Does Class Size Matter? How, and at What Cost?*

Desire Kedagni[1], Kala Krishna[2], Rigissa Megalokonomou[3], and Yingyan Zhao[4]

[1]Iowa State University

[2]The Pennsylvania State University, NBER and CES-IFO

[3]University of Queensland and IZA

[4]George Washington University

December 23, 2020

Abstract

Using Greek high quality administrative data we show that class size has a hump shaped effect on achievement whereas most of the literature assumes linearity. We then embed our estimates for this relationship in a dynamic structural model with costs of hiring and firing. We find that class sizes around 19 maximize the attainment, and the existing costs result in an average class size of 22-23 being chosen. Firing costs are about the same size as hiring costs consistent with the presence of unions. Reducing firing costs to zero raises class size by 4-6 students and reduces achievement by 0.8 to 1.2 GPA points. Reducing hiring costs to zero reduces class size by 3 to 5 students and increases GPA by 0.3 GPA points. Raising wages by 50%, class size increases by 2 students and GPA falls by 0.3 GPA points suggesting that unions (which raise salaries and firing costs) do raise class size, but have a small effect on achievement. We show that class size caps are costly, and more so for small schools, even when they are set at levels well above the average class size.

Keywords: Class Size, Student Performance, Adjustment Costs, Dynamic Model

JEL: C6, IE, JE

*We thank the seminar participants at University of Melbourne and University of Tokyo. We also thank Nino Doghonadze for her excellent work as a research assistant and Ilias Arvanitakis who helped us to compile the data. We are grateful to the Editor, an Associate Editor and two anonymous referees for helpful comments.

1 Introduction

What determines student achievement? In this paper, we focus on the effects of class size on achievement. This area has been widely studied in both labor economics and education. Somewhat surprisingly, the estimates are relatively mixed. Leuven et al. (2008) summarize the state of the debate as follows:

> "One of the still unresolved issues in education research concerns the effects of class size on students' achievement. It is by now well-understood that endogeneity problems may severely bias naive OLS estimates of the class size effect, and that exogenous sources of variation in class size are key for a credible identification of the class size effect. Various recent studies acknowledge this and apply convincing identification methods. This has, however, not led to a definite conclusion about the magnitude or even the sign of the class size effect."

While performance has been related to class size, there has been little attempt to allow for nonmonotonicities.[1] For example, it could be that larger class size first raises (as students learn from each other as well as the teacher) and then lowers achievement (when congestion effects take over). In this paper, we explicitly allow for such possibilities. We argue that not allowing for nonmonotonicities could be why the literature has found mixed results.

We use high quality administrative data from Greece to empirically confirm the existence of a hump-shaped relationship. We estimate the relationship between class size and achievement while carefully dealing with issues of endogeneity of class size. We show that class size does matter and that the linear/monotonic specification form used in past work may be why past results were mixed. After all, if we fit a monotonic regression when the true relationship is non-monotonic, we could get a positive, negative or zero slope depending on the precise shape of the underlying true quadratic relationship. Our estimates suggest that the shape of this relationship is relatively flat in the relevant region, namely the region close to the chosen class size. As a result, a marginal mandatory reduction in class size can have a small positive effect on achievement. Moreover, as the chosen class size, in the presence of adjustment costs, will exceed the class size at which achievement is maximized, a large reduction in class size could easily move achievement to the other side of the hump and have little or no effect on achievement. In addition, the effect of increases versus decreases in class size can be very asymmetric. All of this is consistent with what the literature has found: namely that decreasing class size is a costly way of raising achievement.

[1] Most work assumes a monotonic form.

Our contribution in this paper is to bring together reduced-form analysis that identifies the causal effects of class size (while explicitly allowing for nonmonotonicities) and dynamic structural estimation, which is then used to study possible policy implications using counterfactual exercises.

This paper proceeds as follows. In Section 2, we put our work in perspective relative to the literature. In Section 3, we describe the institutional setup and data, present some summary statistics and descriptive regressions both parametric and nonparametric that suggest a hump shaped relationship in class size and achievement. In Section 4, we take a more parametric approach and use enrollment as an instrument for class size to control for endogeneity.

With the estimates of the effects of class size on achievement in hand, we are in a position to understand how class size is chosen. In Section 5 we use our reduced form estimates in a dynamic structural model of class size to estimate hiring/firing and marginal cost of adding a class. The manner in which the number of classes chosen varies as enrollment fluctuates helps us estimate the costs involved. Our estimates here are sensible in view of the institutional setup. While temporary teachers are easier to fire than permanent ones (who are rarely fired since doing so would involve compensation as well as union pushback), both types are typically used.[2] Firing costs are estimated to be about the same as firing costs and about two and a half to three times a teacher's yearly salary. Finally, in Greece, as in much of the rest of the world, teachers unions are a powerful force to be reckoned with. Their power is expressed not only in terms of wages set but in terms of the ability to fire teachers at will. We use the model to ask whether inflexibility in terms of unions creating high firing (and even maybe hiring) costs might be driving class size choices by school administrators and the impact of this on student achievement if any. We find that unions, even though they raise costs and class size, have a small effect on achievement. Finally, we look at the costs versus benefits of class size requirements. We find that these have significant effects even when set above the targeted class size. Moreover, they have very different effects on small and large schools with small schools being more affected by them. Section 6 concludes.

2 Relation to the Literature

Given the increasing importance of skills in the labor force in this age of robotics and artificial intelligence, there is intense interest in what drives educational attainment. A small part of this debate has focused on the role of class size on achievement. An excellent, though slightly dated survey can be found in Hanushek (2003) and Rivkin et al. (2005).

The main problem is that class size itself is a choice, i.e., it is highly endogenous. Teachers

[2]Temporary teachers are not good substitutes for permanent ones as explained in Section 3.

and headmasters are better informed about students than the econometricians. They may choose, for example, to allocate better students to larger classes or better teachers to larger classes, thus generating a positive correlation between class size and student performance and biased estimates. In this event, OLS estimates of the coefficients on class size cannot be interpreted causally. This is not a problem specific to class size, but is more general. For example, estimating effects of other school inputs on pupil outcomes is also complicated by potential endogeneity issues.[3] The usual way to deal with this problem is to have a good instrument or an experiment and this is essentially the route the literature has taken.

On of the best known experiments is the Tennessee STAR experiment. Students were randomly assigned to different sized classes. This should make it straightforward to estimate at least the policy effect of class size. However, there remain concerns about whether teacher quality changed, and the attrition and entry of students throughout the experiment (which could also have been endogenous) could confound the results (Hoxby, 2000; Hanushek, 1999). Krueger (1999) and Krueger and Whitmore (2001) find that smaller class sizes in kindergarten and first grade seemed to have a significant and lasting positive effect on academic achievement.

More recently, Jepsen and Rivkin (2009) study California's class size reduction program for grades K-3. This reduced class size on average from 30 to 20 thus hiring 50% more teachers at a cost of roughly a billion dollars. They find this policy raised math and reading achievement by roughly .10 and .06 standard deviations of the average test scores respectively, holding other factors constant. This is about the same effect as that of having a teacher with two more years of experience. Assuming teachers' salaries rise at less than 25% per year of experience, class size reductions would seem the more expensive option.[4]

In contrast to much of the work using field experiments above, an elegant and often used quasi experimental approach is based on class size limits which turn out to be relatively common. Angrist and Lavy (1999) noticed that in Israeli public schools, by law, there could be no more than 40 students in a class. Thus, if a cohort grew beyond 40, there would be an exogenous fall in class size from 40 to 21, while if the cohort grew over 80, there would be an exogenous fall in class size from 40 to 27, and so on. They show that without correcting for endogeneity, class size is positively associated with achievement, but when endogeneity is controlled for the sign is reversed. This makes economic sense as when students are good, larger class sizes can be tolerated which will bias OLS estimates upwards. Their estimates are for grades 3, 4 and 5. The coefficient on

[3]School inputs are chosen by parents, school administrators, teachers, and politicians at both local and national levels. For instance, parents locating close to resource abundant schools may have chosen to locate there because they care a lot about their children's education and so also invest more time in their children's education (creating an upward bias).

[4]It is also worth noting that the increase in demand for teachers resulted in a fall in their quality.

class size is not significant for grade 3, but is significantly negative for grades 4 and 5. In general, estimates suggest that class size is a costly way of improving achievement.

Other papers which exploit maximum class-size rules include Bonesrønning (2003) for Norway, Urquiola (2006), Browning and Heinesen (2007) and Bingley et al. (2007) for Denmark. Browning and Heinesen (2007) focus not only on class size but also on teacher hours per student. The class size is limited to 28 students in Denmark. However, Bingley et al. (2007) find that the target class size in the data is closer to 24 suggesting that the limit is not binding and the quasi experimental approach is invalid.

Levin (2001) and Dobbelsteen et al. (2002) use a method similar to the maximum class-size rules to study the class size effect in Netherlands. Dutch rules tie the number of teachers to enrollment and this provides quasi exogenous variation in the number of classrooms. The data they used is PRIMA data. This longitudinal survey of Dutch students in grades 2, 4, 6 and 8 in 1994-5 is rich in information including IQ. Levin (2001) explores peer and quantile effects. Dobbelsteen et al. (2002) also find strong peer effects on student achievement. Controlling for peer effects, they find class size effect to either be insignificant or significantly negative.

The other approach to correct for endogeneity of class size is related to the work of Hoxby (2000). In the absence of binding class size limits, one might think of using variations in overall enrollment as exogenous shocks. Hoxby goes a step further: she fits a quartic to the enrollment data and uses deviations from the quartic as the exogenous variation. In this way, she controls for trends in enrollment.

In our work, as there is no explicit class size cap, we cannot use the Angrist and Lavy approach. As a result, we use enrollment as the instrument. We could have followed Hoxby (2000), who used deviations from predicted enrollment as an instrument.[5] Our results are very similar in either case as shown in the Online Appendix. The reduced-form analysis in the literature identifies the causal effects of class size. Our study complements the literature in two ways. First, it shows that the focus on linear/monotonic functional forms as is prevalent in the literature seems inappropriate. A non-monotonic functional form seems to be called for. Second, by embedding our causal estimates for the effect of class size on performance in a structural model where class size is chosen subject to costs of running a class as well as of adjusting the number of classes, we are able to estimate key structural parameters and to examine the effect of a number of different policies on outcomes by doing counterfactual exercises.

[5] Gary-Bobo and Mahjoub (2013) use data on French junior high schools to look at the effect of class size on promotion, and Urquiola (2006) uses Bolivian data to look at the effect of class size on performance. Both follow Hoxby's approach in spirit. Although the estimated causal effects of larger class size tend to be negative, they remain small.

Most of the literature, including that discussed above, uses a linear/monotonic specification. There are a few papers that allow for nonmonotonic effects Borland et al. (2005) use data from the Kentucky Department of Education for the third grade in 1989-90. They specify a four-equation simultaneous equation system. Class size, achievement, market competition and teacher salary are the endogenous variables and achievement is allowed to be a quadratic function of class size. They argue that class size and GPA could be nonmonotonic. Why? Students learn from peers like themselves and the larger the class, the more likely it is that they have peers like themselves and GPA rises with class size. On the other hand, there is crowding and ultimately these congestion forces dominate so that GPA first rises with class size and then falls which is what they find. Borland et al. (2005) find evidence for a non-monotonic class size effect on achievement. They find the peak of achievement as a function of class size to be 26 even though most class sizes are below this number. This is difficult to reconcile with optimizing behavior as a school could reduce cost and raise achievement by raising class size.[6]

Hojo (2013) also allows for a nonlinear relationship between class size and GPA using a spline regression and accounting for endogeneity. He studies the effect of class size on the performance of fourth grade Japanese students. He finds a negative effect throughout, although the relationship flattens out for larger class sizes. Using his approach does not affect our results. It could be that the difference in our results and his comes from differences in the teaching style and philosophy. With an interactive teaching style, performance may increase as class size rises initially (because having peers to interact with helps learning) and fall when congestion effects start to dominate. With a less interactive approach than that in Greece, there may be no such non monotonicity and student performance may not fall much once class size gets large enough. For such reasons, there is little reason to expect a one size fits all curve that works for all settings.

We find that class size effects are nonmonotonic in our data, with class size initially increasing and then reducing achievement. We are able to control for teacher fixed effects, though only for a limited subsample. We show that our results are quite robust across a range of specifications. It could be that this hump shape is why restricting the functional form to be monotonic gave estimates that were small in size and variable in sign. However, it is worth emphasizing that much

[6] Bandiera et al. (2010) use rich data on student performance in undergraduate classes in the UK. They allow for both nonmonotonicities and quantile effects. However, they assume that assignment of students to classes is random as they have no instrument. Their data has student performance over time as well as teacher assignment so that they can incorporate both teacher and student fixed effects. Though they allow for nonmonotonic effects, they find class size always reduces performance, though the effect is not linear. Moreover, they find that class size seems to affect better students more. Kokkelenberg et al. (2008) use data on a large number of undergraduate observations from a northeastern public university and find that class size negatively affects grades for a variety of specifications and subsets of the data. Their rich data allows for many controls including academic department, peer effects (relative ability in class), student ability, gender, minority status, and other factors. GPA declines monotonically as class size increases but this is less so for larger class sizes.

of the work above uses data on lower grades. In contrast, our data is for high school students in Greece. It may well be that class size effects differ greatly depending on the context: for young students class size may have a large effect while for older students the effect may be smaller or vice versa. Similarly, effects may be subject specific or differ in intensity by sub groups.

3 Institutional Background and Data

The Greek education system is run by the Ministry of Education, Research and Religious Affairs. It exercises control over the state schools in terms of curriculum, staffing and funding. Teachers are civil servants and get a salary based on seniority, location and family size. There are two tracks for teachers: permanent and temporary or substitute teachers. The former got tenure after two years of employment before 2013, though this is no longer the case. Teaching needs are first met by utilizing existing permanent staff, then by hiring temporary staff and only as a last resort adding a permanent teacher. As there is an excess supply of teachers for High School, it is relatively easy to hire on a temporary basis. Temporary jobs are allocated on the basis of date of graduation and experience. A new graduate has to acquire experience as a temporary teacher to land a job as a permanent one. Temporary jobs can be anywhere in Greece and turning down an offer results in being blacklisted. The process is centralized and complex. For more detail on this, see Dinerstein et al. (2020). In the period of our data only about 5-7% of teachers were hired on a temporary basis. This number has increased significantly in recent years after a ban on hiring permanent teachers in 2009 and budgetary pressures facing schools.

Temporary teachers get paid on the same scale as entry level permanent teachers, but only for the work they do. They may work part time at an hourly wage at one or more schools, or full time. Their contract is a 10 month one and as their allocation is done centrally in response to temporary needs on the part of a school, it is unlikely that they would be reassigned to the same school. Thus, the only way to hold on to a specific teacher is to hire them as permanent staff. Permanent staff is very difficult to fire, especially in public schools where firing permanent staff is almost unheard of. Not only is there compensation to be paid, but union involvement results in strikes in response to such actions. Teachers can be fired for an inability to do their job but documenting this is very difficult. Even in private schools, severance pay for permanent teachers includes a month's salary for every year of seniority up to 25 years. Even hiring teachers is costly. The hiring process is centralized and schools have little choice in whom to hire. Whoever graduates with an education degree and wants a public sector teaching position are ordered by the date of their degree and allocated jobs as they arise. There is excess supply in the market, so

that new graduates may wait quite a while to get a job. See OECD (2005) for details of how the system works. The yearly salary for a teacher with 15 years of experience and minimum training was about 20,000-21,000 Euro in 2004.

In Greece the government provides free education up to 12th grade for all students. There is an exam for entrance to university but no tuition is charged. This is because the Greek constitution says that all Greeks (and some foreigners) are entitled to free education. State-run schools and universities even provide textbooks free to all students, although, from 2011 onward, shortages have occurred. There are private cram schools that operate side by side with the high schools where students go for extra tuition to perform better in exams, and this is especially so in the 11th and 12th grades.[7] Most of the students attend such classes in the afternoon and evening in addition to their normal schooling. Private universities and colleges operate alongside the public ones.

In the 10th grade, students have, for the most part, a common curriculum.[8] In the 11th and 12th grade, they start to differ as they choose their tracks.[9] At the end of 12th grade, most students take the university entrance exam. Their performance in this exam, together with their performance in high school determines their placement score for entrance into university.[10]

Students are assigned to a particular class. Students in a class stay together for all non track subjects and teachers move from one class (equivalent to classroom) to another class (classroom). In the 10th grade, there are no track subjects and so students stay together throughout the day. Moreover, they are less likely to attend cram schools or take private tutoring in the 10th grade as the university entrance exam is still some time away. This is relevant because such tutoring would be an omitted variable that affects performance that we cannot control for. Also, there are likely to be more unexpected shocks to enrollment for the 10th grade, than for higher grades as the incoming class comes from several feeder Junior High Schools which makes enrollment a good

[7] Cram schools are popular in a number of OECD countries. Out of all OECD countries, Greece is the country with the second highest number of minutes spent attending after school classes/cram schools, ranking just after Korea. See OECD (2013).

[8] 10th grade compulsory subjects include religion, ancient Greek, literature, modern Greek, history, algebra, geometry, physics, chemistry, economics, technology and one foreign language.

[9] 11th grade compulsory subjects include religion, ancient Greek, literature, modern Greek, history, algebra, geometry, physics, chemistry, biology, introduction to law, a foreign language and 3 track subjects (which are fixed within each track). Students are required to attend these subjects in eleventh grade and they take either school or national exams in each one of them. In the 12th grade, they finalize a specialty/track of which there are three: Classics, Science and Information Technology. 12th grade compulsory subjects are religion, literature, modern Greek, ancient Greek, history, physics, biology, mathematics, a foreign language (either English, or German, or French) and 4 track subjects (which are fixed within each track). Students are required to attend these subjects in twelfth grade and they take either school or national exams in each one of them. All other subjects are optional.

[10] With their placement score in hand they list their preferences. Students are admitted not to schools but to programs within schools. We do not focus on entrance to university here and do not use the data on preferences, entrance exam scores, placements scores and final placements here.

instrument.[11] For all three reasons we focus our attention on the 10th grade data.

The data used in this paper was obtained from the local school authorities and covers 123 public high schools in Greece. Most students in Greece attend public schools. Our data covers roughly 10% of the public high schools in Greece. The time period is 2001-2013.

The data we use includes the following: the exam scores of the student in the school exams in 10th grade for non track subjects. The gender, age, number of classrooms for each grade in the school, class size, cohort size and total enrollment in each school. We also have performance in the first term, the second term, as well as the school annual exams. The annual exams are course and teacher specific. The Principal is supposed to ensure a common standard is adhered to.[12] Performance is measured on a continuous scale from 0-20. We take the simple average of the annual exams across non track compulsory subjects (Ancient Greek, Literature, Modern Greek, History, Algebra, Geometry, Physics, Chemistry, Economics, and Technology) to get the performance measure we call GPA for each student. We choose to use the annual exam as it is less likely to be subjective compared to evaluations based on performance over the term. We know the name of the school, the type of school (public, public elite, evening, private), and whether the school is urban or rural. We chose to not use evening school data as these schools are very different from regular high schools: they have a very different set of guidelines, much larger class sizes and more mature students. Elite schools are entered by passing an exam and are for gifted students but they are few in number and as a result we have no elite schools in our subsample of schools. The inputs available to private schools are likely to be very different and the student mix may also differ. For these reasons we chose to restrict ourselves to public schools. All schools operate under the same guidelines as the educational system is highly centralized.

In Greece, performance in high school matters because university placement depends on the performance in the university entrance exam (70%) and on high school exams (30%). However, performance in 10th grade is not included in this. It matters in terms of which track is chosen in the 11th grade and an average score of 50% in school exams is needed to sit for the university entrance exam.

[11] In the 11th and 12th grade, enrollment tends to lie below the enrollment in the previous year for the grade below, while in the 10th grade enrollment could lie above or below that for the 10th grade in the previous year. Also, class number in the 10th grade is more closely related to the enrollment. This is shown in Table 1 in the Online Appendix.

[12] Nor are teachers evaluated by their students. As a consequence we are not worried about teachers having different grading standards.

3.1 Summary Statistics

Table 1: Sample means and standard deviations

		All Grades	Urban	Rural	Rural - Urban
		Individual Level Data			
GPA	Mean	11.79	11.80	11.61	-0.192**
	Std. Dev.	(3.79)	(3.79)	(3.86)	(-2.73)
Female	Mean	0.54	0.54	0.56	0.0210*
	Std. Dev.	(0.50)	(0.50)	(0.50)	(2.29)
Age	Mean	15.97	15.97	16.01	0.0462***
	Std. Dev.	(0.60)	(0.58)	(1.04)	(3.97)
Obs		81845	78816	3029	
		Class Level Data			
Class Size	Mean	22.62	22.83	18.28	-4.547***
	Std. Dev.	(4.15)	(4.02)	(4.36)	(-14.22)
Obs		3641	3474	167	
		School Level Data			
Cohort Size	Mean	76.17	81.65	27.75	-53.89***
	Std. Dev.	(33.90)	(31.06)	(13.00)	(-18.02)
Class Number	Mean	3.37	3.58	1.52	-2.060***
	Std. Dev.	(1.24)	(1.12)	(0.59)	(-19.06)
Obs		1082	972	110	

(1) The data used in this paper was obtained from the local school authorities and covers 123 public high schools in Greece.

Here we share some patterns in the data that motivate much of what we do below. Table 1 shows the mean and standard deviation for the key variables we use. Note that class size is relatively concentrated around the mean. In fact, 90% of the data has a class size between 16 and 28. The average school has 3 or 4 classes in a grade, but there is a lot of variability here. Rural areas usually have small schools with lower enrollments, smaller class size and number of classes that range from 1 to 3, while urban areas have larger schools with as many as 9 classes. GPA tends to be lower in schools in rural areas. Note that we have school fixed effects in our baseline regressions below which would absorb such differences between schools. We have up to 12 years of data for each school. Table A.1 in Appendix A shows the panel composition over number of years and

cohort size. Larger schools have a slightly longer panel due to data availability.

The sample of schools that we use in this study does not differ systematically from the population of schools in Greece. While we cannot find nationwide data on 10th grade students, we did find such data for students in the 12th grade. To show this, we compare various variables between our sample and the whole population of schools in the country. In particular, we provide comparisons in terms of some predetermined characteristics (gender and age), average performance of the school in terms of the senior year's exams, but also university admission, and percentage of public schools in the sample and the population. The data for the whole population of schools come from the Ministry of Education and Religious Affairs[13] and includes all schools in the country, except for the evening schools (which are only designed for employed students). Our sample is representative in terms of gender and age of students and school performance in the senior year of high school. In particular, the percentage of female students in the sample and the population is 56% and their average age when they start their senior year is 17.3 in the sample and 17.4 in the remaining schools (p-value for the difference=0.52). In terms of the average school performance in the senior year, the sample is also representative, since the average (national and school exams) performance of the sampled schools is 14.45/20, while the average performance of the remaining schools is 14.34 (p-value for the difference=0.33). We also observe the same fraction of public schools in our sample (87%) as in the remaining schools (85%), while the difference is insignificant (p-value=0.49), and the students' (log) university admission score is similar in the sample (9.48) and the remaining schools (9.46).

Furthermore, the OECD[14] reports that the number of students per class for the grades that we are looking at is on average 23 students (in 2005 it was 24 and in 2010 it was 22), a figure that is quite close to the class size that we observe in our data (22.6).

3.2 Data Patterns

3.2.1 Class Size and Enrollment

No cap on class size was in place officially in Greece for the period of our data. A consequence of this is evident in Figure 1 which plots the distribution of class size in the data. As shown in Figure 1, the distribution of class size is smooth. An effective class size cap should lead to a mass point at the cap and no mass afterwards.

Figure 2 plots average class size versus enrollment for grade 10. The red dashed line gives the

[13] We obtained the data from Ministry of Education and Religious Affairs in Greece, and the data is not publicly available.

[14] https://stats.oecd.org/Index.aspx?DataSetCode=EDU_CLASS

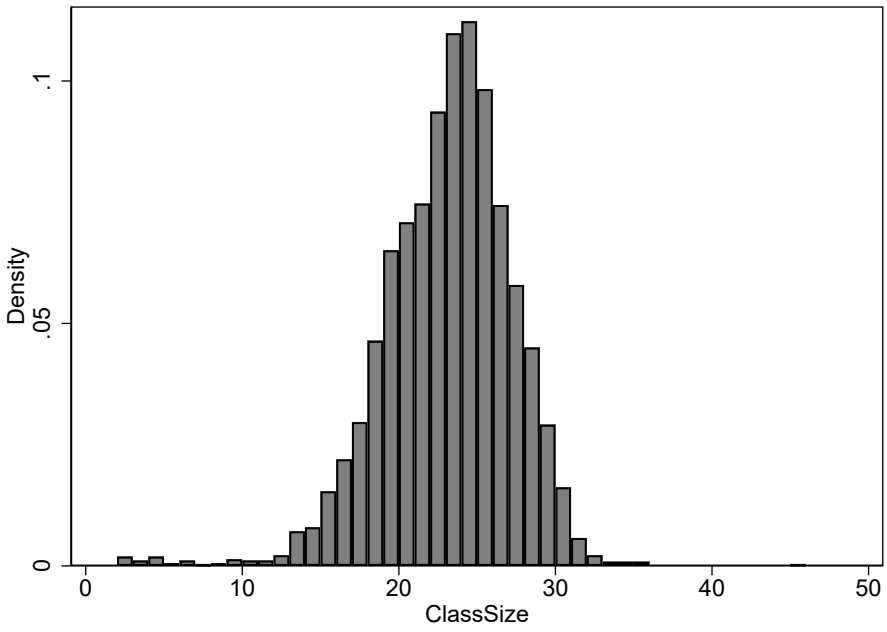


Figure 1: The Distribution of Class Size

predicted class size *had there* been a binding cap on class size of 27. We also plot the green solid fitted line for cohort size and class size for different intervals 0-27, 27-54, ..., 108-135, >135. The class size data loosely follows the red line for low enrollment, but diverges from it when the cohort size becomes larger. What looks like a class size cap is a result of the administrator's choices. For example, if administrators are trying to maximize some increasing function of learning (as measured by GPA) less costs, given student quality, and find it roughly optimal to have a class size close to 27, we might see such a pattern. In Section 5, we show that the targeted class size is close to 27 and that the simulated data from our estimated structural model tracks the actual data on class size reasonably well.

3.2.2 Class Size, Enrollment and GPA

Figure 3 uses the Epanechnikov kernel function to plot a smooth version of the relationship between enrollment and class size given by the solid black line, and between enrollment and GPA given by the dashed red line. The turning points of class size and of GPA seem to be the same so that when one peaks, the other reaches its trough. This relationship becomes much fuzzier for large enrollments.

Figure 2: Class Size versus Enrollment for Grade 10

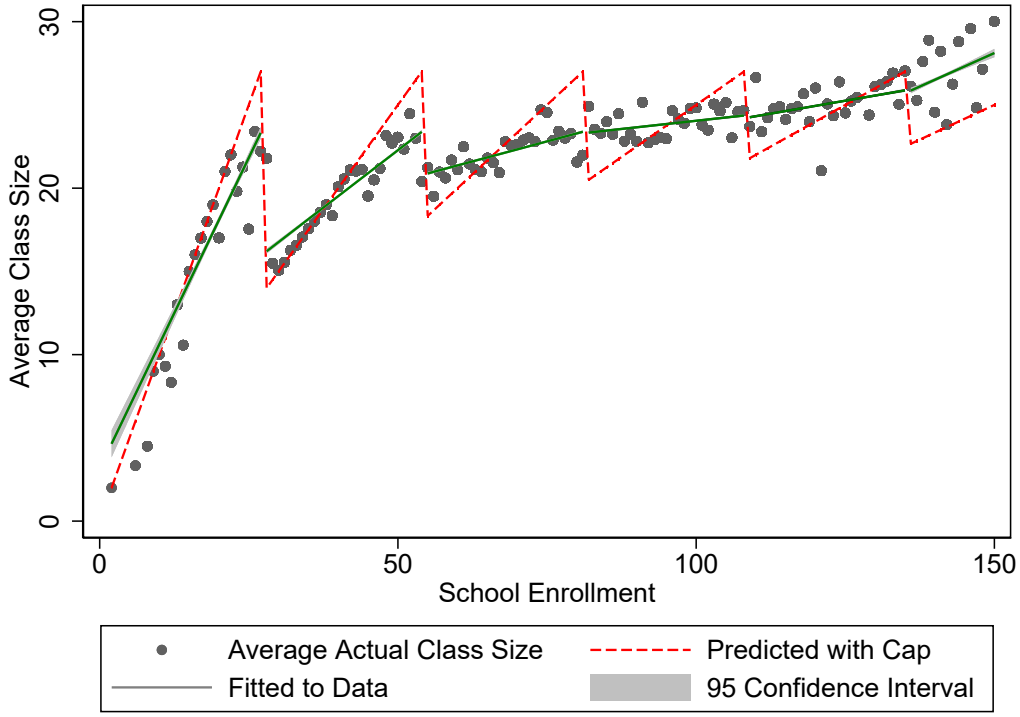
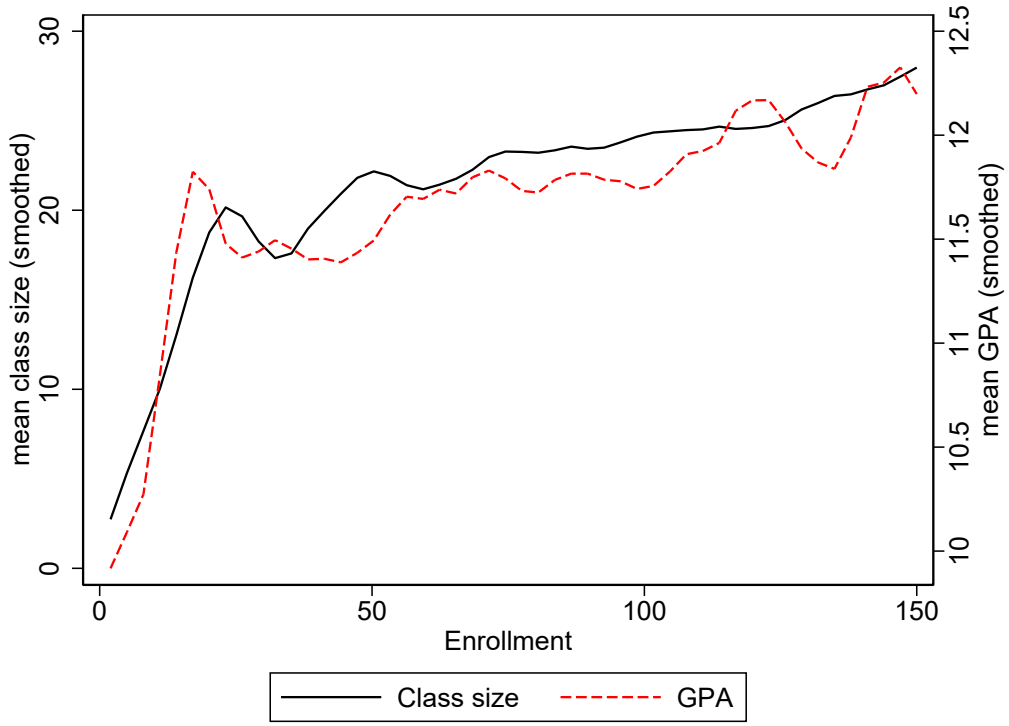


Figure 3: Class Size versus Enrollment for Grade 10 (Smooth Version)



3.2.3 Non Causal Estimates

As a purely descriptive exercise, we next turn to the OLS estimates of the relationship between GPA and class size. We control for school fixed effects so that the estimates are driven by variations over-time within each school. In addition, we include a school-specific linear trend to control for different trends of student characteristics that might affect their academic performance. As is well understood, OLS estimates are likely to be biased and should not be interpreted as causal. Nevertheless, this is the logical starting point for the analysis. Table 2 presents these estimates. Column (1) does not allow for non-linearity and gives a negative and significant coefficient for log of class size. Column (2) adds a quadratic term in class size (ClassSizeSQ). The coefficients now point to a hump shape with a turning point around 18.

Finally, in Appendix D we detrend GPA by regressing it on the variables Female, Age, AgeSQ and the School specific linear trend. We then plot the detrended GPA against class size.[15] We see that the hump shape remains in Figure A.2 in Appendix D.[16]

4 IV Estimates

In view of the results above which suggest an inverse U shape for the effect of class size on GPA, we include a quadratic term in the baseline parametric specification. Our specification is:

$$GPA_{ijt} = \beta CS_{jt} + \gamma \left(CS_{jt}\right)^2 + \alpha_j + \lambda X_{ijt} + \varepsilon_{ijt}.$$

GPA for individual i in school j at time t depends on class size, its square, school fixed effects, and a set of controls, X_{ijt}, which include gender, age, age squared, and the school-specific linear time trends. The school-specific linear trend helps control for different trends of student characteristics that affect their academic performance. Why might class size and GPA be hump shaped? One reason given in the literature, see Borland et al. (2005) and Dobbelsteen et al. (2002), is that students learn from peers like themselves. The larger the class size, the more likely it is that they have peers like themselves. This force makes GPA rise with class size. On the other hand, a larger class size reduces the attention a teacher can give to each student. For low class sizes, the first set of forces dominate but after a point the second does, creating a hump shaped pattern. It has also been argued that a homogeneous class is easier to teach, see Levin (2001) and Dobbelsteen et al.

[15] The size of the bubble represents the mass of data at the point.

[16] We also explore the nonlinear effects with a nonparametric framework provided by Chernozhukov et al. (2013). Their framework is applicable to our setting and does not assume a particular functional form. The approach and results are available in the Online Appendix. It also shows a hump shape relationship between class size and GPA.

(2002).

Since class size could be an endogenous variable, we need an instrument. We cannot use the Angrist and Lavy (1999) approach as there is no maximum class size on the books in Greece in the period that our data covers. The data patterns described in Section 3.2 are consistent with class size being endogenous. From looking at the pattern of enrollment and class size in Figure 2, it seems clear that class size is not allowed to get too large though it is clear there is no class size cap: the actual and predicted class size had there been a cap of 27 are not quite in line, though they are closer together for low enrollment than for high. Instead, we use enrollment as an IV. It is natural to think of overall enrollment as an exogenous shock to class size.[17] In Greece, school enrollment is not a choice as students attend the local school unless they choose to go to private school (which is uncommon) so that we are not worried about parents basing their enrollment on class size. Nor do schools have a say on enrollment as they have to admit all eligible students.[18]

Table 2 gives the IV estimates for grade 10 for the linear and quadratic models in the Column (3) and (4). The standard errors are clustered at the class level. The lower panel of the table gives the relevant estimates for the first stage for convenience. We report the full first stage estimates in the Online Appendix. The upper panel gives the estimates for the second stage.

Recall that the OLS estimate of the coefficient on $ClassSize$ in the linear regression was negative and about -0.02. The coefficient with the IV for the same regression is given in column 1 of Table 2 and is -0.07. Note that this is exactly what one would have expected due to endogeneity bias. If the administrator is choosing class size, classes with better students will tend to be larger as such larger class size has little cost in terms of GPA and OLS is upward biased as in these estimates. Column (4) in Table 2 gives the estimates for the quadratic specification. It clearly has the hump shape expected with a peak at around 19. The bootstrap standard deviation for the turning point is 1.45.[19]

Note that the first stage looks fine: the coefficient on the instrument is positive and significant at the 1% level and the instruments are not weak as the Kleibergen-Paap LM statistic is around 241 in Table 2. It is interesting, and in line with the literature that women have a higher GPA.

While we find strong evidence for a nonmonotonic relationship between class size and achievement, our results are entirely consistent with findings in the literature, see for example Jepsen and Rivkin (2009), that reducing class size is an expensive way of improving achievement. The solid

[17] The approach taken in Hoxby (2000) goes a step further: she fits a quartic to the enrollment data and uses deviations from the quartic as the exogenous variation. We also run the regression using her approach in the Online Appendix and find similar results.

[18] We have school specific linear trends in our model. Had we not put these in, we might have had a problem as enrollment could have been correlated with the error term.

[19] We performed 100 bootstrap replications to estimate the standard deviation.

Table 2: OLS and IV Estimates of Class Size Effects (The Baseline Model)

	(1)	(2)	(3)	(4)
	Dependent Variable: GPA			
	OLS		IV: Second Stage	
ClassSize	-0.021***	0.054	-0.066***	0.37**
	(0.006)	(0.03)	(0.02)	(0.2)
ClassSizeSQ		-0.0016**		-0.0099***
		(0.0007)		(0.004)
Female	0.89***	0.89***	0.90***	0.90***
	(0.03)	(0.03)	(0.03)	(0.03)
Age	-1.78***	-1.78***	-1.78***	-1.79***
	(0.10)	(0.10)	(0.10)	(0.10)
AgeSQ	0.028***	0.028***	0.028***	0.028***
	(0.002)	(0.002)	(0.002)	(0.002)
Kleibergen-Paap Statistic			240.9	45.5
p-value			0.000	0.000
School FE	YES	YES	YES	YES
School-Specific Linear Time Trend	YES	YES	YES	YES
R-sq	0.067	0.068	0.067	0.064
N	81845	81845	81845	81845
			IV: First Stage	
			ClassSize	ClassSize	ClassSizeSQ
Enrollment			0.070***	0.16***	5.56***
			(0.004)	(0.01)	(0.7)
Sq of *Enrollment*				-0.00047***	-0.013***
				(0.00007)	(0.003)

(1) *ClassSizeSQ* is the square of *ClassSize*. Female = 1 if a student is female. Age and AgeSQ control for students' age and its square.

(2) Standard errors are clustered at the class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

line in Figure 4 depicts the quadratic relation we estimate. Note that the value of the intercept is not meaningful as we have school fixed effects and other controls. We choose to center the figure at class size 5 and GPA zero. The curve is relatively flat in the region near the peak by definition. As a result, changing the class size in this region would give small effects. If the curve is not too peaked, this region could be quite large. This might be why even the experimental literature, see Jepsen and Rivkin (2009) for example, found small effects on performance of fairly large changes in class size.

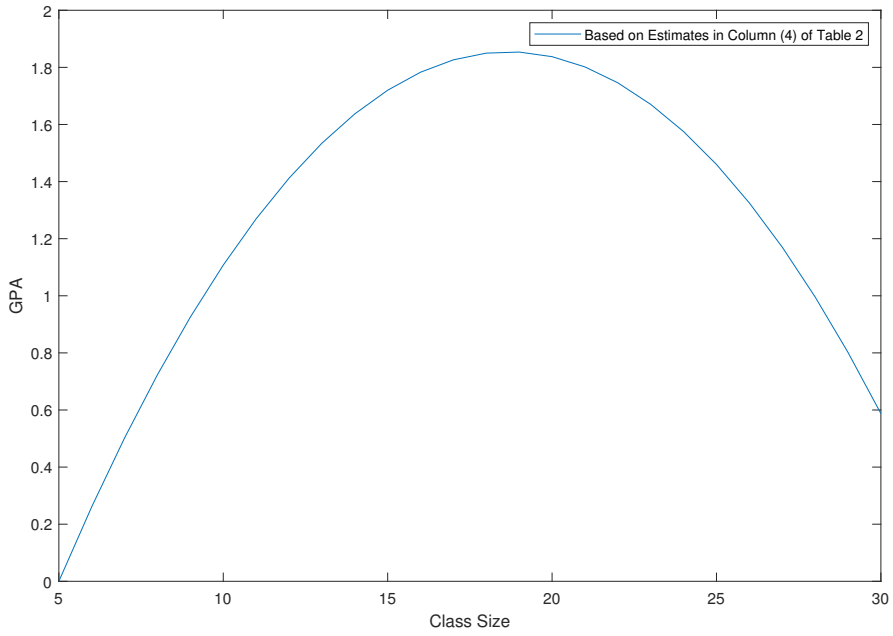


Figure 4: Estimated GPA production function

One concern might be that new teachers with less experience may be hired when the class number rises. As a result, the estimated effect in this table should be interpreted as a net effect, since we are not directly controlling for teacher quality. In the robustness checks, we do however estimate the model with teacher fixed effects for a small subsample and the results are similar.

4.1 Robustness Checks

Our results are robust to various specifications. Here we show that our results are immune to a variety of robustness checks.

Using the Average Class as the Regressor: One might be concerned that better teachers might be assigned more difficult to deal with or in some way different classes. One way to avoid

this problem is to run the regression at the school level so that such effects get averaged out. This is what we do in Columns (1) and (2) in Table 3.

The results are extremely significant and very close to what we have in the baseline model with coefficients and the turning point roughly unchanged.

Teacher Quality: A possible concern might be that we have so far not controlled for teacher quality. For example, are good teachers assigned to larger classes in the hope they can manage them? Such a pattern could bias our estimates, especially in the absence of teacher fixed effects. We examine this in Appendix B. If we think of estimated teacher fixed effects as a proxy for teacher quality, we could ask if there is a correlation between them and class size. We see no such patterns. The correlation between the estimated teacher fixed effects and class size is insignificantly different from zero. Teachers may also have different grading standards. Teacher fixed effects will help take care of this. We were able to obtain and digitize teachers' assignment data for 9 schools and run our baseline regression (subject by subject) including teachers' fixed effects. The results are in Appendix B. The hump shape remains, though the coefficients are not significant. This is not surprising since the sample is small.

Other Specifications: We run a number of alternative specifications. First, we run the regression using log of class size and its square on the RHS using enrollment as the IV, which is reported in Columns (3) and (4) in Table 3. Second, we run the same regressions using Hoxby's instrument which are to be found in the Online Appendix. Our results are robust to these variations with the inverse U shape present everywhere and with the coefficients being highly significant with no sign of weak instruments.

Small Schools: Another concern might be that there are far fewer classes/students in the lower class size bins raising the concern that a quadratic term just adds curvature to a fitted linear regression line with the results being driven by a small number of small schools. The histogram of class size at the class level is given in Figure 1. Roughly 21% of the classes have class size below the turning point of 19. To deal with this issue, we see if our results remain when we divide the sample into large/small schools for each cutoff where the cutoff of cohort size is limited to 30, 50 and 70. The number of small schools in terms of students is .8%, 5.7% and 13% respectively of the total. We show that our results are not driven by small schools. Table A.3 in Appendix C presents the estimation results for large schools where large is defined as having an average cohort size more than 30, 50 and 70 in the three columns respectively. The hump shape remains and all the coefficients are very significant. Moreover, the peak occurs around 18.[20]

Flexible Functional Forms: We also explore more flexible functional forms of class size

[20]The results for small schools are in the Online Appendix. The hump shape remains but the coefficients are no longer significant and the peak occurs around 17.

Table 3: IV Estimates of Class Size Effects with Alternative Specifications

	(1)	(2)	(3)	(4)
		Dependent Variable: GPA		
		Second Stage		
Avg *ClassSize*	-0.063***	0.38**		
	(0.02)	(0.1)		
Sq of Avg *ClassSize*		-0.010***		
		(0.004)		
log(*ClassSize*)			-0.67**	8.68***
			(0.3)	(2.9)
Sq of log(*ClassSize*)				-1.67***
				(0.5)
Female	0.89***	0.89***	0.89***	0.89***
	(0.03)	(0.03)	(0.03)	(0.03)
Age	-1.78***	-1.79***	-1.78***	-1.79***
	(0.10)	(0.10)	(0.10)	(0.10)
AgeSQ	0.028***	0.028***	0.028***	0.028***
	(0.002)	(0.002)	(0.002)	(0.002)
Kleibergen-Paap Statistic	275.6	183.4	179.9	155.1
p-value	0.000	0.000	0.000	0.000
School FE	YES	YES	YES	YES
School-Specific Linear Time Trend	YES	YES	YES	YES
R-sq	0.067	0.066	0.067	0.066
N	81845	81845	81845	81845
		First Stage		
	Avg ClassSize	Avg ClassSize	Sq of Avg ClassSize	log(*ClassSize*)	log(*ClassSize*)	Sq of log(*ClassSize*)
Enrollment	0.073***	0.18***	6.05***			
	(0.004)	(0.01)	(0.6)			
Sq of *Enrollment*		-0.00054***	-0.015***			
		(0.00006)	(0.003)			
log(*Enrollment*)				0.34***	1.38***	5.78***
				(0.02)	(0.1)	(0.7)
Sq of log(*Enrollment*)					-0.13***	-0.47***
					(0.02)	(0.08)

(1) Female = 1 if a student is female. Age and AgeSQ control for students' age and its square.
(2) Standard errors are clustered at the class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

effects. In Appendix D, we run the spline regression of GPA on $ClassSize$ allowing for a more flexible form. There is a clear hump shape and the coefficients are for the most part significant.

General Concerns: One might worry that class size in earlier grades might also have had an effect on achievement in grade 10. Since class size depends on cohort size, one might be concerned that students in large grade 10 classes were also in large classes in all the previous grades, so that the effect of class size might be overstated by being compounded through many years. In Greece, students move to senior school when they enter the 10th grade and senior schools are usually significantly larger and they merge students from several feeder junior schools. To some extent, this alleviates this concern.[21]

Yet another concern might be that GPA may not be a good measure of students' performance because of grade curving so that students are graded relative to their peers. We cannot identify such curving using the data, but curving is not a standard technique in Greece and should not happen in any of the grades. There is nothing about curving grades on the guidelines that are circulated to schools. Even though the 10th grade exam is not standardized across schools, the curriculum covered during the school year is the same across schools (they cover the same books), and schools and teachers follow the same instructions and guidelines about the difficulty of the questions, structure of questions, etc.

A final concern might be one raised in Bach and Sievert (2019). They show that when holding back students is common, large birth cohorts lead to large enrollment and this tends to lead to large classes. However, if enrollment is large, a smaller fraction of students in a grade will be those who were held back at some time. As a result, the average quality of the class will be higher. This will bias the coefficient on class size upwards in the GPA equation. It can be dealt with by adding age as a control. While they show this is the case in Germany, holding back students is very unlikely in Greece.[22] As a result, this is not a concern for us. Moreover, the regressions for the limited sample where we have teacher data also controls for age and age squared which should take care of any such bias anyway. Our results remain so that this concern does not seem an issue with the Greek data.

5 The Structural Model

In the previous sections, we have shown that GPA has an inverse U shape with respect to class size. In this section, we develop a simple stylized structural model that uses our reduced form estimates as an input. The model is kept as simple as possible to highlight the consequences of

[21] Unfortunately, we do not have information for lower grades.

[22] At most 5-7% of students are held back.

various policies. As we cannot distinguish between temporary and permanent teachers in the data we are unable to allow for different adjustment costs for them. We do not allow for differential teacher quality as we have data on teachers for only 9 out of 123 schools.

In our structural model, we ask what an administrator who is trying to do his best for his students but subject to constraints would choose to do. We posit that the administrator is trying to maximize a welfare/objective function that depends on the mean GPA of the students enrolled, as well as the number of students enrolled. Enrollment, e_t, is taken as an exogenous AR1 process and estimated from the data.

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \mu_t \tag{1}$$

where, e_t is assumed to follow a Poisson distribution with mean $\gamma_0 + \gamma_1 e_{t-1}$ with the error term μ_t. We estimate the enrollment process separately for schools of different sizes.

The constraints the administrator faces are of two kinds. To begin with, he faces the trade off we have estimated between class size and GPA. In addition, he faces the enrollment process which is exogenously given to him. We can think of these as technical constraints.[23] He also faces costs associated with the choices he makes. In our model, the only choice the administrator makes is the number of classes, n_t, to have at a point of time. Each additional class has a given cost which can be thought of as the cost of the teachers needed for the additional class. Since teachers unions are prevalent in Greece, firing teachers is costly. Moreover, finding a new teacher also involves a number of costs including sending a vacancy request to the Ministry of Education, advertising the position, and so on. The empirical transition probabilities across the number of classes are in Table 4. Note that schools tend to keep the same number of classes across years. This is especially so for schools with a small number of classes.

For these reasons we allow for hiring and firing costs in the model. This makes the problem dynamic. At any point of time, the administrator is forward looking when choosing the number of teachers/classes. Given the enrollment today, the enrollment process the school faces, as well as the marginal and adjustment costs it faces, the school makes its decision. We do not impose any assumptions on the size of hiring and firing costs, but let the data pin them down.

[23] The trade-off between GPA and class size the administrator faces is analogous to the production function a manager choosing inputs would face. The enrollment process (e_t) can be thought of as similar to an exogenous Total Factor Productivity (TFP) process.

Table 4: Transition Matrix of the Number of Classes

n_{t-1} \ n_t	1	2	3	4	5	6	7
1	0.7143	0.2727	0.0130	0.0000	0.0000	0.0000	0.0000
2	0.0813	0.6986	0.2010	0.0144	0.0048	0.0000	0.0000
3	0.0034	0.1399	0.6007	0.2389	0.0171	0.0000	0.0000
4	0.0000	0.0036	0.2456	0.5979	0.1459	0.0071	0.0000
5	0.0000	0.0000	0.0405	0.2162	0.6622	0.0743	0.0068
6	0.0000	0.0000	0.0000	0.0645	0.3871	0.4194	0.1290
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.7500	0.2500

The administrator cares about the mean GPA. As this has a quadratic form, and as $\frac{e}{n}$ is the average class size, we have

$$GPA_t = a\frac{e_t}{n_t} + b\left(\frac{e_t}{n_t}\right)^2 + A.$$

A is the value of the other variables in the regression at their mean levels. It is worth noting that its value will not affect the choice of the number of classes below.

We assume that having twice the students with the same GPA gives the administrator twice the utility. This makes sense as the object is to educate students and educating twice as many to the same level gives twice the utility. Thus, we have the administrator's utility as

$$e_t\left[a\frac{e_t}{n_t} + b\left(\frac{e_t}{n_t}\right)^2 + A\right].$$

The administrator faces hiring and firing costs of H and F, and a variable cost per class of c which we interpret as the salary of the additional teacher(s) needed for one more class. The administrator knows the realization of e_t and knows the state variable, n_{t-1}, and the vector of random utility shocks associated with the utility of each number of classes, $\boldsymbol{\varepsilon}_t = \{\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}...\varepsilon_{10t}\}$, when the school makes its choices.[24] This shock is not observed by the econometrician. Each element of $\boldsymbol{\varepsilon}_t$ is drawn from a type 1 generalized extreme value distribution. For example, an administrator who has one class in the 10th grade may find that a room has opened up. This

[24]One might worry that the decision of number of classes is made for the 10th, 11th and 12th grade jointly by the administrator. For example, if there is a surge of enrollment in the 11th grade, the number of classes in the 10th and 12th grade might fall. While we cannot rule this out completely, we show that when we regress the change in the number of classes on the change in enrollment, that variation in enrollment in grade 10 has a significantly larger impact on class number than does variation in enrollment across grades. The result is presented in Table 1 on the Online Appendix.

can be interpreted as a high realization of ε_{2t} which makes it more likely the school will open an additional class. $\boldsymbol{\varepsilon}_t$ helps fit the model's predictions to data since, given enrollment and the number of classes, there is considerable randomness in terms of whether the number of classes is changed. Since no school has more than 10 classes, we restrict the size of the vector to be 10. This assumption allows us to use the logit setup and fit the data parsimoniously. In some periods we may see a larger class size, i.e., fewer classes, than in others. The reason for this comes partly from enrollment declines, partly because fewer classes were present in the past and there are hiring costs, and partly from the shock.[25]

Thus, the administrators value function is:

$$\begin{aligned} V(e_t, n_{t-1}, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}) &= \underset{n_t}{Max} \Bigg\{ e_t \left[a\frac{e_t}{n_t} + b\left(\frac{e_t}{n_t}\right)^2 + A \right] - cn_t \\ &\quad -H \cdot \max(n_t - n_{t-1}, 0) - F \cdot \max(n_{t-1} - n_t, 0) + \varepsilon_{n_t t} \\ &\quad + \delta \mathbb{E}_{\boldsymbol{\varepsilon}_{t+1}, e_{t+1}} V(e_{t+1}, n_t, \boldsymbol{\varepsilon}_{t+1}; \boldsymbol{\theta}) \Bigg\} \\ \text{where } e_{t+1} &= \gamma_0 + \gamma_1 e_t + \mu_{t+1} \text{ and where } \boldsymbol{\theta} = (c, H, F, \sigma) \end{aligned}$$

where c is the marginal cost of a class, H and F are the hiring and firing costs respectively and σ^2 denotes the variance of the utility shock.[26] We set the discount factor $\delta = 0.95$. The parameters a and b are taken from Column (4) of Table 2. Note that the expectation is taken over both $\boldsymbol{\varepsilon}_{t+1}$ and e_{t+1}, the shock to utility and the shock to enrollment, respectively. Note that though Ae_t enters the objective function, it will not affect the optimal choice of n_t as it is exogenous. In what follows conditioning on the set of parameters is implied but omitted for clarity of exposition.

Rewriting this slightly for notational ease we define $u(e_t, n_t, n_{t-1})$ as the deterministic component of current period contribution to the objective function and $\overline{V}(e_t, n_{t-1})$ as the ex ante value function, i.e., the value of behaving optimally from tomorrow onwards before knowing the realization of the utility shock.

[25] We are not modeling the budget constraints of schools. Since schools are smoothing their budgets over years, the yearly budget is not strictly binding and the administrator can be thought of as maximizing attainment less the costs of obtaining it as we do.

[26] It is well understood that there is a degree of freedom in terms of what we can estimate: the weight on GPA and the variance of the utility shock cannot be separately identified. See Greene (2009) for a discussion. We choose to set the weight on GPA at unity and estimate the variance of the utility shock as it is easier to interpret.

$$
\begin{aligned}
u(e_t, n_t, n_{t-1}) &= e_t \left[a\frac{e_t}{n_t} + b\left(\frac{e_t}{n_t}\right)^2 + A \right] - cn_t \\
&\quad -H\max(n_t - n_{t-1}, 0) - F\max(n_{t-1} - n_t, 0) \\
\overline{V}(e_{t+1}, n_t) &= \mathbb{E}_{\varepsilon_{t+1}}[V(e_{t+1}, n_t, \boldsymbol{\varepsilon}_{t+1})] \\
v(e_t, n_t, n_{t-1}) &= u(e_t, n_t, n_{t-1}) + \delta \mathbb{E}_{\mu_{t+1}}[\overline{V}(e_{t+1}, n_t)|e_t] \qquad (2)
\end{aligned}
$$

so that

$$V(e_t, n_{t-1}, \varepsilon_t) = \max_{n_t} v(e_t, n_t, n_{t-1}) + \varepsilon_{n_t t}.$$

Thus we have rewritten the value function as a base utility and a shock. Since $\varepsilon_{n_t t}$ follows an independent and identically distributed (iid) type 1 generalized extreme value distribution with variance σ^2, the probability of observing n_t is

$$p(n_t|n_{t-1}, e_t; \boldsymbol{\theta}) = \frac{exp(v(e_t, n_t, n_{t-1}; \boldsymbol{\theta})/\sigma)}{\sum_{n=1}^{10} exp(v(e_t, n, n_{t-1}; \boldsymbol{\theta})/\sigma)}.$$

5.1 Identification and Estimation

We first provide some intuition behind what pins down $\boldsymbol{\theta}$ before we turn to the estimation part. The problem is modeled as a dynamic discrete choice problem. We bring the estimates of the quadratic model for achievement from the reduced form regressions to the structural model. As estimated in the parametric quadratic model, $b = -0.0099$ and $a = 0.37$. It remains to estimate $\boldsymbol{\theta} = (c, H, F, \sigma)$. How can we identify $\boldsymbol{\theta}$? One way to get some intuition about which features of the data would help identify which parameters is to ask how a simulation based approach might pin down the parameters. We do not use this approach, but nevertheless, this is a useful exercise.

To see how the optimization works, it is useful to think of the problem in a slightly different way where we first define the pre-value function as $W(e_t, n_t, \varepsilon_{n_t t})$. $W(.)$ is the value of the flow utility today (excluding the adjustment costs) and behaving optimally from tomorrow onwards for every value of n_t chosen today.

$$
\begin{aligned}
W(e_t, n_t, \varepsilon_t) &= e_t \left[a\frac{e_t}{n_t} + b\left(\frac{e_t}{n_t}\right)^2 \right] - cn_t + \varepsilon_{n_t t} \\
&\quad +\delta \mathbb{E}_{\boldsymbol{\varepsilon}_{t+1}, \mu_{t+1}} V(e_{t+1}, n_t, \boldsymbol{\varepsilon}_{t+1}) \qquad (3)
\end{aligned}
$$

Then,

$$V(e_t, n_{t-1}, \varepsilon_t) = \max_{n_t} \{W(e_t, n_t, \boldsymbol{\varepsilon}_t) - H \cdot \max(n_t - n_{t-1}, 0) - F \cdot \max(n_{t-1} - n_t, 0)\}$$

To begin with, let us see how the model works when we take n to be continuous, the pre-value function to be concave, and set the utility shocks to zero. In this case, the current period problem can be depicted as in Figure 5 where $W(e_t, n_t, \boldsymbol{\varepsilon}_t = 0)$ is depicted by the concave curve. Consider such a school with a given enrollment as well as utility shocks set at zero. Anchor the linear adjustment costs to n_{t-1} as depicted. The cost of increasing the number of classes has slope H and decreasing it has slope F while making no change in their number has no cost. The optimal choice of n_t is that which maximizes the difference in the pre-value function and these adjustment costs. Let n^L be where the slope of the pre-value function is H and n^H be where the slope is $-F$. It is obvious from the picture that if n_{t-1} exceeds n^H, it is optimal to reduce n_t to n^H, i.e., increase class size, and if n_{t-1} falls short of n^L, to raise n_{t-1} to n^L, that is reduce class size. If n_{t-1} lies in the interval $\left[n^L, n^H\right]$ it is optimal to keep $n_t = n_{t-1}$. This region of inaction is created by these adjustment costs which are not differentiable at 0. The higher the adjustment costs, the larger this region of inaction.[27] The size of H and F will be pinned down by the bounds of the region of inaction, $\left[n^L, n^H\right]$ which would be observed in the data.[28]

When we add back the utility shocks and the discreteness of n, the stark predictions of the restricted model above are tempered. The depiction in Figure 5 changes so that the pre-value function is discretized. At each of the ten values taken by n, there is a base value plus the shock. As a result, the curve connecting the grid points of the analogue of the pre-value function need not be concave. However, the optima choice will still be such that given n_{t-1}, the difference in the pre-value and adjustment costs is maximized. And an increase in the adjustment costs would increase the region of inaction and affect the probability of transitioning into this region. In this way, the empirical transition probabilities help pin down H and F in the data.

These same empirical transition probabilities also help pin down the variance of the shock. Given the enrollment process, when the variance of the utility shock rises, the probability of moving from one class size to another also rises.

How is the final parameter, c, pinned down? As c rises, having more classes becomes more

[27] It is worth noting that a change in H or F will also shift the pre-value function as it will change the continuation value. However, this effect will be second order relative to the direct effect of H and F.

[28] It can be shown that the effect of an increase in hiring costs will be greater for n^L, the hiring cutoff, than for n^H, the firing one. Similarly, an increase in F will have a greater effect for n^H than for n^L.Thus, if we think of the combinations of H and F that are consistent with a given value for the hiring cutoff as well as those consistent with the firing cutoff we will get a unique value of H and F which are consistent with both. As a result, there is a unique H and F that correspond to give values of $\left[n^L, n^H\right]$ which would be observed from the data.

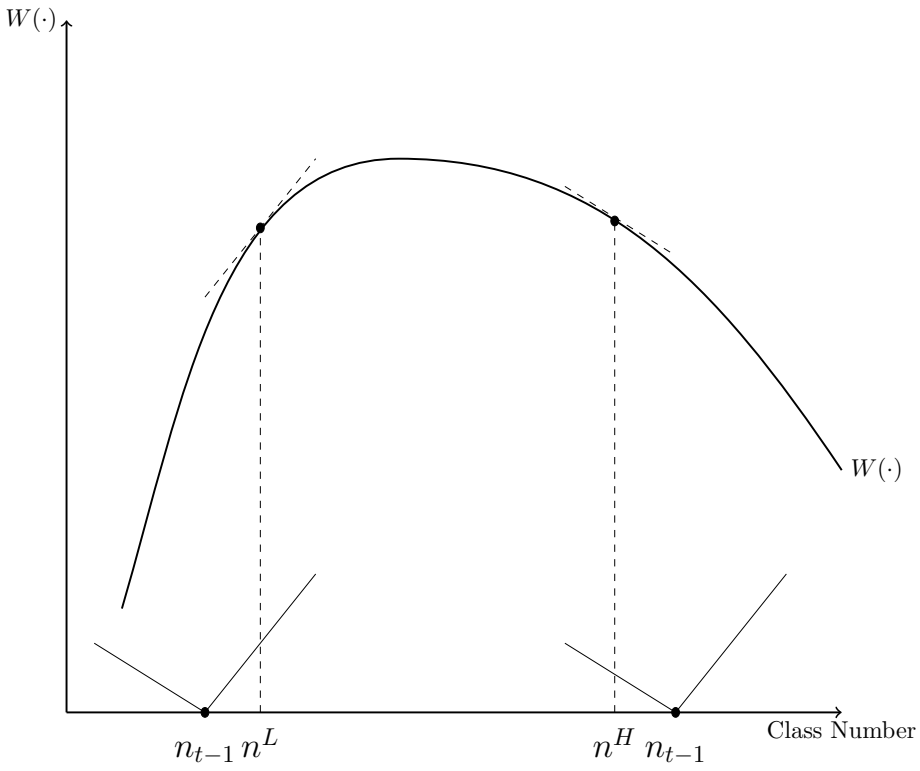


Figure 5: Identification of Adjustment Cost H, F

expensive and the number of classes falls. Thus c is pinned down by the average number of classes given enrollment, or equivalently, the average class size.

Having sketched out the intuition behind identification, we move on to the details of estimation. The estimation can be thought of as proceeding in two steps. First, we estimate the process for enrollment. Then, we estimate $\boldsymbol{\theta}$.[29]

5.2 Estimating Enrollment

The enrollment process is given by Equation (1). If there was no random component, μ_t, and we applied a common enrollment process to all schools, then this enrollment process results in the data generated by it being on the straight line with slope less than 1 depicted in Figure 6. This would result in a steady state at point A in Figure 6. This means that all schools would have the same enrollment in steady state. Adding a random component will make the process generate data that falls in a band around the straight line in Figure 6. The width of this band depends on the variance of μ_t. This will give a distribution of steady states in Figure 6. This distribution will

[29]The structural estimation imposes a stylized structural model on schools' behavior. We drop the school fixed effects we had in the reduced form regression to achieve tractability.

be more concentrated around A as the variance of μ_t falls. Note that in this case, schools will *not* tend to stay in their own rough groups over time.

Alternatively, suppose we run the AR1 process separately for the three quantiles. Then, in the absence of a random component, we would get three lines, one above the other, for the three groups and get three different steady state points as in Figure 6. Adding back randomness would create bands around the lines as before and create a distribution of steady states for each group size. If these intervals overlapped, as they would if the variance of μ_t was large, there could be some movement between groups in steady state. For low variance of μ_t, schools would tend to stay in their own group in terms of size.

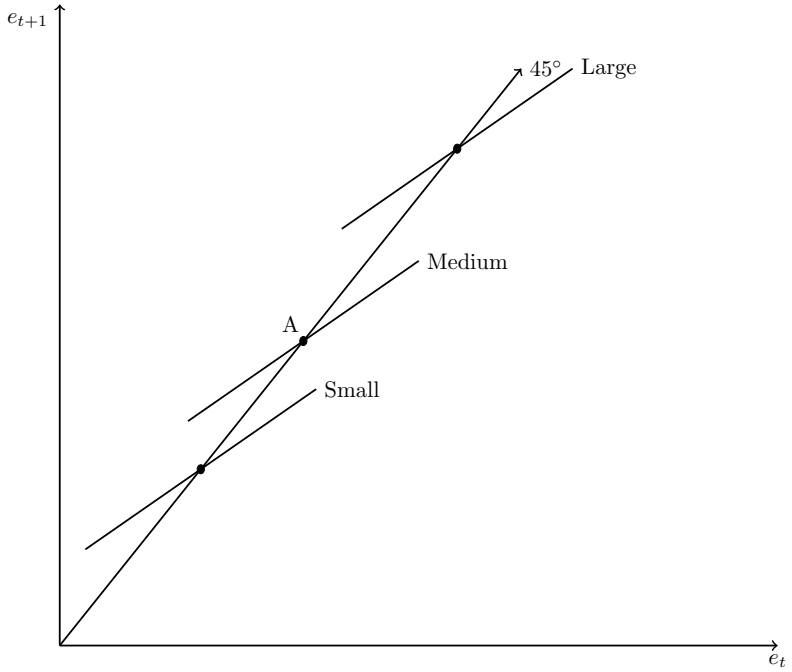


Figure 6: Enrollment Process

We do not want to estimate a single enrollment process for the entire data as in the data, schools tend to stay in the same rough size groups, though their enrollment fluctuates year by year. We do not want to estimate the enrollment process at the school level as this would give us at most 10 data points to work with. We group schools by their mean enrollment and estimate an enrollment process for each group.[30] We break the data into K quantiles with schools in the

[30]We do not estimate the enrollment process school by school for two reasons. First, we have at most 12 years of data for a school. Second, estimating the enrollment process school by schools makes the estimation of the structural model very computationally intensive as the value functions and decision rules for each school have to be calculated separately. Moreover, the identification comes from the choices made, and particularly from the changes in the number of classes. With at most 12 years of data, there are few instances of a change in class number. For these reasons we choose to group schools together.

lowest quantile in group 1 and the highest quantile in group K. This gives us the cutoffs for mean enrollment for each group. How might one choose K? In our counterfactuals below we look at the effect of various policies on three groups of schools: small, medium and large. For this reason, we would like K to be a multiple of 3. We would also like to choose K to match the transitions between deciles from t to $t-1$. To do so, we construct the transition matrices across deciles in the data and in the simulated enrollment when we estimate the enrollment process for $K = 1, 3, 6, 9, 12$ groups. The estimated enrollment processes for different choices of K are to be found in the Online Appendix. We choose $K = 6$ as our baseline.[31] The more disaggregated enrollment process, the better the simulations match the data. However, a more disaggregated enrollment process also leads to more computational difficulties in the structural estimation as value functions differ by enrollment process. Our choice of $K = 6$ balances these two forces.[32]

The estimates for the enrollment process from the actual data for each of the 6 groups are presented in Table 5.

Table 5: Estimates of Enrollment Process

Enrollment Quantile	0-17	18-33	34-50	51-67	68-83	84-100
γ_1	0.58	0.54	0.36	0.38	0.31	0.64
sd	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)
γ_0	12.14	24.25	43.65	50.12	65.90	44.55
sd	(0.70)	(1.53)	(1.70)	(2.52)	(2.04)	(3.87)
N	167	177	173	182	184	179

(1) The standard errors are presented in parentheses.

5.3 Estimation of $\boldsymbol{\theta}$

Recall that since $\boldsymbol{\varepsilon}_t$ is assumed to have an iid type I generalized extreme value distribution, we know that equation (4) holds.

$$p(n_t|n_{t-1}, e_t, \boldsymbol{\theta}) = \frac{exp(v(e_t, n_t, n_{t-1}))}{\sum_{n=1}^{10} exp(v(e_t, n, n_{t-1}))}. \tag{4}$$

[31] The Online Appendix presents the simulated transition matrix between groups from the first period to the last period with different groupings.

[32] The estimates when we use 1, 3, 9 and 12 groups of roughly equal size are presented in the Online Appendix. With more groups, the estimated slopes are slightly flatter. The estimates of the structural parameters corresponding to these estimates of the AR1 process are presented in the Online Appendix. Note that there is not much difference in them.

$$\begin{aligned}\overline{V}(e_t, n_{t-1}) &= \mathbb{E}_{\boldsymbol{\varepsilon}_t} \max_{n_t} \left[v(e_t, n_{t,} n_{t-1}) + \varepsilon_{n_t t}\right] \\ &= \sum_{n_t=1}^{10} p(n_t | n_{t-1}, e_t, \boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\varepsilon}_t} \left(v(e_t, n_{t,} n_{t-1}) + \boldsymbol{\varepsilon}_{n_t t} | n_t \; is \; optimal\right)\end{aligned}$$

since $\mathbb{E}_{\boldsymbol{\varepsilon}_t} \max_n f(n, \varepsilon_t) = \sum_n p(n) \mathbf{E}_{\boldsymbol{\varepsilon}_t}[f(n, \boldsymbol{\varepsilon}_t) | n \text{ being the maximum}]$ where $p(n)$ is the probability that n is the maximum at a particular value. In other words, the ex-ante value function is just the probability of n being the optimal choice (given enrollment today and the number of classes inherited) times the payoff from then on.

Using the form of the distribution of $\boldsymbol{\varepsilon}_t$ and some calculations yields

$$\overline{V}(e_t, n_{t-1}) = \ln \left(\sum_{n_t=1}^{10} exp \left[u(e_t, n_t, n_{t-1}) + \delta \mathbb{E}_{e_{t+1}} [\overline{V}(e_{t+1}, n_t) | e_t]\right] \right) + \gamma$$

where γ is Euler's constant.

By value function iteration, we can solve $\overline{V}(e_t, n_{t-1})$, and thus $v(e_t, n_{t-1}, n_t)$. This is essentially finding a fixed point of a function. By taking a grid and guessing values of the function $\overline{V}(e_{t+1}, n_t)$ over the grid, this reduces the problem to a finite dimensional one. e is allowed to take values from 1 to 1000 since the largest school in the data has far less than 1000 students. This guess, together with the estimated process for enrollment gives a numerical value of $\mathbb{E}_{\mu_{t+1}}[\overline{V}(e_{t+1}, n_t) | e_t]$ over the grid. For given parameter values, we can calculate $u(e_t, n_t, n_{t-1})$ so that we get a numerical value for the RHS over the grid which is the new guess. We stop when the guess and the new guess are close enough, i.e., when we have a fixed point. Since $\delta < 1$ and the enrollment process is stable, i.e., $\gamma_1 < 1$, this is a contraction mapping and this process converges to the fixed point. Having solved for $\overline{V}(e_t, n_{t-1})$ we use equation (2) to solve for $v(e_t, n_t, n_{t-1})$, which in turn gives the value for $p(n_t | n_{t-1}, e_t; \boldsymbol{\theta})$. Finally, we choose $\boldsymbol{\theta}$ to maximize the likelihood of the empirical transition probabilities

$$L = \Pi_j \; p(n_{jt} | n_{jt-1}, e_{jt}; \boldsymbol{\theta})$$

to get the estimated $\boldsymbol{\theta}$.

The estimates are presented in Table 6. A larger variance indicates that idiosyncratic shocks matter more when schools choose the number of classes. Idiosyncratic shocks could be the availability of spaces and teachers. The variable cost of adding a class is given by c. The fixed cost of adding a class is H while the fixed cost of subtracting a class is F. Note that the hiring cost is insignificantly different from zero, while the firing cost is substantial and significantly different

Table 6: Estimates of the Structural Dynamic Model

	c	H	F	σ
All	87.90	250.09	232.58	213.05
sd	(56.32)	(67.19)	(63.55)	(5.39)
Euro	€20,572	€58,531	€54,435	

(1) The standard errors are presented in parentheses.

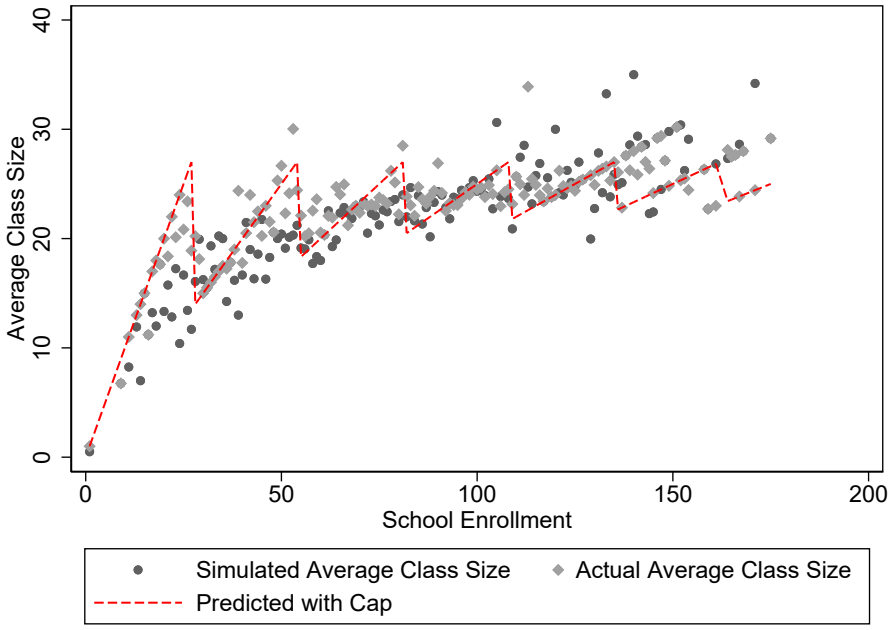
from zero. Recall that there is an excess supply of temporary teachers in the period of our data, which might drive the low hiring cost estimated. The marginal cost is roughly twice of the firing cost.

We can transform these numbers into Euros as follows. Suppose that the cost of an additional class is one teacher's salary in Greece. The salary after 15 years' experience with minimum training for a high school teacher is about €20,572 in 2004 (OECD, 2005). H and F are the adjustment costs per class. To get the adjustment cost in Euros, we divide F by c and then multiply by the Euro cost of an additional class. These Euro cost estimates are given in the lower part of the entry in Table 6. It costs €54,435 to drop a class and €58,531 to add a class. The cost of dropping a class is high. This is reasonable in Greece as firing a teacher is hard for public schools. The hiring cost is also high as a lot of bureaucracy is involved in hiring a teacher as well. The class size chosen in the absence of any adjustment costs is 25-26.

One might be concerned that the choice of K makes a big difference in terms of the structural parameters estimated. It turns out that it has little effect on these and our basic patterns persist. In the Online Appendix, we explore the impact of different groupings when estimating the enrollment process on our structural parameter estimates of interest. It turns out that these estimates are not very sensitive to the choice of enrollment process used. Marginal cost estimates with different grouping range from 83 to 90, hiring costs estimates range from 246-256, and firing cost estimates range from 235 to 247. Estimates of σ vary from 200 to 219. All of these are not significantly different from each other.

We simulate the data using the estimated model given the enrollment and class number in the previous period in the data. We compare the class size pattern of the simulated data to the actual data. Figure 7 presents both the simulated and actual class size patterns. First, the result shows that the pattern of the simulated class size matches the actual data well. Second, we use the simulated data to show that even without the class cap, average class size is close to 22.

Figure 7: The Simulated vs Actual Class Size Pattern



5.4 Counterfactual Exercises

We use the estimated processes for enrollment for each size school to simulate the model. We simulate 1000 schools for each school size. For each simulated school, we simulate 100 periods. We only use the last 10 periods as the data is by then invariant to choice of a starting point. We use the simulated data to study class size and GPA changes for three different school groups. These groups are formed by putting the lowest two quantiles in group 1 and the highest two quantiles in group 3. We can think of these as small schools with one or two classes per grade, medium schools with two classes or three classes a grade, and large schools with four or more classes per grade. We choose to do this to allow schools of different sizes to be differentially impacted by any of our counterfactual polices.[33] It is thought that larger schools tend to have more flexibility in adjusting to enrollment shocks than smaller ones and our results tend to confirm this belief.

In Greece, as in many countries, teachers are unionized and as a result, firing a teacher is quite costly. The first counterfactual exercise we consider is the effect of reducing firing cost to zero. How would this affect the class size and GPA? On the one hand, firing teachers will be easy which will raise class size relative to the status quo. This is the direct effect. On the other hand, since its easy to fire teachers, it is more likely they will be hired, which reduces class size. This is the indirect effect. Ex ante, the net effect is not obvious. The results of this counterfactual are

[33]It is worth noting that the parameter estimates and counterfactual as shown in Table A.5 and Table A.6 in Appendix E would be very different if we use the linear GPA function of class size.

presented in Table 7. Reducing firing cost to zero raises average class size by 5-6 students and reduces GPA by about 0.76-1.18 point (recall the scale was from 1-20). Since class size is larger, fewer teachers are hired, and thus, variable cost is lower. The total costs, variable and adjustment, fall by €13,370 for small schools, €17,797 for medium size schools and €21,939 for large schools on a per school per year basis. Finally we calculate the welfare change, i.e. the change in the objective function. The total welfare increases by €283,729 for small schools, €321,517 and €352,778 for large schools.

If the government was able to reduce hiring costs, it could potentially reduce costs and improve welfare. The next counterfactual looks at the case where the hiring cost is set at 0. These results are presented in Table 7. A zero hiring cost reduces class size by 4-5 students as more teachers are hired and increases GPA by 0.25-0.33. The total cost goes up as schools hire more teachers (so that variable costs increase). As shown in the last column, total welfare, which includes the flow utility, adjustment costs and the continuation value, fall for each size school by 119 to 151 thousand euros.

The next counterfactual looks at the case where the variable cost increases by 50%, i.e., the teachers' salary is raised by 50%. Class size rises by about 2 students and GPA falls by 0.21-0.32 points. The total cost goes up by 12 to 35 thousand euros and welfare decreases considerably: by €394,512 for small schools and €928,099 for large schools.[34] Note that costs overall rise by less than 50% as there are adjustments on the hiring and firing margins. It is well understood that teacher quality has a large impact on students performance. Higher salary attracts better teachers. Our calculations do not include any improvement in teacher quality due to higher wages paid and so are likely to over estimate the welfare losses of this policy.

The next counterfactual is to look at the effects of a class size cap at 25, 30 and 35 on class size. Class caps cause schools to add a class well before the cap is reached when the enrollment is more volatile and adjustment costs are large. The effects of such caps are stronger for small schools since they have smaller margins to adjust. As a result, such caps will impact small schools more. Consider a class size cap of 25. For small schools, this reduces class size dramatically by 10 students, while large schools have class size falling by 8 students. Welfare falls by €610,474 for small schools and €800,877 for large ones while costs rise by €27,949 for small schools and €39,960 for large ones. Even when a class cap of 35, which is well above 27 which looks like the targeted class size found in the data, has a considerable impact, especially for small schools. Class size falls by 7-8 for small schools and 2 for large ones. The literature has found almost uniformly that changing class size tends to be a costly way of raising academic achievement. We also find

[34] Cost increases in this counterfactual are per year, while welfare increases are the present discounted value over time. This is why welfare decreases are so much larger in magnitude here.

this. In addition, we find that even caps which seem non binding have very significant impacts, especially for small schools.

Table 7: Counterfactuals

School Size	Δ Average Class Size (#)	Δ Average GPA (points)	Δ Cost (€)	Δ Welfare (€)
		$F' = 0$		
Small	4.58	-0.76	-13370.48	283729.17
Medium	6.19	-1.18	-17797.30	321517.61
Large	5.88	-1.14	-21939.85	352778.34
		$H' = 0$		
Small	-5.29	0.25	15928.30	118889.64
Medium	-4.12	0.31	16209.60	137061.84
Large	-3.83	0.33	19571.97	150951.82
		$c' = 1.5c$		
Small	2.02	-0.21	11842.24	-394512.26
Medium	2.12	-0.31	24026.09	-655661.78
Large	2.14	-0.32	35397.92	-928099.70
		Class Cap = 25		
Small	-10.29	0.29	27949.46	-610474.12
Medium	-8.18	0.41	27548.14	-651820.43
Large	-7.99	0.44	39960.00	-800877.55
		Class Cap = 30		
Small	-7.94	0.45	13128.12	-318849.03
Medium	-6.10	0.45	15836.46	-381950.90
Large	-5.33	0.45	20186.85	-385247.57
		Class Cap = 35		
Small	-7.66	0.46	12251.13	-291750.19
Medium	-3.73	0.39	7522.60	-180494.71
Large	-2.20	0.28	6079.26	-116188.29

(1) These groups are formed by putting the lowest two quantiles in group 1 and the highest two quantiles in group 3.

(2) Cost increases in this counterfactual are per year, while welfare increases are the present discounted value over time. This is why welfare decreases are so much larger in magnitude here.

6 Conclusions

Our work shows a clear hump shaped relationship between class size and GPA. We speculate that the mixed results prevalent in the literature on the relationship between class size and achievement is due to the focus on a monotone specification.

Our estimates also help explain why changes in class size in practice did not have a large effect on student achievement. See Jepsen and Rivkin (2009) who find small effects of a reduction of class size from 30 to 20 for students in kindergarden to third grade. This could come from the relationship between class size and GPA being hump shaped and from moving from one size of the hump to the other or from the slope being small in absolute terms. Of course, the shape of this relationship could vary by subject and grade. If the relationship had more curvature, then class size might be a far less costly way of improving achievement than previously thought, but there is little work on this in the literature.

Our structural estimates and the simulations allow us to closely match the data as shown in Figure 7. Our results suggest that reducing firing costs actually hurts achievement so that teachers' unions may not be as pernicious as might be thought. Raising hiring cost from the low initial level increases class size and thus, worsens students' achievement. Class size caps have large effects even when they are set above average levels, and their effects are more pronounced for small schools. A class size cap forces schools to add a class before they would want to do so in order to not cross the cap if enrollment surges.

A channel we could not fully explore and is potentially more important, is the effect of teacher quality on achievement and how this varies by the ability of the students. Does having a good teacher in a core subject like Mathematics have spillover effects on performance in other subjects like Physics? We know from past work (see Chetty et al. 2014) that the effect of teacher quality on achievement is large. Further work that controls for both student ability and teacher ability and spillovers across subjects taken to better understand the impact of better teachers on students of different abilities would be valuable.

References

Angrist, Joshua D. and Victor Lavy, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 1999, *114* (2), 533–575.

Bach, Maximilian and Stephan Sievert, "Birth Cohort Size Variation and the Estimation of Class Size Effects," 2019.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul, "Heterogeneous class size effects: New evidence from a panel of university students," *The Economic Journal*, 2010, *120* (549), 1365–1398.

Bingley, Paul, Vibeke Jensen, and Ian Walker, "The effect of school class size on post-compulsory education: Some cost benefit analysis," 2007.

Bonesrønning, Hans, "Class size effects on student achievement in Norway: Patterns and explanations," *Southern Economic Journal*, 2003, *69* (4), 952–965.

Borland, Melvin V., Roy M. Howsen, and Michelle W. Trawick, "An investigation of the effect of class size on student academic achievement," *Education Economics*, 2005, *13* (1), 73–83.

Browning, Martin and Eskil Heinesen, "Class size, teacher hours and educational attainment," *The Scandinavian Journal of Economics*, 2007, *109* (2), 415–438.

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey, "Average and quantile effects in nonseparable panel models," *Econometrica*, 2013, *81* (2), 535–580.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Dinerstein, Michael, Rigissa Megalokonomou, and Constantine Yannelis, "Human Capital Depreciation," Technical Report, National Bureau of Economic Research 2020.

Dobbelsteen, Simone, Jesse Levin, and Hessel Oosterbeek, "The causal effect of class size on scholastic achievement: Distinguishing the pure class size effect from the effect of changes in class composition," *Oxford Bulletin of Economics and Statistics*, 2002, *64* (1), 17–38.

Gary-Bobo, Robert J. and Mohamed-Badrane Mahjoub, "Estimation of class-Size effects, using "Maimonides' Rule" and other instruments: The case of French junior high schools," *Annals of Economics and Statistics*, 2013, (111/112), 193–225.

Greene, William, "Discrete choice modeling," in "Palgrave handbook of econometrics," Springer, 2009, pp. 473–556.

Hanushek, Eric A., "Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects," *Educational Evaluation and Policy Analysis*, 1999, *21* (2), 143–163.

_ , "The failure of input-based schooling policies," *The Economic Journal*, 2003, *113* (485), 64–98.

Hojo, Masakazu, "Class-size effects in Japanese schools: A spline regression approach," *Economics Letters*, 2013, *120* (3), 583–587.

Hoxby, Caroline M., "The effects of class size on student achievement: New evidence from population variation," *The Quarterly Journal of Economics*, 2000, *115* (4), 1239–1285.

Jepsen, Christopher and Steven Rivkin, "Class size reduction and student achievement: The potential tradeoff between teacher quality and class size," *Journal of Human Resources*, 2009, *44* (1), 223–250.

Kokkelenberg, Edward C, Michael Dillon, and Sean M Christy, "The effects of class size on student grades at a public university," *Economics of Education Review*, 2008, *27* (2), 221–233.

Krueger, Alan B., "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 1999, *114* (2), 497–532.

_ **and Diane M. Whitmore**, "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*, 2001, *111* (468), 1–28.

Leuven, Edwin, Hessel Oosterbeek, and Marte Ronning, "Quasi-experimental estimates of the effect of class size on achievement in norway," *The Scandinavian Journal of Economics*, 2008, *110* (4), 663–693.

Levin, Jesse, "For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement," *Empirical Economics*, 2001, *26* (1), 221–246.

OECD, "Attracting, Developing and Retaining Effective Teachers—Final Report: Teachers Matter," 2005.

_ , "PISA 2012 results: What makes schools successful?: Resources, policies and practices (volume IV)," 2013.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, "Teachers, schools, and academic achievement," *Econometrica*, 2005, *73* (2), 417–458.

Urquiola, Miguel, "Identifying class size effects in developing countries: Evidence from rural Bolivia," *The Review of Economics and Statistics*, 2006, *88* (1), 171–177.

Appendix A The Number of Observations for Each School

Table A.1 shows the panel composition over number of years and cohort size. We have up to 12 years of data for each school. Larger schools have a slightly longer panel.

Table A.1: Number of Years Available and School Size

	Number of Years Available						
Quantile of Cohort Size	6	7	8	9	10	11	12
10	0	0	1	7	1	3	0
20	1	1	1	7	0	2	0
30	0	0	0	8	1	3	0
40	0	1	0	4	1	7	0
50	1	1	0	8	1	1	0
60	0	0	0	7	1	4	0
70	0	0	0	4	3	6	0
80	0	1	0	6	2	3	0
90	0	0	0	7	1	4	0
100	1	0	3	3	1	4	1
Total	3	4	5	61	12	37	1

Appendix B The Restricted Sample when Teacher Data is Available

Table A.2: Estimates of the Baseline Model with Teachers' Fixed Effects.

	(1)	(2)	
Dependent Variable: Subject GPA			
		Second Stage	
ClassSize	0.086***	1.27	
	(0.03)	(0.8)	
ClassSizeSQ		-0.028	
		(0.02)	
Female	0.95***	0.96***	
	(0.08)	(0.08)	
Age	5.67***	5.56***	
	(1.1)	(1.1)	
AgeSQ	-0.20***	-0.20***	
	(0.03)	(0.03)	
Kleibergen-Paap Statistic	5632.2	624.0	
p-value	0.000	0.000	
School FE	YES	YES	
R-sq	0.224	0.223	
N	17212	17212	
		First Stage	
	ClassSize	ClassSize	ClassSizeSQ
Enrollment	0.17***	0.041***	0.47
	(0.002)	(0.009)	(0.4)
Sq of Enrollment		0.00074***	0.038***
		(0.00005)	(0.002)

(1) ClassSizeSQ is the square of ClassSize. Female = 1 if a student is female. Age and AgeSQ control for students' age and its square.

(2) Standard errors are clustered at the class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

For this subsample we are able to add teacher fixed effects as a robustness check to control for the teachers' quality and their different grading standards. Table 3 in the Online Appendix gives the summary statistics for these schools. These 9 schools are clearly different from the full sample. They have a smaller (by 24 students) cohort size. Their class size is smaller (by 2.87 students) as well and number of classes is also smaller by (0.73). Students in these schools are also slightly older.

We run our baseline regression for this subsample where we have teacher data. The results are reported in Table A.2. The same hump shape can be observed, though the coefficients are not significant. This is not surprising given the sample size is much smaller. Note that the dependent variable in this regression is GPA for each subject and for each student, not the average GPA since teachers vary by subject. For this small sample, we include both teachers' fixed effects and subject and school fixed effects.

Appendix C Parametric Estimates: Small Schools and Large Schools

A concern might be that there are few small classes in the sample. Could it be that these few data points are adding curvature to what is really a linear/monotonic regression? As shown in Figure 1, there are indeed few data points below the turning point in the estimated regression.

Figure A.1 plots the relationship between the average cohort size over years and class size. The average cohort size is divided into categories, 0-15, 15-30, ... 120-135, and greater than 135. It is clear that small class sizes tend to occur in small cohorts.

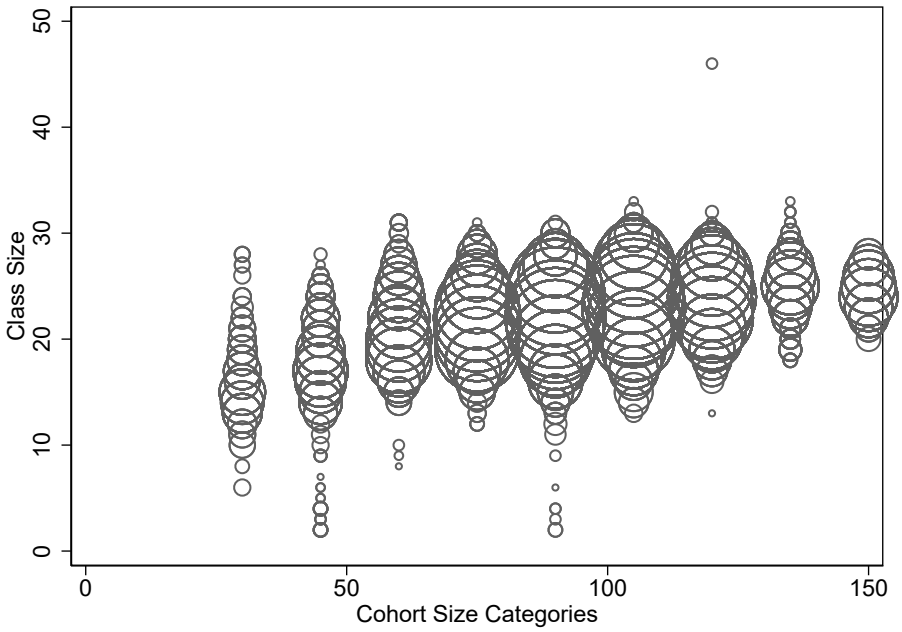


Figure A.1: The Relationship between Cohort Size and Class Size

To deal with this concern, we estimate the parametric model for different sized schools separately. We use the average enrollment to define the size of schools. We use three definitions of small and large schools using the average enrollment over the years of less than (more than) 30, less than (more than) 50 and less than (more than) 70 to define small (large) schools. Table A.3 presents the estimation results for large schools. The hump shape remains and all the coefficients are significant. Moreover, the peak occurs around 19-20 in all three cases.

Appendix D Nonparametric Evidence

We first detrend gpa by regressing on $Female$, Age, $AgeSQ$ and School specific linear trend. We then plot the detrened gpa against class size. The graph is presented in Figure A.2. We see the hump shaped result.

We also use a spline regression to relax the assumptions on the functional form and provide causal evidence. We follow Hojo (2013) in this. We divide class size into three sections, $\leq 18, 18-23$ and > 23. We deliberately choose 18 as our first cutoff since 18-19 is the turning point in our IV regression. The second cutoff 23 is the 50 percent quantile of class size. Since $Enrollment$ is positively correlated with $ClassSize$, we divide our instrument $Enrollment$ into three sections $\leq 50, 50-95$ and > 95. Table A.4 reports the spline regression. It shows that GPA is increasing with small class size while decreasing with large class size. The coefficients for the first region

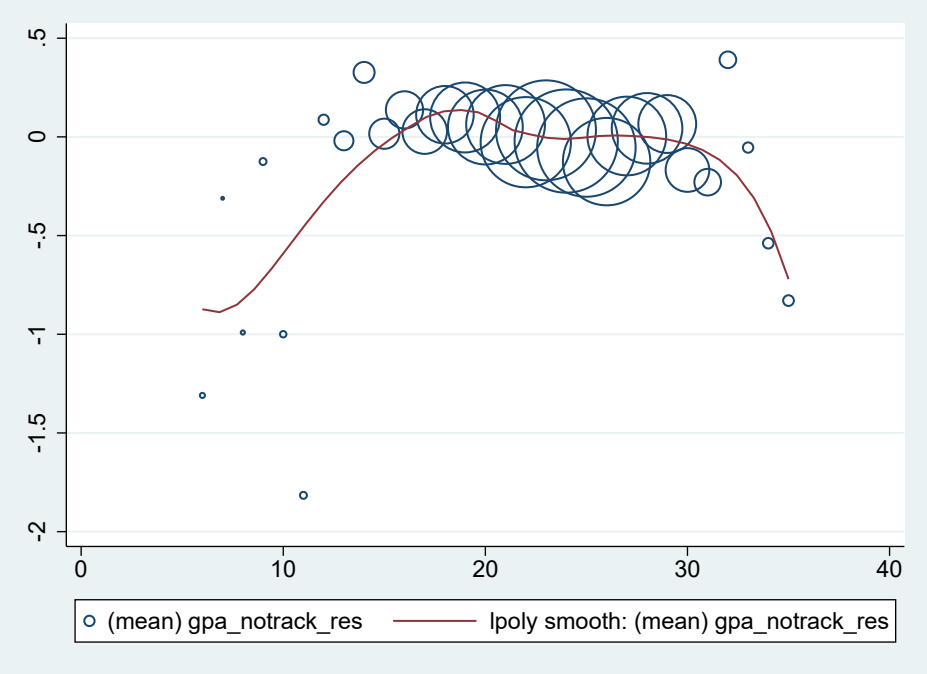


Figure A.2: Class Size and GPA

is significant and positive, and the one for the third region is significantly negative. The second region is not significantly different from zero.

Table A.3: IV Estimates for the Baseline Model: Large Schools

	(1)		(2)		(3)	
Dependent Variable: GPA						
			Second Stage			
Average CohortSize	≥ 30		≥ 50		≥ 70	
ClassSize	0.39**		0.39**		0.29*	
	(0.2)		(0.2)		(0.2)	
ClassSizeSQ	-0.010***		-0.010**		-0.0082**	
	(0.004)		(0.004)		(0.004)	
Female	0.89***		0.88***		0.89***	
	(0.03)		(0.03)		(0.03)	
Age	-1.78***		-1.71***		-1.65***	
	(0.10)		(0.1)		(0.1)	
AgeSQ	0.028***		0.027***		0.025***	
	(0.002)		(0.002)		(0.003)	
Kleibergen-Paap Statistic	44.7		39.4		42.1	
p-value	0.000		0.000		0.000	
School FE	YES		YES		YES	
School-Specific Linear Time Trend	YES		YES		YES	
R-sq	0.063		0.062		0.062	
N	81178		77167		71419	
			First Stage			
	ClassSize	ClassSizeSQ	ClassSize	ClassSizeSQ	ClassSize	ClassSizeSQ
Enrollment	0.15***	5.30***	0.15***	5.33***	0.16***	5.35***
	(0.01)	(0.7)	(0.02)	(0.7)	(0.02)	(0.7)
Sq of Enrollment	-0.00044***	-0.011***	-0.00043***	-0.011***	-0.00045***	-0.012***
	(0.00007)	(0.003)	(0.00007)	(0.004)	(0.00008)	(0.004)

(1) ClassSizeSQ is the square of ClassSize. Female = 1 if a student is female. Age and AgeSQ control for students' age and its square.
(2) Standard errors are clustered at the class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table A.4: Spline Regression of Class Size Effect

	(1)		
	Second Stage		
ClassSize $\times$ $\mathbb{1}\{ClassSize \leq 18\}$	0.17**		
	(0.08)		
ClassSize $\times$ $\mathbb{1}\{18 < ClassSize \leq 23\}$	-0.059		
	(0.07)		
ClassSize $\times$ $\mathbb{1}\{ClassSize > 23\}$	-0.14**		
	(0.06)		
Female	0.90***		
	(0.03)		
Age	-1.79***		
	(0.10)		
AgeSQ	0.028***		
	(0.002)		
Kleibergen-Paap Statistic	95.6		
p-value	0.000		
School FE	YES		
School-Specific Linear Time Trend	YES		
R-sq	0.064		
N	81845		
	First Stage		
	ClassSize $\times$ $\mathbb{1}\{CS \leq 18\}$	ClassSize $\times$ $\mathbb{1}\{18 < CS \leq 23\}$	ClassSize $\times$ $\mathbb{1}\{CS > 23\}$
Enrollment $\times$ $\mathbb{1}\{Enrollment \leq 50\}$	0.14***	0.090***	0.044***
	(0.01)	(0.01)	(0.009)
Enrollment $\times$ $\mathbb{1}\{50 < Enrollment \leq 95\}$	0.0061***	0.045***	0.011***
	(0.002)	(0.004)	(0.004)
Enrollment $\times$ $\mathbb{1}\{Enrollment > 95\}$	0.000089	0.011***	0.046***
	(0.0003)	(0.002)	(0.005)

(1) Female = 1 if a student is female. Age and AgeSQ control for students' age and its square.
(2) Standard errors are clustered at the class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Appendix E Counterfactual with the Linear Model

Table A.5 reports the parameter estimates and Table A.6 reports the results of counterfactual exercises when gpa is a linear function of class size. First note that the estimated hiring and firing costs are much smaller than in the baseline model. The hiring cost is €9,731 versus €58,531 and the firing cost is €8,846 versus €54,435. This makes sense. The pre value function is less peaked with the linear model estimated. As a result, the adjustments seen in the data can be rationalized by lower hiring and firing cost. The counterfactuals show that the change in student performance is larger and the change in class size is smaller in the quadratic model than that in the linear model. This is to be expected since greater concavity reduces responses and increases impact of a response. The welfare change in the euro value is much smaller in the linear model.

Table A.5: Estimates of the Structural Dynamic Model When Using linear GPA Function

	c	H	F	σ
All	480.59	227.32	206.65	219.25
sd	(55.12)	(54.12)	(52.47)	(6.22)
	€20,572	€9,731	€8,846	

(1) The standard errors are presented in parentheses.

Table A.6: Counterfactuals When Using linear GPA Function

School Size	Δ Average Class Size (#)	Δ Average GPA (points)	Δ Cost (€)	Δ Welfare (€)
		$F' = 0$		
Small	2.58	-0.17	-372.25	2643.65
Medium	3.78	-0.25	-721.62	3405.55
Large	3.60	-0.24	-1010.80	4003.04
		$H' = F$		
Small	-2.33	0.15	345.08	1248.92
Medium	-1.85	0.12	467.57	1591.53
Large	-1.74	0.12	610.89	1869.33
		$c' = 1.5c$		
Small	3.99	-0.26	671.06	-26592.00
Medium	5.49	-0.36	1121.17	-47330.85
Large	5.40	-0.36	1657.54	-68712.03
		Class Cap = 25		
Small	-9.37	0.62	1609.09	-16845.83
Medium	-6.86	0.45	1896.82	-15525.18
Large	-6.02	0.40	2430.32	-17733.49
		Class Cap = 30		
Small	-7.77	0.51	1007.19	-8176.88
Medium	-4.70	0.31	1106.81	-7698.82
Large	-3.64	0.24	1177.12	-6497.33
		Class Cap = 35		
Small	-7.36	0.49	939.87	-7312.50
Medium	-2.39	0.16	474.57	-2615.16
Large	-0.92	0.06	233.81	-1141.06

(1) These groups are formed by putting the lowest two quantiles in group 1 and the highest two quantiles in group 3.

(2) Cost increases in this counterfactual are per year, while welfare increases are the present discounted value over time. This is why welfare decreases are so much larger in magnitude here.