0.005	(0.004)
Mother educ high		0.323	(0.004)	0.222	(0.003)	0.323	(0.004)	0.231	(0.003)
	x class size	-0.016	(0.005)	-0.017	(0.004)	-0.021	(0.005)	-0.015	(0.004)
Log(income)		0.107	(0.002)	0.099	(0.001)	0.099	(0.002)	0.090	(0.001)
	x class size	0.005	(0.002)	0.009	(0.002)	0.006	(0.002)	0.011	(0.002)
Female		-0.083	(0.002)	-0.097	(0.002)	0.178	(0.002)	0.157	(0.002)
	x class size	-0.021	(0.003)	-0.003	(0.002)	-0.020	(0.003)	-0.006	(0.002)
Year × school FE		Yes		Yes		Yes		Yes	
R-squared		0.226		0.252		0.189		0.189	
Observations		776,697		1,106,724		776,697		1,106,724	

First stage results. Class size divided by 10 and subtracted by its mean. Log of income subtracted by its mean.

(reading). For private schools, they are in Table 4 (math) and 6 (reading). In specification (1), we show a naive OLS regression with year fixed effects. We add school fixed effects in (2) to account for unobserved time-invariant heterogeneity at the school-level, and we estimate a GMM model with time-varying heterogeneity in (3). Finally, (4) allows for heterogeneous class size effects by schools' unobserved types.

We first discuss the math results for public schools and then highlight the main differences in the three other tables. Moving from OLS to FE, the sign of class size changes. In the FE regression, we find that an increase of 10 students, leads to a decrease in test scores of 1.7% of a SD. Also the effects of teacher experience and financial resources are reduced. Adding autocorrelated time-varying unobserved heterogeneity in the GMM estimators turns out to be important. To see this, note first that the autocorrelation parameter is significantly different from 0, showing that fixed effects are not sufficient to capture persistent unobserved heterogeneity. Second, the result for class size changes substantially when allowing for correlation with the residual: 4.8% of a SD. It also increases the impact of teacher experience to be significant and larger than in other specifications. A 10-year increase in average experience leads to a 6.3% of a SD increase in scores (or a 5.2% if we consider the final specification with added flexibility on class size effects). We find the opposite effect for financial resources as their effect is reduced and no longer significant (but also more imprecisely estimated). Finally, specification (4) shows that class size effects are driven entirely by a single type of school: type 3. Type 4 schools also have a negative effect, but not statistically significant. Note that we order types from a high average math fixed effect (1) to a low one (4). It looks like the best-performing schools do not experience any class size effects. Only the third category turns out to be statistically significant with a strong effect of 14% of a SD. This suggests

Table 3: Second stage results math: public

	OLS	FE	GMM	GMM +het
	(1)	(2)	(3)	(4)
Class size	0.064	-0.017	-0.048	
	(0.008)	(0.006)	(0.033)	
Type 1				0.019
[33.7%]				(0.043)
Type 2				0.063
[23.8%]				(0.047)
Type 3				-0.142
[36.9%]				(0.033)
Type 4				-0.078
[5.6%]				(0.051)
Teacher exp	0.029	0.013	0.063	0.052
	(0.015)	(0.011)	(0.041)	(0.030)
Log(money/student)	0.193	0.124	0.063	0.166
	(0.031)	(0.024)	(0.093)	(0.075)
ρ			0.540	0.408
			(0.033)	(0.026)
Year FE	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
Schools	1,589	1,589	1,576	1,576
Years	14	14	12	12
Observations	20,764	20,764	16,924	16,924

Second stage results. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses. Estimated share of students in each school type in brackets. Clusters ordered from highest school fixed effect in math scores (1) to lowest (4).

Table 4: Second stage results math: private

	OLS	FE	GMM	GMM +het
	(1)	(2)	(3)	(4)
Class size	0.161	-0.026	-0.154	
	(0.007)	(0.005)	(0.032)	
Type 1				-0.108
[50.4%]				(0.047)
Type 2				-0.097
[12.5%]				(0.040)
Type 3				-0.109
[33.7%]				(0.037)
Type 4				-0.230
[3.4%]				(0.049)
Teacher exp	0.062	0.020	0.065	0.066
	(0.015)	(0.013)	(0.058)	(0.042)
Log(money/student)	-0.067	0.207	0.113	0.149
	(0.025)	(0.022)	(0.043)	(0.037)
ρ			0.642	0.540
			(0.031)	(0.024)
Year FE	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
Schools	1,952	1,952	1,927	1,927
Years	14	14	12	12
Observations	25,185	25,185	20,360	20,360

Second stage results. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses. Estimated share of students in each school type in brackets. Clusters ordered from highest school fixed effect in math scores (1) to lowest (4).

Table 5: Second stage results reading: public

	OLS	FE	GMM	GMM +het
	(1)	(2)	(3)	(4)
Class size	0.051	-0.011	-0.098	
	(0.007)	(0.005)	(0.029)	
Type 1				0.041
[33.7%]				(0.036)
Type 2				0.001
[23.8%]				(0.042)
Type 3				-0.129
[36.9%]				(0.030)
Type 4				-0.134
[5.6%]				(0.044)
Teacher exp	0.031	0.007	0.054	0.048
	(0.013)	(0.010)	(0.028)	(0.022)
Log(money/student)	0.242	0.093	0.094	0.115
	(0.027)	(0.022)	(0.078)	(0.064)
ρ			0.332	0.198
			(0.046)	(0.033)
Year FE	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
Schools	1,589	1,589	1,576	1,576
Years	14	14	12	12
Observations	20,764	20,764	16,924	16,924

Second stage results. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses. Estimated share of students in each school type in brackets. Clusters ordered from highest school fixed effect in math scores (1) to lowest (4).

Table 6: Second stage results reading: private

	OLS	FE	GMM	GMM +het
	(1)	(2)	(3)	(4)
Class size	0.114	-0.023	-0.146	
	(0.006)	(0.004)	(0.029)	
Type 1				-0.004
[50.4%]				(0.037)
Type 2				0.010
[12.5%]				(0.035)
Type 3				-0.088
[33.7%]				(0.033)
Type 4				-0.297
[3.4%]				(0.051)
Teacher exp	0.076	0.036	0.042	0.014
	(0.012)	(0.011)	(0.039)	(0.032)
Log(money/student)	-0.056	0.124	0.108	0.135
	(0.021)	(0.019)	(0.039)	(0.034)
ρ			0.402	0.331
			(0.041)	(0.031)
Year FE	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
Schools	1,952	1,952	1,927	1,927
Years	14	14	12	12
Observations	25,185	25,185	20,360	20,360

Second stage results. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses. Estimated share of students in each school type in brackets. Clusters ordered from highest school fixed effect in math scores (1) to lowest (4).

that class size effects are particularly a concern in the most underperforming schools.

The biases in OLS and FE are similar for the four other outcomes. Reading scores in public schools show very similar patterns in the effects of all inputs. Class size is still mainly affected through types 3 and 4. Math scores in private schools are more strongly affected by class size: -15.4% with homogeneous types, and more evenly distributed over the different types in the final specification. Large negative effects of class size on reading scores in private schools, however, are also concentrated in the underperforming types. Regarding other inputs, we find little differences between public and private schools. Money per student generally has a larger impact for private schools. A 10% increase in funding raises math scores by 1.5% of a SD and reading scores by 1.4 %. Teacher experience has small and insignificant effects in all cases.

Distribution of treatment effects To obtain a better overview of the heterogeneity in class size effects (both through school types and student heterogeneity), we calculate average effects. Note that these are distributions by students, not schools, to weigh schools of different sizes correctly. For math, Figure 3 shows average treatment effects in math of -5.1 % in public schools and -12 % in private schools. Consistent with the estimation results, we see that a large group of students in public schools experience (close to) 0 effects, while in private schools almost every student has an effect size that is more negative than -10%, and some around -25%. For reading, Figure 4 shows average treatment effects of -5.8% in public and -4.9% in private schools and effects are more spread out over the distribution.

Figure 3: Distribution of treatment effects in math

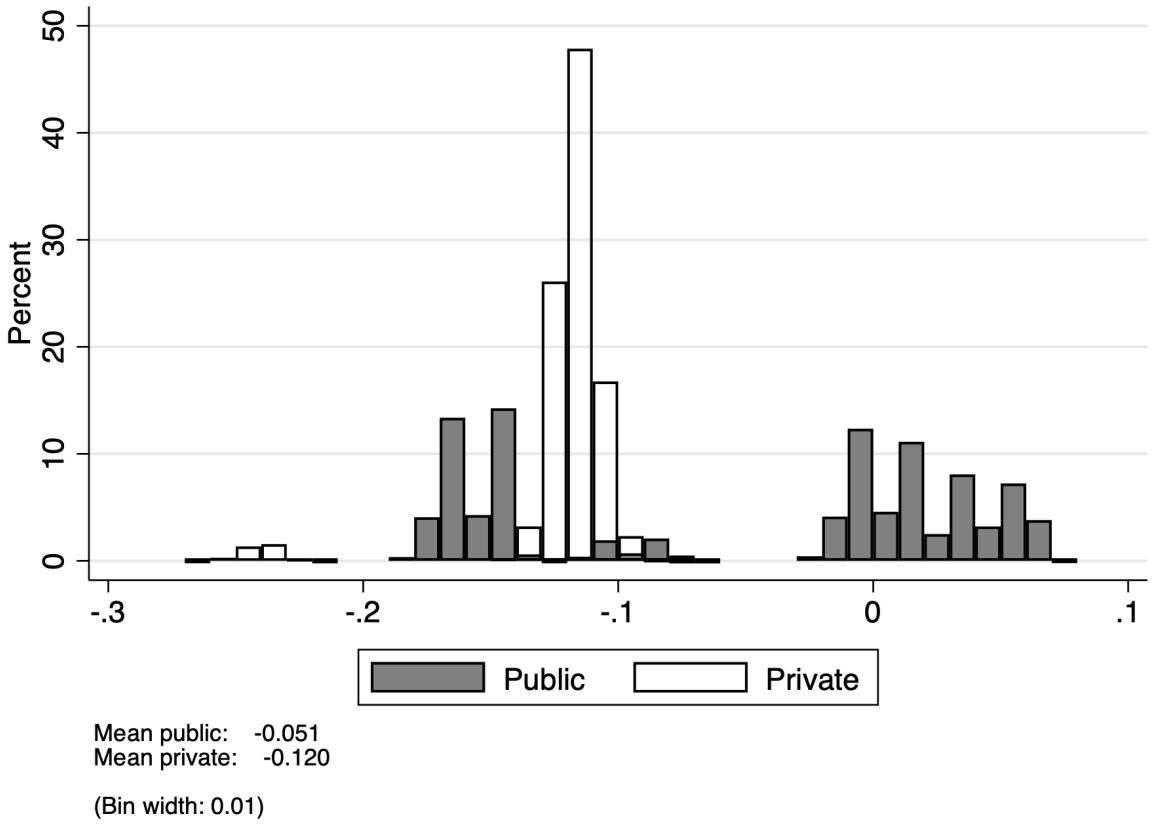
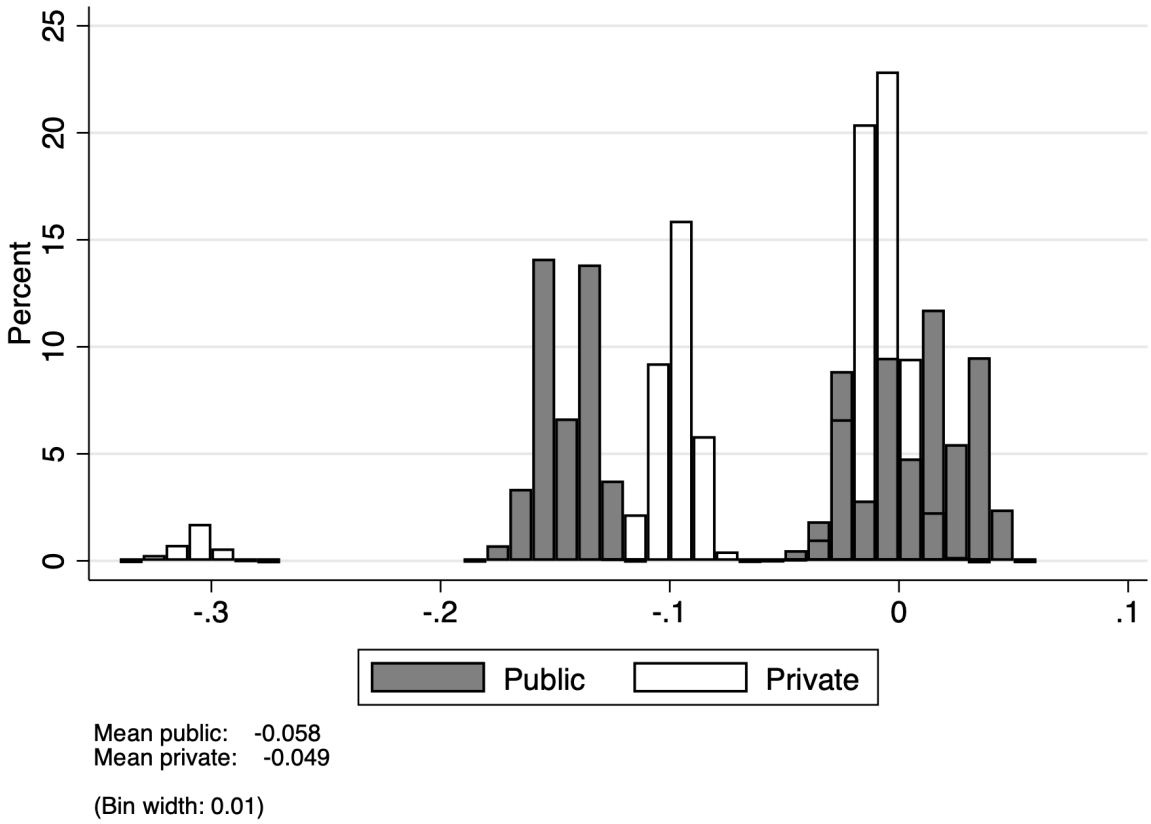


Figure 4: Distribution of treatment effects in reading



5.2 Sensitivity analysis

This section discusses the sensitivity of our results to different modeling choices. For ease of exposition, we show the difference with one of the outcomes looked at in this paper: math outcomes in public schools. We also focus the discussion on our main variable of interest: class size. This section serves two purposes. First, it shows that our results are conservative, finding more negative effects for reasonable changes of our assumptions. Second, it shows that assumptions often used in other work as a result of less rich data can be problematic.

Alternative GMM estimators Table 7 discusses alternative GMM estimators of the same model. The first exercise is to avoid the use of a two-step efficient GMM estimator and use the first-step results instead. While theoretically this is less efficient, this could perform better in finite samples when the number of instruments is large. This suggests our results are conservative as this shows a more negative effect of class size (and slightly larger standard errors).

The next column restricts the set of instruments to only contain lags 3 until 7, while the last column instead starts at lag 7. While both lead to more noisy results, we find that it is important to use deeper lags. While results are more imprecise with fewer instruments, it does suggest that the first-order Markov assumption might be too strong. This is what validates lags of 3 as instruments. It is possible that parents and teachers use information about older cohorts in their school choices and not just restricted to one cohort before (even after controlling for the fixed effect). Using deeper lags allows us to be more flexible in the timing assumptions regarding what information parents and teachers use when choosing schools.

Alternative production functions Table 8 discusses deviations of the production function.

In the second column of results, we use the average test scores by year and school as the outcome instead of first filtering out individual observed heterogeneity. Furthermore, we only include class size as an input. This means that the unobservable now captures much more school, teacher and student heterogeneity. While the estimate for class size gets more noisy, it remains robust. This suggests that our estimator does not suffer from omitting other characteristics of students or schools that we do not have in our dataset. Note that we obtain a more negative result here, which is what we would expected as it now captures an overall effect over the sample, while in the first column, the coefficient corresponds to the effect of a baseline category of male students with average incomes and without highly educated parents. The first step results indeed showed they have less negative class size effects.

The last column abstracts from school fixed effects. This is often done in the firm production function estimation literature due to the limited within-variation of inputs (De Loecker and Syverson, 2021). We have significant within-school variation in our context due to the voucher policy change. As the time-varying unobserved heterogeneity now needs to capture all school-level

Table 7: Sensitivity: second stage math results in public schools with different instruments

	Baseline	One-step	Recent lags	Non-recent lags
Class size	-0.048	-0.082	-0.006	-0.097
	(0.033)	(0.039)	(0.040)	(0.055)
Teacher exp	0.063	0.102	0.080	0.194
	(0.041)	(0.040)	(0.047)	(0.073)
Log(money/student)	0.063	0.087	0.104	-0.022
	(0.093)	(0.106)	(0.106)	(0.214)
ρ	0.540	0.425	0.566	0.279
	(0.033)	(0.042)	(0.040)	(0.075)
Year FE	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
Schools	1,576	1,576	1,576	1,576
Years	12	12	12	12
Observations	16,924	16,924	16,924	16,924

Second stage results of sensitivity analysis for math scores in public schools (without school types). First column repeats main result. Second column uses the one-step GMM estimator. Third column constructs instruments using lags of 3 to 7 periods. Fourth column constructs instruments using lags of 7 periods and earlier. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses.

unobserved heterogeneity, it is natural to see higher persistence in column 3. However, this is not sufficient to capture all endogeneity concerns as we now obtain a zero class size effect.

5.3 Understanding passive effects of the policy change

In this section we use the estimated effects to understand how passive effects of a policy change can influence the results. First, we show that part of the increase in test scores can be accounted for by changes in observable inputs. Then, we dig deeper into the school type-specific effects. Finally, we do a counterfactual exercise where we consider test scores if class size would have remained fixed.

Unobserved changes in school quality We use the estimates of specification (4) to decompose the evolution of test scores and obtain a better idea of the passive effects of the policy change. Figure 5 plots the estimated school-year fixed effects in differences with respect to 2007 (last year pre-policy), weighted by the number of students. This captures the change in test scores in the different types of schools that cannot be accounted for by observable student characteristics. The two graphs at the bottom show the differences in the estimated residual of the production function, which additionally filters out the estimated impact of observable inputs (money per student, teacher

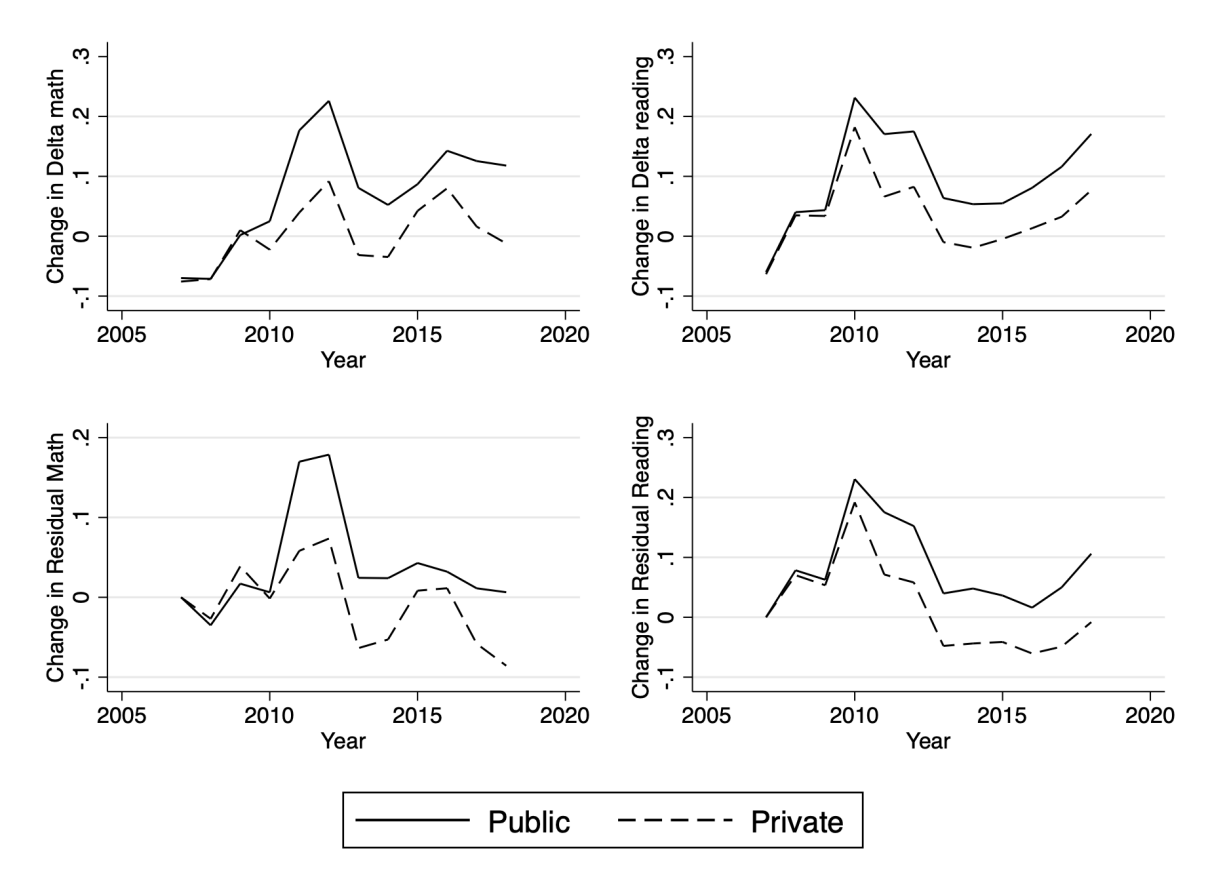
Table 8: Sensitivity: second stage math results in public schools with different controls

	Baseline	No controls	No FE
Class size	-0.048	-0.118	0.001
	(0.033)	(0.049)	(0.027)
Teacher exp	0.063		0.067
	(0.041)		(0.039)
Log(money/student)	0.063		0.071
	(0.093)		(0.094)
ρ	0.540	0.733	0.910
	(0.033)	(0.039)	(0.006)
Year FE	Yes	Yes	Yes
School FE	Yes	Yes	No
Schools	1,576	1,576	1,582
Years	12	12	12
Observations	16,924	16,924	17,257

Second stage results of sensitivity analysis for math scores in public schools (without school types). First column repeats main result. Second column omits observables in first and second stage, except for class size. Third column omits school fixed effects. Class size and teacher experience divided by 10 and subtracted by their mean. Standard errors in parentheses.

experience and class size).

Figure 5: School-year fixed effects and unobserved inputs



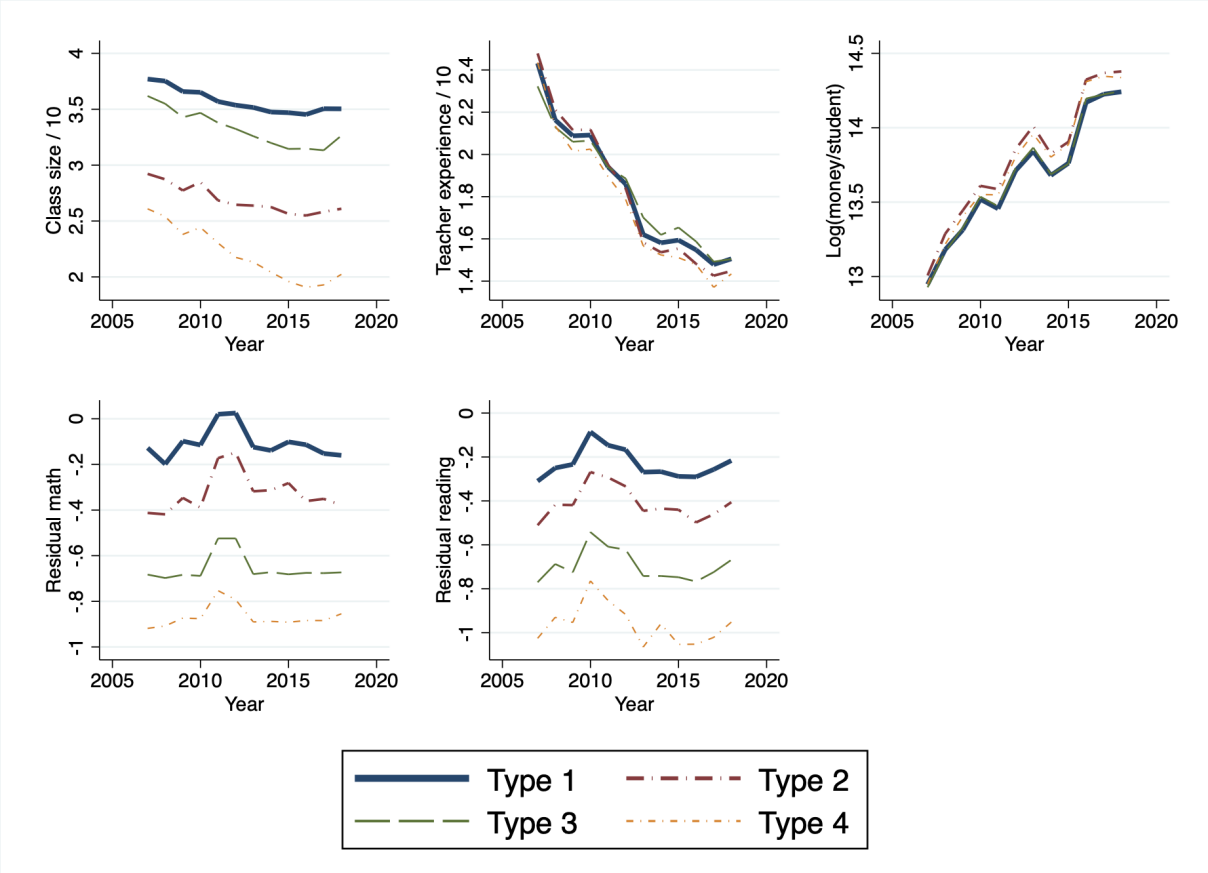
From the first row of this figure, we see that on average test scores in public schools increase substantially after 2007 (the first year we run our GMM estimation and the year before the policy started), while test scores in private schools show a much more moderate improvement. The evolution is very similar to what is observed for average test scores in Figure 1, suggesting that changes in test scores are not simply the result of changes in student sorting, at least not based on observable characteristics.

In the second row of this figure, we strip the fixed effects from the estimated effects through funding, teacher experience, and class size, also relative to 2007. Interestingly, the lines are much flatter. For public schools we observe a slight increase, while for private schools we see a decrease in the residual term. This suggests that schools did not improve scores by raising their unobserved quality, nor attracting students or teachers with better unobserved traits. Instead, effects are mostly driven by observed inputs. As can be observed from Figure 1, class size and teacher experience decreased, and financial resources increased. Since class size dropped significantly in public schools and not in private schools, we hypothesize this was an important driver for the convergence of test scores between the different types. To investigate this further, we first explore how class size effects are related to different school inputs, and then use these heterogeneous

estimates to do a counterfactual exercise for class size being fixed over time. Additionally, we want to relate our estimates to the changes in test scores for students of different socioeconomic status.

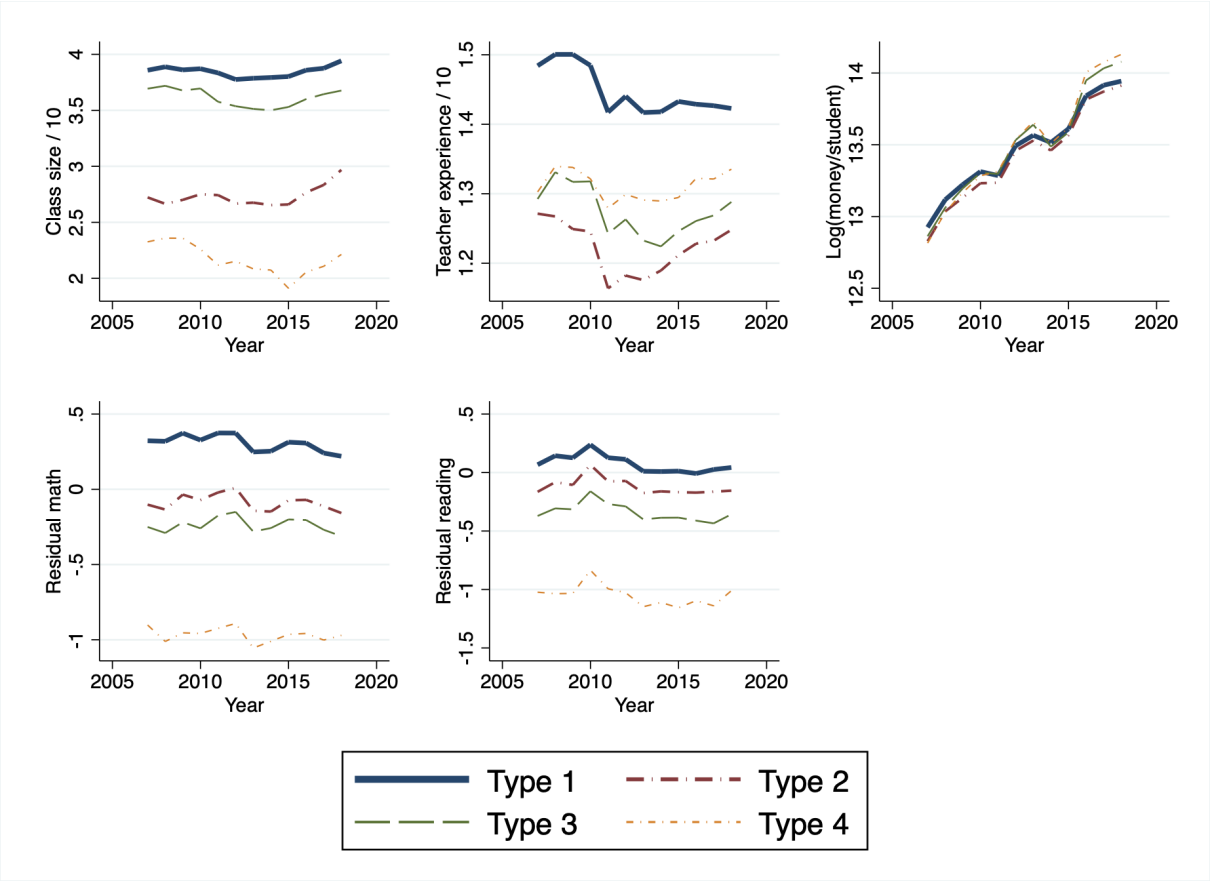
School-type specific decomposition Figure 6 and Figure 7 show these patterns for the specific school types (respectively in public and private schools). Before discussing the differential changes over time, we first compare how they differ from each other. Since the name of each type is arbitrary, we order them based on their average math fixed effects from best results (type 1) to worst results (type 4). While clustering happens independently in both types of schools, we do see a lot of similarities. Not surprisingly, the order by math fixed effects is identical to the ordering in the math residual (which includes the fixed effect, but also the time-varying AR(1) unobservable) and the reading residual. More interesting, the best type also has the largest classes, while the third type has only slightly smaller classes, but much larger than the second type. Money per student is similar over types. Teacher experience is similar for public schools, but in private schools, there is more heterogeneity. The best schools (based on unobservables) also have the more experienced teachers, while the second best schools have the least experienced.

Figure 6: Heterogeneity by school type: public schools



When we compare these patterns to the heterogeneity in class size effects, we can conclude the following. First, we should keep in mind that negative class size effects in public schools are

Figure 7: Heterogeneity by school type: private schools



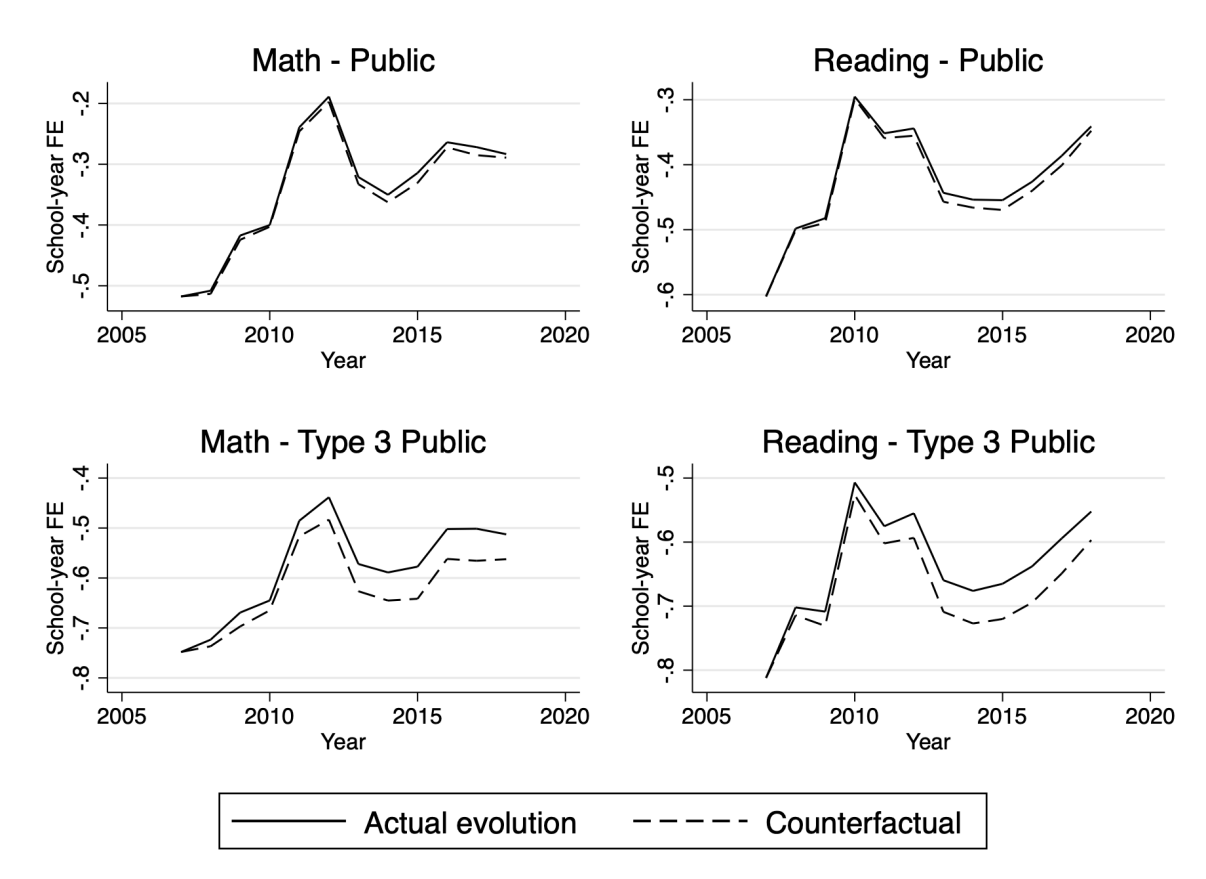
primarily driven by type 3, especially for math. This seems to point to a school type that is characterized by large classes, but low unobserved quality. Comparing this with type 1, it suggests that unobserved school, teacher, and student quality (which enter additively) in test scores, is related to the ability to handle large classes. Second, while type 4 only represents a small share of students, it is interesting to zoom in on its results in private schools. These schools show very negative treatment effects, while they also have very young teachers. This suggests that experience can play an important role in handling large classes.

Finally, we focus on the trends over time. We see that class size drops in all public schools, but particularly in the schools with the lowest performance (types 3 and 4). Apart from sorting to likely better schools, the decrease in class size for type 3 schools is particularly welcome as this type struggled most with handling large classes.

Counterfactual simulation of class size effect The heterogeneity in effects, combined with the changes in characteristics, can explain how the resorting of students due to the policy and its effects through class size contributes to the improvements we see in test scores, especially in public schools. In Figure 8 we plot the counterfactual evolution of δ_{jt} for public schools if class sizes had remained fixed at their 2007 level. We focus on public schools as class sizes do not change much

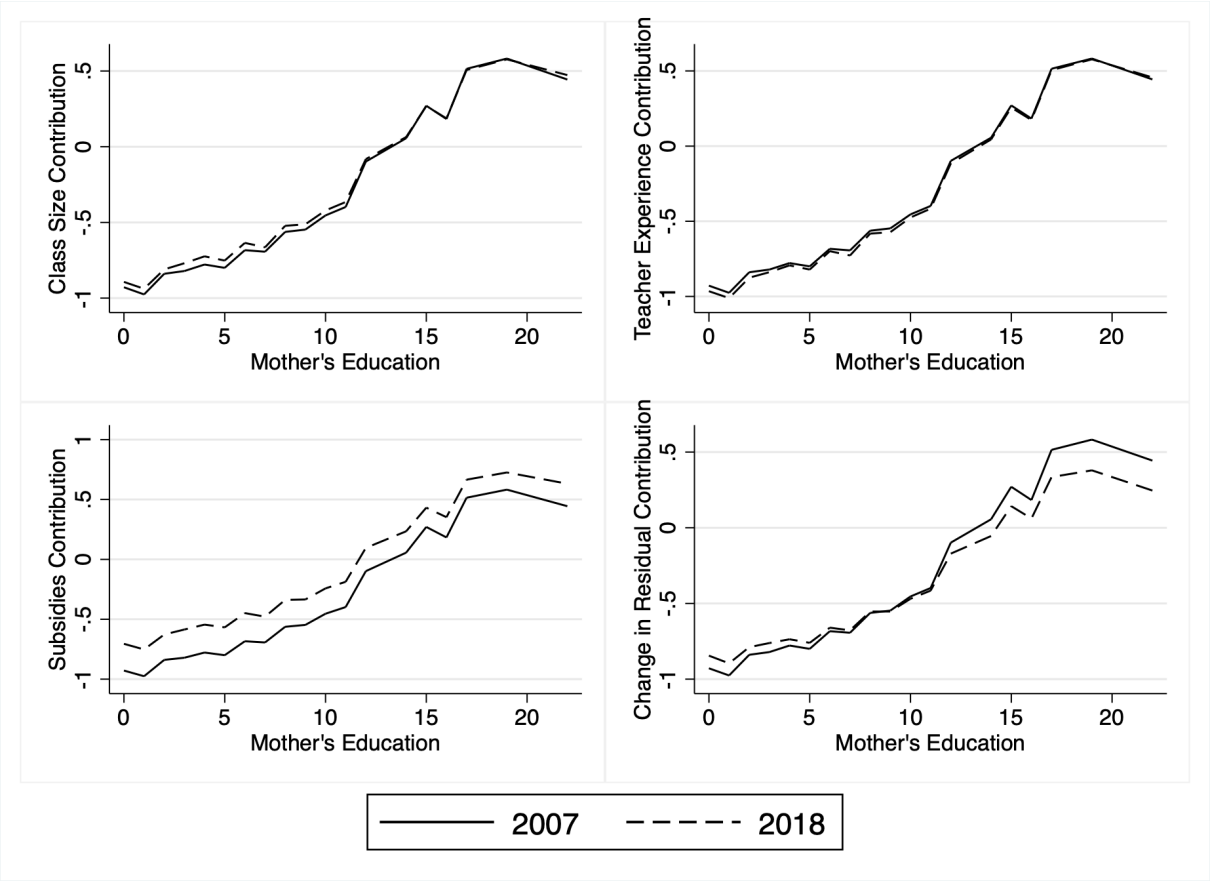
in private schools. The top row shows the comparison for all public schools and the bottom row focus on the type 3 schools. Looking at all schools, the overall class size effect is quite small, but we find a noticeable impact on type 3 schools. These schools account for 37% of all schools, have a very negative class size effect combined with quite large classes.

Figure 8: Counterfactual test score with constant class size: Public Schools



Input Contribution by Student Socioeconomic Status Figure 9 shows the contribution of each of our three observed inputs to the changes in test scores from 2007 to 2018 by student mother's education. We can see that the gains in scores explained by class size are concentrated in the bottom part of the socioeconomic distribution. The effect of teacher experience goes in the opposite direction, as experience decreases for low-SES students. The increase in subsidies has a large impact, but it is mostly the same for everyone. The unexplained residual component seems to be contributing to the decrease in test score inequality shown in Figure 2 as it is positive for low-SES and negative for high-SES.

Figure 9: Input contribution by student socioeconomic status



6 Conclusion

We provide a new education production function that can be used in a context with school choice with rich data on student test scores and characteristics of schools with repeated data of the schools. We used it to study the channels behind changes in school quality following a voucher reform in Chile and found important effects through class size reductions in public schools. Our paper highlights the importance of taking into account the "passive effects" of policy changes as class size reductions might simply have been the result of driving out students because of the voucher reform.

Future research could use our production function estimation approach in other contexts where rich panel data is available and the researcher wants to allow for unobserved quality differences in production factors.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 2411–2451.

Adusumilli, K., F. Agostinelli, and E. Borghesan (2024): "Heterogeneity and Endogenous Compliance: Implications for Scaling Class Size Interventions," Tech. rep., National Bureau of Economic Research.

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011): "Do value-added estimates add value? Accounting for learning dynamics," *American Economic Journal: Applied Economics*, 3, 29–54.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly journal of economics*, 114, 533–575.

Arellano, M. and S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The review of economic studies*, 58, 277–297.

Blundell, R. and S. Bond (2000): "GMM estimation with persistent panel data: an application to production functions," *Econometric reviews*, 19, 321–340.

Bonhomme, S. and E. Manresa (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

Contreras, D., P. Sepúlveda, and S. Bustos (2010): "When Schools Are the Ones that Choose: The Effects of Screening in Chile," *Social Science Quarterly*, 91, 1349–1368.

Correa, J. A., F. Parro, and L. Reyes (2014): "The effects of vouchers on school results: evidence from Chile's targeted voucher program," *Journal of Human Capital*, 8, 351–398.

De Loecker, J. and C. Syverson (2021): "An industrial organization perspective on productivity," in *Handbook of Industrial Organization, Volume 4*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, Elsevier, vol. 4 of *Handbook of Industrial Organization*, 141–223, iSSN: 1573-448X.

Ding, W. and S. F. Lehrer (2014): "Understanding the role of time-varying unobserved ability heterogeneity in education production," *Economics of Education Review*, 40, 44–75.

Finn, J. D. and C. M. Achilles (1990): "Answers and questions about class size: A statewide experiment," *American Educational Research Journal*, 27, 557–577.

Gazmuri, A. (2018): "School segregation in the presence of student sorting and cream-skimming: Evidence from a school voucher reform," Tech. rep., TSE Working Paper.

Hsieh, C.-T. and M. Urquiola (2006): "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's voucher program," *Journal of Public Economics*, 90, 1477–1503.

Krueger, A. B. and D. M. Whitmore (2001): "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR," *The Economic Journal*, 111, 1–28.

Mineduc (2015): "Subvencion Escolar Preferencial, Superintendencia de Educacion Escolar, Ministry of Education, Government of Chile," `https://www.ayudamineduc.cl/ficha/subvencion-escolar-preferencial`.

Murnane, R. J., M. R. Waldman, J. B. Willett, M. S. Bos, and E. Vegas (2017): "The Consequences of Educational Voucher Reform in Chile," Tech. rep., National Bureau of Economic Research.

Navarro-Palau, P. (2017): "Effects of differentiated school vouchers: Evidence from a policy change and date of birth cutoffs," *Economics of Education Review*, 58, 86–107.

Neilson, C. (2017): "Targeted Vouchers, Competition Among Schools, and the Academic Achievement of Poor Students," .

Todd, P. E. and K. I. Wolpin (2003): "On the specification and estimation of the production function for cognitive achievement," *The Economic Journal*, 113, F3–F33.

Urquiola, M. and E. Verhoogen (2009): "Class-size caps, sorting, and the regression-discontinuity design," *American Economic Review*, 99, 179–215.

Weinstein, J., A. Fuenzalida, and G. Muñoz (2010): "La Subvención Preferencial: desde una Difícil Instalación hacia su Institucionalización," *Fin de ciclo*, 161–282.

Word, E., J. Johnston, H. P. Bain, B. Fulton, J. B. Zaharias, C. M. Achilles, M. N. Lintz, J. Folger, and C. Breda (1990): "The State of Tennessee's student/teacher achievement ratio (STAR) Project," *Tennessee Board of Education*.

Table 9: Strength of instrument

	Math	Reading	Class size	Teacher exp	Log(money/student)
Public schools					
Lag 3 coefficient	-0.062	-0.060	-0.034	-0.126	-0.043
Lag 3 p-value	0.000	0.000	0.000	0.000	0.000
Lag 7 coefficient	-0.033	0.024	-0.014	-0.101	-0.021
Lag 7 p-value	0.000	0.006	0.009	0.000	0.000
Lag 3 to 6 p-value	0.000	0.000	0.462	0.000	0.000
Lag 7 to 10 p-value	0.290	0.364	0.306	0.000	0.000
Private schools					
Lag 3 coefficient	-0.045	-0.054	-0.023	-0.061	-0.012
Lag 3 p-value	0.000	0.000	0.000	0.000	0.000
Lag 7 coefficient	-0.025	0.008	-0.009	-0.035	0.003
Lag 7 p-value	0.000	0.210	0.015	0.000	0.321
Lag 3 to 6 p-value	0.003	0.000	0.036	0.000	0.000
Lag 7 to 10 p-value	0.030	0.284	0.062	0.009	0.000

Tests indicating instrument strength. For each school type and variable, we run four OLS regressions with clustering at the school level. In a first regression we regress the difference over time of the variable in the top row on its third lag. We report its coefficient and p-value. In a second regression we do the same but with the seventh lag. A third regression includes all lags 3 to 6 and reports the p-value of a joint singificance test. The final regression repeats this for lags 7 to 10.