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

August 7, 2019

Abstract

Using high quality administrative data on Greece we show that class size has a hump shaped effect on achievement. We do so both nonparametrically and parametrically, while controlling for potential endogeneity and allowing for quantile effects. We then embed our estimates for this relationship in a dynamic structural model with costs of hiring and firing.

We argue that the linear specification form used in past work may be why it found mixed results. Our work suggests that while discrete reductions in class size may have mixed effects, discrete increases are likely to have very negative effects while marginal changes in class size would have small negative effects.

We find optimal class sizes around 27 in the absence of adjustment costs and achievement maximizing ones around 15, and firing costs much larger than hiring costs consistent with the presence of unions. Despite this, reducing firing costs actually reduces achievement. Reducing hiring costs raises achievement and reduces class size. We show that class size caps are costly, and more so for small schools, even when set at levels well above average.

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

JEL: C6, IE, JE

1 Introduction

What determines student achievement? The usual approach is to think of achievement as the output of an educational production function. Inputs into this educational production function include teacher quality, class size, resources, peer effects (possibly positive spillover effects and negative disruption effects), as well as past achievement since achievement builds on the past knowledge.

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. A recent paper, 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 on Greece to first show nonparametrically that there does indeed seem to be such a hump shape in the data. Following this, we estimate a parametric 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 specification form used in past work may be why past results were mixed. After all, if we fit a linear regression when the true relationship is quadratic, 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 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

[1] Most work assumes a linear form.

to the other side of the hump and have little or no effect on achievement. For these reasons, 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.

We further explore the data to look for evidence of quantile effects. We find that the hump shape is present across all quantiles, i.e., for students of all abilities. The hump shape is however more pronounced for worse students.

This paper proceeds as follows. In Section 2, we put our work in perspective relative to the literature. In Section 3, we describe where the data came from, 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 parametric approach and use the Hoxby instrument, see Hoxby (2000), to control for endogeneity. In Section 5, we present our quantile IV results.

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. If the government cares about achievement, and faces costs of adding classes, its behavior in terms of the number of classes it chooses as enrollment fluctuates helps us estimate the costs involved. In Section 6 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. Our estimates here are in line with actual teacher salaries. 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 government and the impact of this on student achievement if any. We find that unions, even if they raise costs and class size, have a small effect on achievement. Finally, we look at the costs versus benefits of class size requirements. Section 7 concludes.

2 Relation to the Literature

Given the increasing importance of skills in the labor force in the 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 and headmasters are better informed about students than the econometricians. Based on students' characteristics that the econometricians do not observe, headmasters tend to allocate better students to larger classes, thus generating a positive correlation between class size and student performance. As a result, 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.[2] 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 as described below.

The best known experiment 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 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 15% per year of experience, class size reductions would seem the more expensive option.[3] Chetty et al. (2014) shows that teacher quality measured as value added has a huge effect on outcomes. Using U.S. data on over a million primary school children, he shows that replacing a teacher in the lowest 5% of value added with the average teacher would have significant positive effects on outcomes like college attendance and teenage pregnancy and increase the lifetime earnings of the students in a classroom by $250,000.

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

[2] 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 a upward bias).

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

seems to be 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 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 seems to be closer to 24 suggesting that the limit is not binding and the quasi experimental approach is invalid.

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.

Gary-Bobo and Mahjoub (2013) use data on French junior high schools and Urquiola (2006) use Bolivian data and follow Hoxby's approach. Though the estimated causal effects of larger class size tend to be negative, they remain small. In the context of the literature, our approach follows Hoxby (2000). In our work, as there is no explicit class size cap, we cannot use the Angrist and Lavy approach. As a result, we use Hoxby's instrument.

Levin (2001) and Dobbelsteen et al. (2002) use a third source of quasi experimental variation. They use PRIMA data. This longitudinal survey of Dutch students in grades 2,4, 6 and 8 in 1994-5 is rich in information including IQ as well as a new instrument for class size. Dutch rules link the number of teachers that can be hired to enrollment and this provides quasi exogenous variation in the number of classrooms. Levin 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.

Even with a good experiment, the literature has made a clear distinction between interpreting the coefficients as structural parameters (i.e., the causal effect of class size) and policy estimates (i.e., the expected effect of an exogenous policy on class size). For example, suppose we changed class size experimentally (so that one group of students was in large classes and another was in small classes) and parental behavior responded to these changes, the estimated effects would be compound effects including the pure effect of changing class size and the induced one on parental behavior. Todd and Wolpin (2003) in particular emphasize that estimates of the class-size effect even using experimental data should be interpreted as policy effects. In contrast, estimates aimed at identifying structural parameters of the education production function could be interpreted as the pure effect of class size, if other channels, like parental inputs in the example above, are accounted for.

It is worth noting that we could find only two papers that allowed for nonmonotonic effects. Borland et al. (2005) uses 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. There are a number of issues with the paper. First, the economic model behind their simultaneous equation system and the exclusion restrictions used is far from clear. Second, their estimates are difficult to reconcile with their data patterns. Their estimates suggest a peak of achievement around class size 26. If class size was being chosen to maximize achievement subject to costs, the optimal class size must be to the right of the peak of achievement. The optimal class size cannot be below 26 as raising class size would raise achievement and reduce costs. However, 99% of data has class size below 26 which is hard to explain in terms of economics. The paper also only presents the achievement equation and even for this equation, presents only a subset of the estimated coefficients.

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.

We argue that class size effects seem to be nonmonotonic, with class size initially increasing and then reducing achievement. It could be that this hump shape might be why restricting the functional form to be monotonic gave estimates that were small in size and variable in sign. In addition, it is worth emphasizing that much 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 it 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. We allow for one such channel of heterogeneous class size effects in our quantile IV regressions. We are also able to control for teacher fixed effects, though only for a limited subsample. Neither teacher fixed effects nor heterogeneous class size effects change our basic point and results regarding nonmonotonicity.

3 Data and Institutional Background

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 teachers get paid on the same scale as entry level permanent teachers, but only for the work they do. Permanent staff is very difficult to fire, especially in public schools where firing permanent staff is almost unheard of. Not only is there compensation, 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. See Stylianidou et al. (2004) for details of how the system works.

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.[4] 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.

[4] 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).

In the 10th grade, students have, for the most part, a common curriculum.[5] In the 11th and 12th grade, they start to differ as they choose their tracks.[6] 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.[7]

Students are assigned to a class (1,2,3,4, etc.). 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 through 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.[8] This is likely to make the Hoxby instrument work better in the 10th grade. For all three reasons we focus our attention to 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 and the school annual exam. The school annual exam is course and teacher specific. Performance is measured on a continuous scale from 0-20. We take the simple average of the annual exam across non track compulsory subjects (Ancient Greek, Literature, Modern Greek, History, Algebra, Geometry, Physics, Chemistry, Economics, and Technology) to

[5] 10th grade compulsory subjects include religion, ancient Greek, literature, modern Greek, history, algebra, geometry, physics, chemistry, economics, technology and one foreign language.

[6] 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.

[7] 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.

[8] 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.

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 to choose in the 11th grade and an average score of 50% in school exams is needed to sit for the university entrance exam.

3.1 Summary Statistics

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 lies between 16 and 28 class size. 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 rural area. Note that we have school fixed effects in our baseline regressions below which would absorb 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 lightly longer panel.

The sample of schools that we use in this study does not differ systematically from the population of schools in Greece. To show this, we compare various variables between our sample and the whole population of schools in the country. In particular, we provide some 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

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.

[2] Standard deviations are clustered at class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Ministry of Education and Religious Affairs[9] and include all schools in the country, except for the evening schools (which are only designed for employed students). Our sample seems to be 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 commence the 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).

The OECD[10] 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 actually really close to the class size that we observe in our data (22.6).

3.2 Data Patterns

3.2.1 Class Size and Enrollment

Figure 1 plots class size versus enrollment for grade 10. The red dashed line gives the predicted class size had there been a binding cap on class size of 27. The data loosely follows this red line, but since no cap on class size was in place officially, this "targeted" class size may be a result of administrators 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.

3.2.2 Class Size, Enrollment and GPA

Figure 2 plots a smooth version of the relationship between enrollment and class size given by the black line, and between enrollment and GPA given by the red line. It is worth noting that especially for low enrollments, class size and GPA seem to be negatively related. As enrollment rises, class size first rises till enrollment reaches the mid 20's. After this class size falls and then rises again near 45 and so on. The turning points of class size and of GPA seem to be the same

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

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

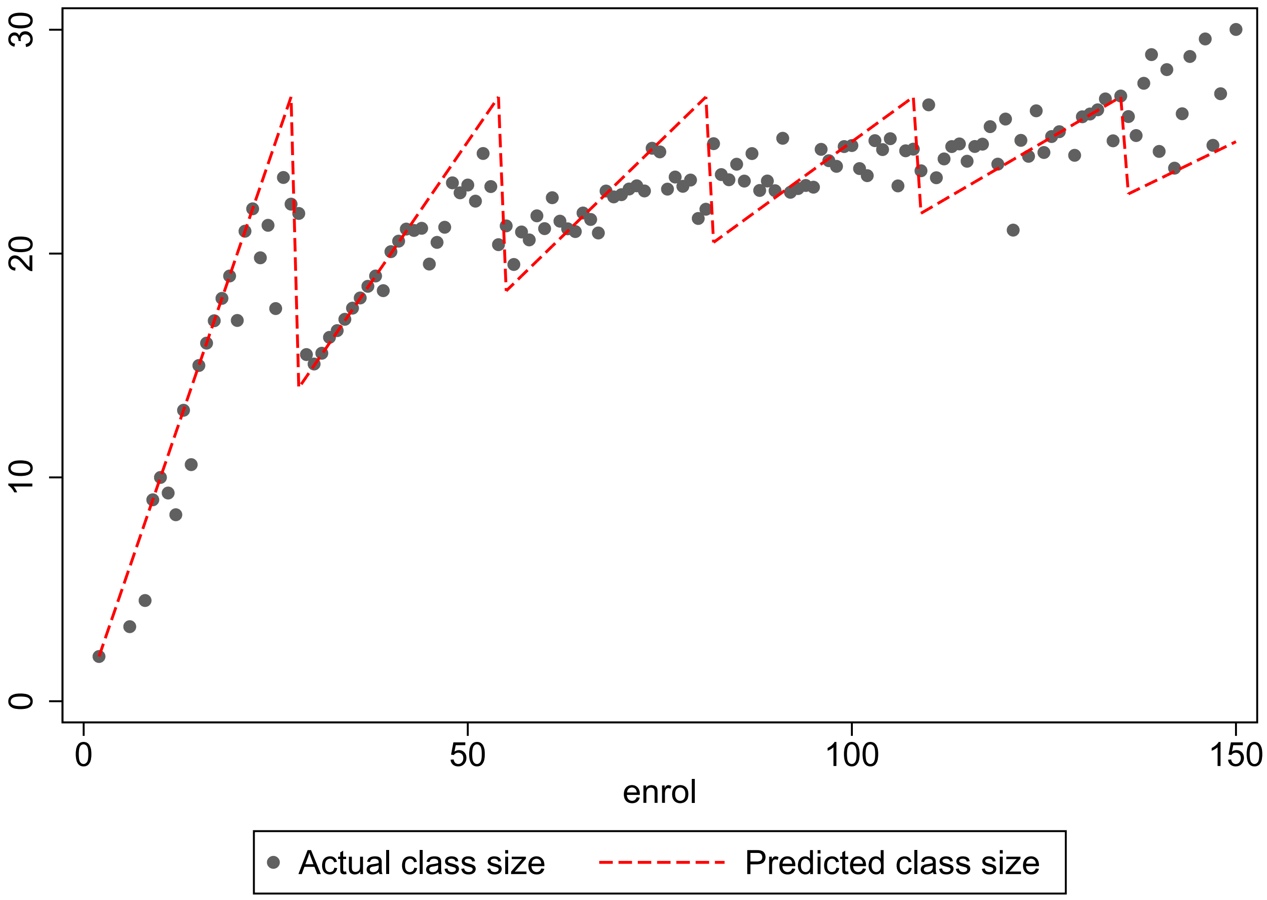

Figure 1: Class Size versus Enrollment for Grade 10

so that when one peaks, the other reaches its trough. This relationship becomes much fuzzier for large enrollments.

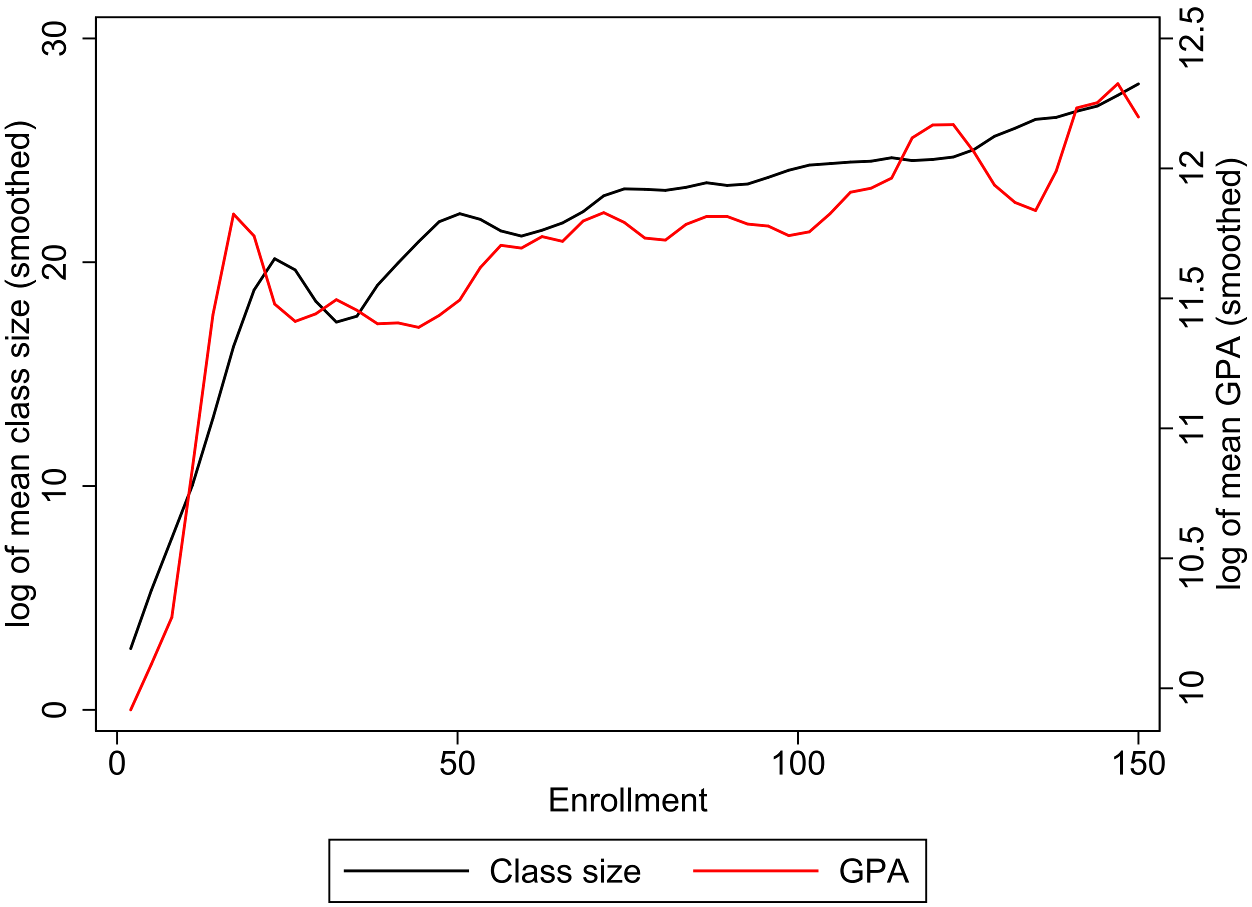

Figure 2: Class Size versus Enrollment for Grade 20 (Smooth Version)

3.2.3 OLS Estimates

As a purely descriptive exercise, we next turn to the OLS estimates of class size and GPA. 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 nonmonotonicity and gives a negative and significant coefficient for class size. Column 2 adds a quadratic term in class size. The coefficients now point to a hump shape with a turning point around 12.

3.2.4 Nonparametric Evidence

Recent work by Chernozhukov et al. (2013) provides a framework applicable to our setting and which does not assume a particular functional form. In essence, it allows for school fixed effects in a non linear manner. The approach needs panel data, which we have, and the endogenous regressor (class size) has to be discrete.

Following their approach we specify the following model:

$$GPA_{jt} = g(CS_{jt,}\alpha_j, \varepsilon_{jt})$$

which has achievement as measured by the *average* GPA of school j in period t being a function of the *average* class size in school j in period t, a school fixed effect, α_j, and a shock, ε_{jt}, that is school and time specific. Though class size is discrete, average class size is a continuous variable. For this reason we discretize the class size into bins below.

We specify this relationship to be at the school level, because nonparametric estimation of this kind needs a long panel for each j.[11]

The assumption needed for this approach is the following:

Assumption 1 (time-homogeneity)

$$\varepsilon_{jt} \mid \boldsymbol{CS}_j,\ \alpha_j \sim F(. \mid \boldsymbol{CS}_{j,}\alpha_j).$$

In other words, the distribution of the shock ε_{jt}, conditional on the *vector* of average class sizes for the school at all periods (denoted by $\boldsymbol{CS}_j$) and the school itself, is time independent as the function F has no time subscript. Stated slightly differently, whatever the distribution of the shock is, its conditional distribution given the vector of average class sizes for school j does not depend on t. Chernozhukov et al. (2013) interprets this as time being randomly assigned or time being an instrument along with the distribution of factors other than class size not changing over time.

One might be concerned that Assumption 1 does not hold in the data and as a result, the approach of Chernozhukov et al. (2013) cannot be used. Fortunately, we need not take the assumption on faith. A recent paper, see Ghanem (2017) derives a statistical test to check the validity of Assumption 1. In the Appendix C we show that using this methodology, we cannot

[11] If we had specified the model to hold at the individual level, we would have a panel of length three, though we would have a lot of students. Similarly, we could have specified the model to be at the class level if we had information on which teacher was assigned to which class. In this case, we would have a panel of the same length as that for the model we use, assuming the teacher was there throughout. We do not have data on teachers and their assignment to classes we cannot use this approach.

reject the hypothesis that Assumption 1 holds in the data.[12]

Estimates We choose to discretize class size into three bins. We do so as we will need to estimate the effect going from each bin to the other so that the number of coefficients rises rapidly with the number of bins. The first is class size below a cutoff s_0. The second bin is from s_0 to s_1, and the third is more than s_1. Since these switches are identifying the effects of interest, we need to choose s_0 and s_1 to ensure that the bins are such that these switches occur.

Let δ_{lk} be the average effect on mean GPA in a school of switching from bin l to k and $\hat{\delta}_{lk}$ be its consistent estimator. In effect, what is done is the following. In each year, each school has a mean GPA and a mean class size and so falls into one of the three bins. Over the entire sample period, each school indexed by j can be in either bin 1 only, bin 1 and 2 only, bin 1 and 3 only, bin 2 and 3 only or in all three bins. We calculate the mean GPA *over time* for school j when in bin $l = 1, 2, 3$. For example, if there were 5 periods and the school was in bins $(1, 2, 1, 3, 2)$ over these time periods with GPA $(g_1, g_2, g_3, g_4, g_5)$, then the mean GPA in bin 2 would be $(g_2 + g_5)/2$ while the mean GPA in bin 1 would be $(g_1 + g_3)/2$. Their difference would capture Δ_{12}^j for school j. The estimated $\hat{\delta}_{12}$ would then be

$$\hat{\delta}_{12} = \frac{\sum_{j=1}^{N} d_j \Delta_{12}^j}{\sum_{j=1}^{N} d_j}$$

where d_j is 1 if the school was ever in both bin 1 and 2 over the entire sample period.

We set s_0 at 15 and s_1 at 22. $\hat{\delta}_{12} = 1.55$ and $\hat{\delta}_{23} = -.18$. Both are significantly different from zero at the 1% level. In Table A.3 in the Appendix D, we vary s_1 from 21 to 24 along the rows and s_0 from 12 to 17 along the columns. For each value of s_0 and s_1 we give the estimate of $\hat{\delta}_{12}$ and $\hat{\delta}_{23}$. Note that no matter what s_0 to s_1 are set at, $\hat{\delta}_{12} > 0, \hat{\delta}_{23} < 0$ and significant. This is consistent with a nonmonotonic relationship between GPA and class size.

3.3 Endogeniety Issues

Should we interpret the estimates in this section as representing the production technology in the classroom between class size and GPA or learning? The answer is no.

Why? Suppose class size is being chosen to maximize an increasing function of the learning, i.e., the GPA of the school, less costs of operation, and say enrollment is exogenously given. For a given quality of students, or teacher, denoted by q and given enrollment, e, Figure 3 depicts the

[12] We thank Dalia Ghanem for sharing programming code with us.

Table 2: OLS Estimation of Class Size Effects

	(1)	(2)	(3)	(4)
lnClassSize	-0.57	3.23	-0.49	3.64
	(0.1)***	(1.0)***	(0.1)***	(1.1)***
lnClassSizeSQ		-0.65		-0.70
		(0.2)***		(0.2)***
female	0.89	0.89	0.88	0.88
	(0.03)***	(0.03)***	(0.03)***	(0.03)***
Age	-1.55	-1.56	-1.48	-1.50
	(0.09)***	(0.09)***	(0.09)***	(0.09)***
AgeSQ	0.024	0.024	0.023	0.023
	(0.002)***	(0.002)***	(0.002)***	(0.002)***
sd of ln GPA			-5.77	-5.79
			(0.3)***	(0.3)***
School FE	YES	YES	YES	YES
R-sq	0.057	0.057	0.064	0.064
N	81845	81845	81845	81845

[1] lnClassSize is log of Class Size, and lnClassSizeSQ is the square of lnClassSize. Female is the dummy for students' sex. Female = 1 if a student is female. Age and AgeSQ controls for students'. sd of ln GPA controls for the standard deviation of ln GPA within a class.

[2] Standard deviations are clustered at class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

benefits (B) and costs (C) as class size rises. As class size rises, the benefits may first rise but ultimately fall and this is depicted as concave function.[13] However, as class size rises, costs fall and this is depicted as a downward sloping convex function. Optimal class size, i.e., the chosen class size, is where the differences in the two is largest. This occurs at $CS^*(e, q)$. Thus, in the data we will see $(e, CS^*(e, q), gpa(e, q, CS^*(e, q))$. Moreover, we will most likely not observe q. What happens if q is higher? If for example, students are better, then the benefit curve will shift up and become flatter as depicted by the curve B' since better students would learn more (have a higher GPA) at any given class size and suffer less from larger classes. But faced with a better class, the chosen class size will rise as depicted to $CS^*(e, q')$ and in the data we will see $(e, CS^*(e, q'),\ gpa(e, q, CS^*(e, q'))$. Though we want to estimate the curve B, the data will trace out a flatter curve than B. This is the essence of the bias in the OLS estimates and the upward bias explains why OLS coefficients on class size often turn out to be positive.

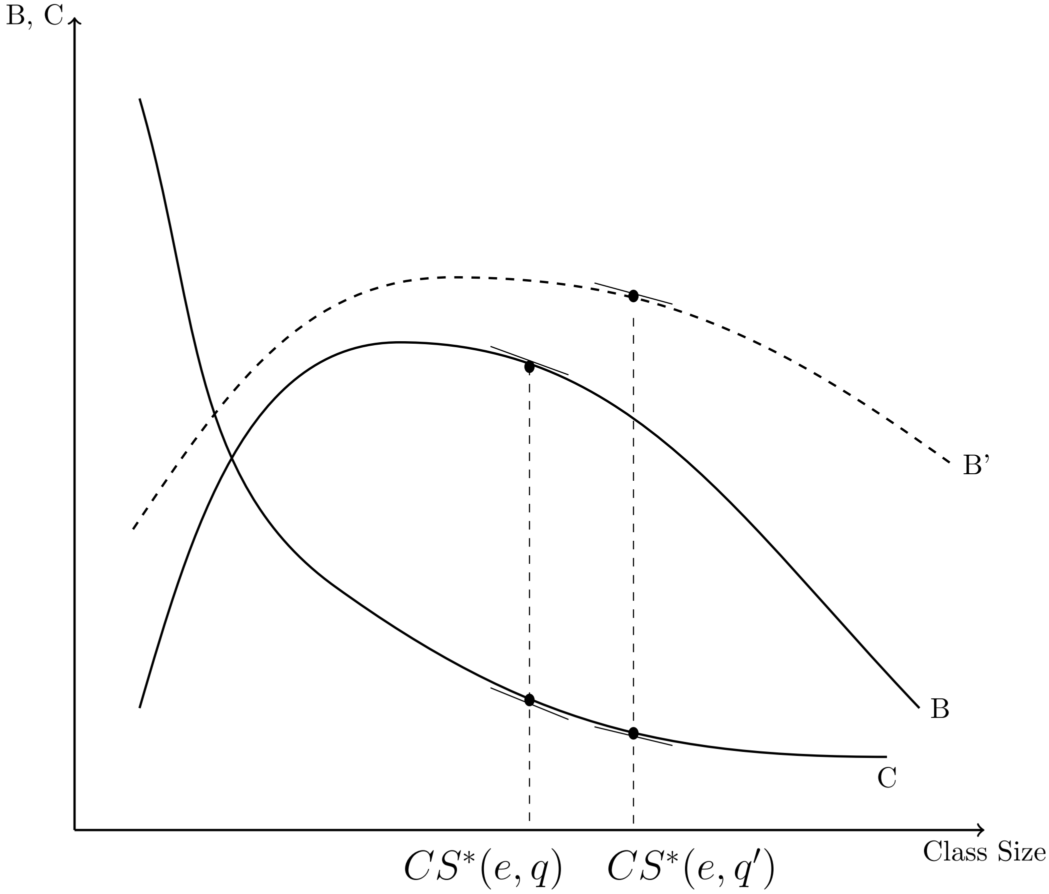

Figure 3: Tradeoff between GPA and Class Size

[13] Even if GPA first rises and then falls with class size, one will never choose to be on the upward sloping part.

4 Linear and Nonlinear 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 parametric specification. Our specification is:

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

GPA for individual i in school j at time t depends on the log of class size, its square, school fixed effects, and a set of control, X_{ijt}, which include gender, age, age squared, the standard deviation of the GPA in the class. 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. (2002). For this reason we include the standard deviation of GPA in the class as a control.

Since class size could be an endogenous variable, we need an instrument. We cannot use the Angrist and Lavy (1999) approach. There is no maximum class size on the books in Greece in our period. The data patterns described in Section 3.2 clearly suggest that class size is endogenous. From looking at the pattern of enrollment and class size in Figure 1, it seems clear that class size is not allowed to get too large: 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.[14] We follow the approach of Hoxby (2000). It is natural to think of overall enrollment as an exogenous shock to class size. Hoxby's approach 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. We run the analogue of Hoxby's approach running the following.

$$\ln e_{it} = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \varepsilon_{it}$$

In the Table 3 and 4, $\ln U$ denotes the deviations of the data from the predicted values, while $\ln USQ$ denotes the square of $\ln U$ in Table 4 and 5.

Table 3 and Table 4 give the IV estimates for grade 10 for the linear and quadratic models respectively. We present estimates when the standard deviation of GPA in the class is controlled

[14] In fact, when we tried using the Angrist Lavy approach, though the first stage did not cause any problems, the second stage gave insignificant/mixed sign coefficient estimates.

for and when it is not. 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 Table A.4 and Table A.5 in Appendix E. The upper panel gives the estimates for the second stage in all these tables.

Recall that the OLS estimate of the coefficient on $\ln ClassSize$ in the linear regression was negative and about -0.57. The coefficient with the IV for the same regression is given in column 1 and is -3.25. 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. If we add the standard deviation of GPA as a control, as in column 2, the coefficient on lnClassSize is slightly smaller. The coefficient on standard deviation is negative, consistent with more diverse students being harder to teach. Table 4 gives the estimates for the quadratic specification. It clearly has the hump shape expected with a peak at around 14.9.

We use Hoxby's instrument so that we would expect the shock in enrollment to be positively correlated with class size as we find.[15] We also are not concerned that enrollment is endogenous. 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. Note that the first stage looks fine: the coefficient on the instrument is positive significant at 1% and the instruments are not weak as the Kleibergen-Paap LM statistic is 114.7. It is interesting, and in line with the literature that women have a higher GPA. The standard deviation of ln GPA in a class is added as an explanatory variable in the second column of Table 3 and Table 4. It is significant at the 1% level and negative. This suggests that the more homogeneous the class, the higher the GPA as in Levin (2001) and Dobbelsteen et al. (2002).

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. 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.

[15] It is worth pointing out that we have at most 12 years of data for a school while Hoxby had 24.

Table 3: Parametric Estimation of Linear Class Size with IVs

	(1)	(2)
Dependent Variable: GPA		
	Second Stage	
lnClassSize	-3.25	-2.85
	(0.5)***	(0.4)***
female	0.89	0.88
	(0.03)***	(0.03)***
Age	-1.59	-1.53
	(0.10)***	(0.10)***
AgeSQ	0.024	0.023
	(0.002)***	(0.002)***
sd of ln GPA		-5.56
		(0.3)***
Kleibergen-Paap Statistic	1029.0	1030.6
p-value	0.000	0.000
School FE	YES	YES
R-sq	0.047	0.056
N	81845	81845
	First Stage	
	lnClassSize	lnClassSize
lnU	0.25	0.24
	(0.01)***	(0.01)***

[(1)] lnClassSize is log of Class Size, and lnClassSizeSQ is the square of lnClassSize. Female is the dummy for students' sex. Female = 1 if a student is female. Age and AgeSQ controls for students'. sd of ln GPA controls for the standard deviation of ln GPA within a class.

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

Table 4: Parametric Estimation of Nonlinear Class Size with IVs

	(1)		(2)	
	Dependent Variable: GPA			
	Second Stage			
lnClassSize	29.1		29.0	
	(5.9)***		(5.9)***	
lnClassSizeSQ	-5.44		-5.36	
	(1.0)***		(1.0)***	
Female	0.89		0.88	
	(0.03)***		(0.03)***	
Age	-1.70		-1.64	
	(0.10)***		(0.09)***	
AgeSQ	0.027		0.026	
	(0.002)***		(0.002)***	
sd of ln GPA			-5.76	
			(0.4)***	
Kleibergen-Paap Statistic	114.7		114.7	
p-value	0.000		0.000	
School FE	YES		YES	
R-sq	0.032		0.042	
N	81845		81845	
	First Stage			
	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ
lnU	0.45	2.22	0.45	2.22
	(0.06)***	(0.3)***	(0.06)***	(0.3)***
lnUSQ	-0.065	-0.24	-0.064	-0.24
	(0.02)***	(0.10)**	(0.02)***	(0.10)**

[1] lnClassSize is log of Class Size, and lnClassSizeSQ is the square of lnClassSize. Female is the dummy for students' sex. Female = 1 if a student is female. Age and AgeSQ controls for students'. sd of ln GPA controls for the standard deviation of ln GPA within a class.

[2] Standard deviations are clustered at class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 5: Nonlinear Class Size with Teachers' Fixed Effects

	(1)		(2)	
	Dependent Variable: Subject GPA			
	Second Stage			
lnClassSize	50.2		40.3	
	(21.9)**		(21.6)*	
lnClassSizeSQ	-8.19		-6.46	
	(3.6)**		(3.6)*	
Female	0.96		0.95	
	(0.08)***		(0.08)***	
Age	5.64		5.69	
	(1.1)***		(1.1)***	
AgeSQ	-0.20		-0.21	
	(0.03)***		(0.03)***	
sd of ln GPA			-5.62	
			(0.3)***	
Kleibergen-Paap Statistic				
p-value	1648.8		1649.5	
School FE	YES		YES	
Subject FE	YES		YES	
Teacher FE	YES		YES	
R-sq	0.224		0.240	
N	17212		17212	
	First Stage			
	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ
lnU	-0.0079	-0.27	-0.0061	-0.26
	(0.02)	(0.09)***	(0.02)	(0.09)***
lnUSQ	0.17	1.12	0.17	1.12
	(0.006)***	(0.03)***	(0.006)***	(0.03)***

[1] lnClassSize is log of Class Size, and lnClassSizeSQ is the square of lnClassSize. Female is the dummy for students' sex. Female = 1 if a student is female. Age and AgeSQ controls for students'. sd of ln GPA controls for the standard deviation of ln GPA within a class.

[2] Standard deviations are clustered at class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

A possible concern might be that we have so far not controlled for which teachers taught which class. We were able to obtain and digitalize teachers' assignment data for 10 schools. Table A.6 in Appendix F gives the summary statistics for these schools. Note that these 10 schools are clearly different from the full sample. They have smaller cohort size by 22 students. Their class size is smaller as well by 2.56 students and the class number is 0.69 smaller. Students in these schools are also slightly older. We also run the same regression as Table 4 for this subsample where we have teacher data. The results are reported in Table A.7. The same hump shape can be observed through, though it becomes insignificant when the standard deviation of log GPA is added[16]. For this subsample, we are able to control for teacher fixed effects as a robustness check, to control for the teachers' quality. The dependent variable is GPA for each subject and for each student, not the average GPA since we are controlling for teacher fixed effects. In addition to teachers' fixed effects, we include subject and school fixed effects in Table 5. Though the point estimates do change a bit, the quadratic form remains.

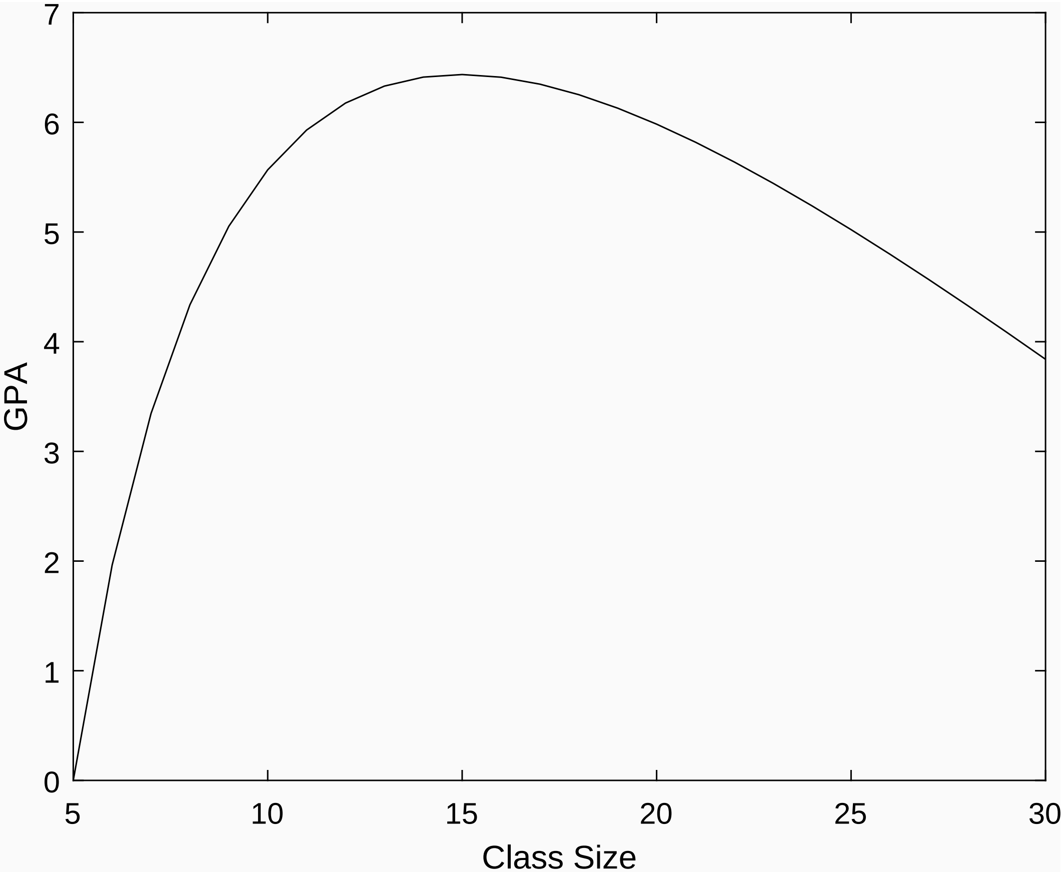

Figure 4: Estimated GPA production function

[16] The coefficients of controls for age and age square are reversed in the restricted sample with or without teachers fixed effects.

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. The histogram of class size at the class level is given in Appendix B as Figure A.1. Roughly 5% of the classes have class size below the turning point of 15. To deal with this issue we see if our results remain when we restrict the sample to small schools where the maximum cohort size is limited to 30, 50 and 70. This reduces the sample size in terms of students to .8%, 5.7% and 13% respectively of the total. The estimates are presented in Table A.2 in Appendix B. The hump shape remains, the coefficients are significant once the cohort size allowed reaches 50, and the peak occurs around 15 in all cases.

One might also worry that class sizes 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 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. To some extent, this alleviates this concern.[17] For the rest of the paper, we return to using the full sample.

5 Quantile Effects

So far we have allowed for nonmonotonicities in how class size affects GPA. In this section, we extend our approach to allow for quantile effects. We do so as it is possible that students of different abilities respond differently to larger class size. One hypothesis is that better students are less affected by class size as they are able to “do it alone”. If this were the case, we might expect a hump shape for all quantiles, but with the hump being flattened out at higher quantiles.

It is worth recalling that a quantile regression does not simply take the data and split it into quantiles, conditional on the independent variable, as doing so would result in selection bias. Rather, it does something more subtle. Suppose we had a linear regression model,

$$y = X\beta + u$$

where we observe data (X_i, y_i) for $i = 1...n$ individuals, and we wanted to allow for quantile effects. Since $\left(y - X\hat{\beta}\right)$ would be the best proxy for u, the coefficient $\hat{\beta}_\alpha$ for the linear regression for the

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

αth quantile is obtained as

$$\hat{\beta}_\alpha = arg\ \min \sum_i \alpha \left| y_i - X_i\beta \right| I(y_i - X_i\beta > 0) + (1-\alpha) \left| y_i - X_i\beta \right| I(y_i - X_i\beta < 0).$$

Suppose $\alpha = .1$. Then, for individual i there is a 10% probability that $y_i - X_i\beta > 0$ and a 90% probability that it is negative. The above formula chooses $\hat{\beta}_{.1}$ so that if 10% of the deviations are positive and 90% are negative, the expected deviations are minimized. Note the contrast to the simple minded approach described above. In fact, even if the model were truly linear, choosing 10% of the data at any X to lie below the estimated regression would not even give a linear regression coefficient.

We use the approach of Lee (2007) to estimate the quantile IV regressions. Note that we do not include school fixed effects in the quantile IV regression since we were unable to include such a large number of fixed effects. We use the same instrument as in the previous section. We present the quadratic version of the regression.

Table 6 gives the estimates for the quadratic form and Figure 5 depicts these estimates graphically. Note that in order to be able to compare the three curves visually we anchor them to zero at class size 10. It is worth noting that the hump shape remains. However, it is clear from this depiction that the 10% quantile (the worst students) is the most hump shaped with the 90% quantile (the best students) is a little less hump shaped, while the 50% quantile is the leasy hump shaped. This suggests that the worst and best students are more affected by class size than those in between.

While we have specified a quadratic form for our model which allows for nonmonotonic relationship between GPA and class size in our quantile regressions, it would be ideal to do this fully nonparametrically. Unfortunately, we are unaware of any technique to do so at this time and this is left for future work. Note that we did estimate the regression, though not allowing for quantile effects, nonparametrically in Section 3.2.4.

With the estimates of the effects of class size on achievement in hand, we are in a position to understand how class size might be chosen. If the government cares about achievement, and faces costs of adding and removing classes, its behavior in terms of the number of classes it chooses as enrollment fluctuates helps us estimate the costs involved. We use our reduced form estimates with a dynamic structural model of class size to estimate hiring/firing and marginal cost of adding a class. 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

Table 6: Quantile Regressions

	(1) Quantile 10%	(2) Quantile 50%	(3) Quantile 90%
lnClassSize	4.73	3.15	4.14
	[3.80,6.41]	[1.79,5.72]	[1.04,5.85]
lnClassSizeSQ	-0.79	-0.45	-0.71
	[-1.07,-0.62]	[-0.89,-0.19]	[-0.99,-0.20]

[1] We also control for other variables in the baseline regression, Female, Age, AgeSQ and sd of ln GPA.

[2] We do not include school fixed effects since we were unable to include such a large number of fixed effects.

[3] The brackets are the 95% confidence intervals.

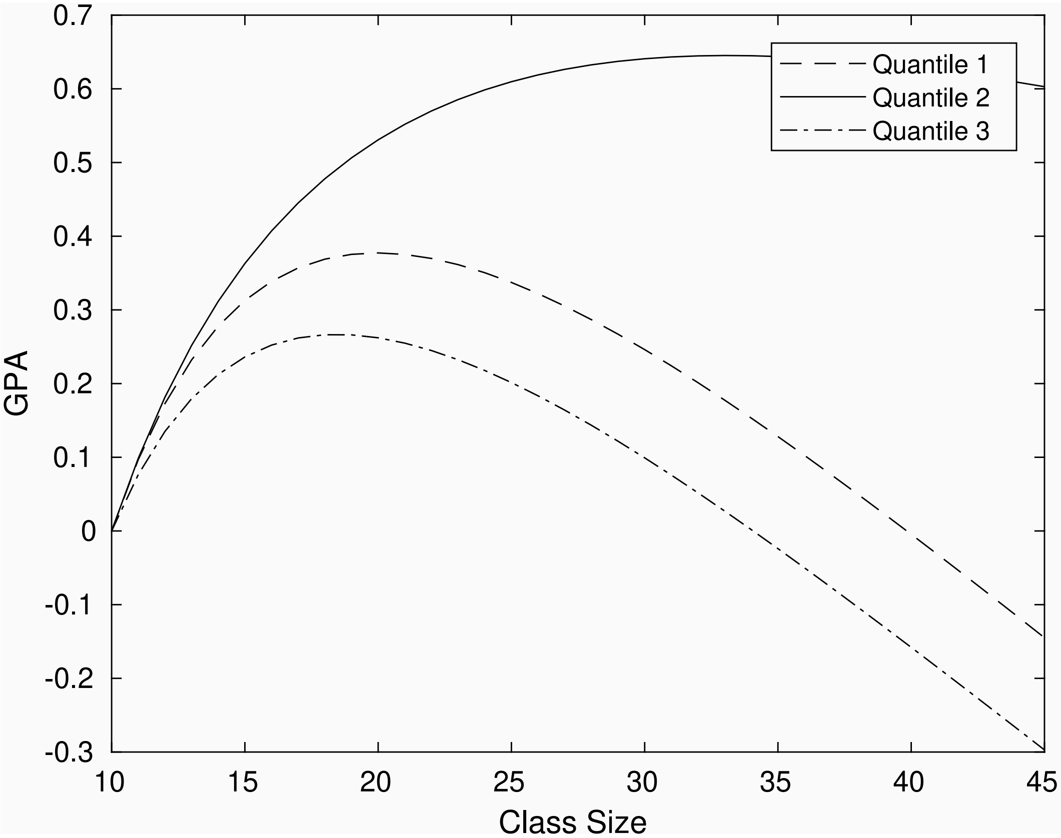

Figure 5: Quantile Effects of Class Size (Three quantiles)

high firing (and even maybe hiring) costs might be driving class size choices by government and the impact of this on student achievement if any. We find that unions, even if they raise costs and class size have a small effect on achievement.

6 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 various policies. For this reason we do not allow for quantile effects. Also, 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 12 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 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. We put roughly 25% in the small and large enrollment groups and 50% in the middle enrollment group.

The constraints the administrator faces are of two kinds. First, he faces the trade off we have estimated between class size and GPA and the enrollment process which is exogenously given to him. We can think of these as technical constraints.[18] Second, he 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 advertising the position, interviewing, and so on. The empirical transition probabilities are in Table 7. Note that schools tend to keep the same number of classes across years. This is

[18] 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 TFP process.

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 must consider the number of teachers he has, the enrollment today and the enrollment process he faces, as well as the range of costs and look forward to find his best decision today.

Table 7: Transition of Class Number

n_t n_{t-1}	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

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

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, so far we have the administrator's utility as

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

The administrator faces hiring and firing cost 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. We assume that this decison is made by the administrator in the 10th grade.[19] The administrator knows the realization of e_t and knows the state variable, n_{t-1}, and the random utility shock ε_{nt}, when he makes

[19] 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. This is presented in Table A.8 in Appendix G.

his choices. This shock is not observed by the econometrician. $\boldsymbol{\varepsilon}_t = \{\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t} ... \varepsilon_{10t}\}$ is a vector of shocks and each element is drawn from a type 1 generalized extreme value distribution. 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.

Thus, the administrators value function is:

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

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. From here on we may not explicitly condition on $\boldsymbol{\theta}$ as above, but it should be taken for granted.

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.

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

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 iid type 1 generalized extreme value distribution with variance σ^2, the probability of n_t is

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

6.1 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, $b = -5.36$ and $a = 29.0$. 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. Note that we have a choice in terms of the parameters to estimate: the weight on GPA or the variance of the utility shock since both cannot be separately identified. We choose to set the weight on GPA at unity as the variance of the utility shock is easier to interpret.

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

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 6 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 differentiated at 0. The higher the adjustment costs, the larger this region of inactionue.[20] 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.[21]

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 6 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 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 estimate the process for enrollment. Then estimate $\boldsymbol{\theta}$.

6.2 Estimating Enrollment

Ideally, we would want to estimate the enrollment process at the school level. However, this would give us at most 10 data points to work with. For this reason we decided to group schools by their mean enrollment and estimate an enrollment process for each group. We break the data into three

[20] 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.

[21] 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.

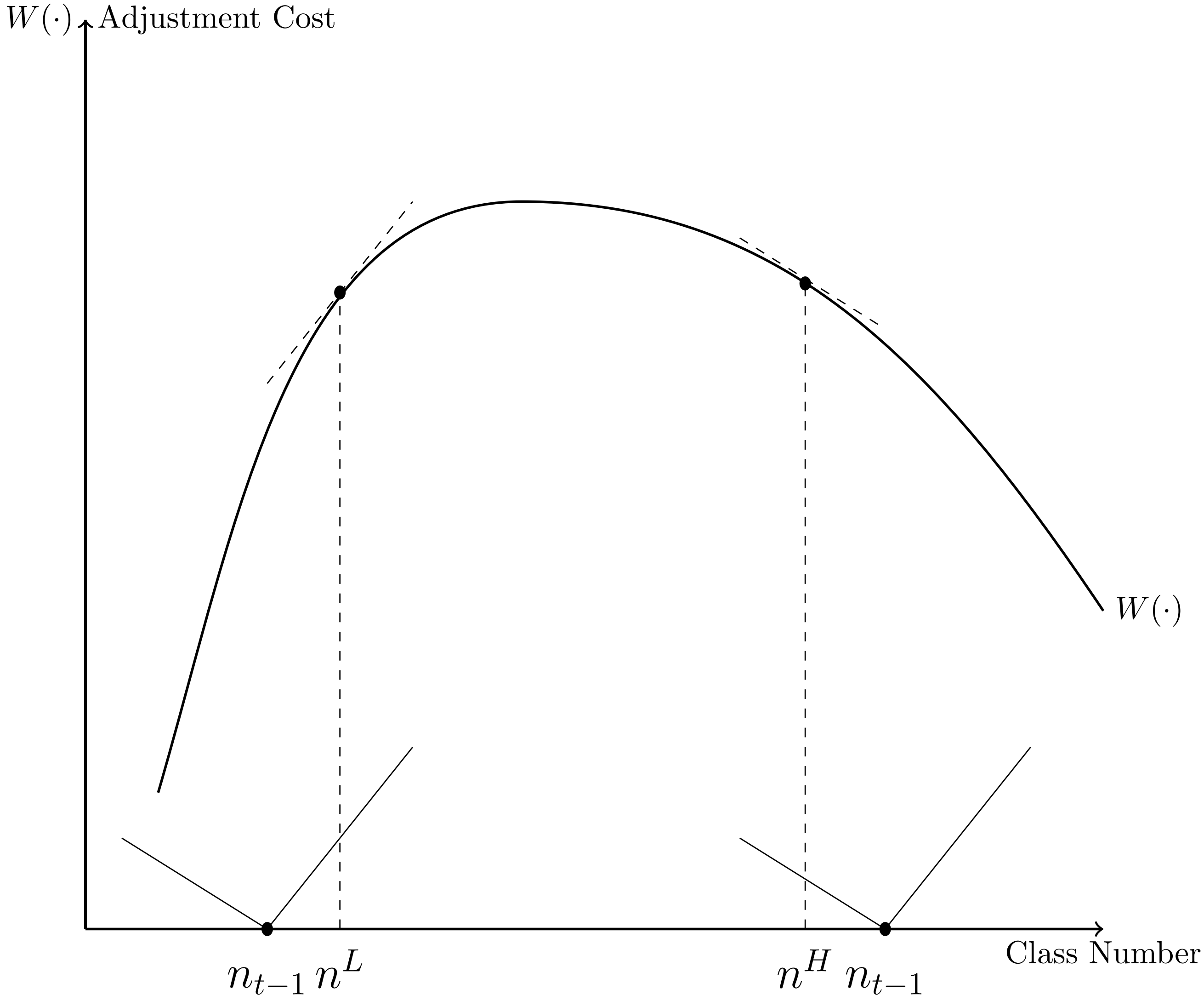

Figure 6: Identification of Adjustment Cost H, F

parts with schools in the lowest quartile in group 1 and the highest quartile in group 2. This gives us cutoffs for mean enrollment of 53 and 99.[22] Think of these as small schools with one class per grade, medium schools with one or sometimes two classes a grade, and large schools with two or more classes per grade.

We first test for whether the AR1 process we specify fits the data. We run the AR1 process separately for each of these three groups. Finally we test whether the estimates for the three groups differ from each other. We expect that γ_1 might be the same, but that γ_0 is likely to be lower for smaller schools. Our reason for expecting this is that these schools tend to stay in the same rough size groups, though their enrollment fluctuates year by year. The enrollment process is given by Equation (1).

If there was no random component, μ_t, then this enrollment process results in the data generated by it being on the straight line with slope less than 1 depicted in Figure 7. This would result in a steady state at point A in Figure 7. 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 7. The width of this band depends on the variance of μ_t. This will give a distribution of steady states in Figure 7. Note that in this case, schools will *not* tend to stay in their own rough groups over time.

What would be consistent with schools staying in their own group? If γ_0 was different (and higher for larger schools) even if γ_1 and the distribution of μ_t was the same across groups, then the process without a random component would be depicted by the lines in Figure 7. Note that as depicted, each group has a different steady state size interval. Adding back randomness would create bands around the lines as before and create a distribution of steady states for each group size, as shown in Figure 8. If these intervals overlapped, there could be some movement between groups in steady state. The estimates for the estimated enrollment process from the actual data for each group are presented in Table 8. Figure 9 depicts both the actual data and the estimated lines. Note that the actual data and lines look a lot like the simulated data. In particular, the slopes are not significantly different from one another while the intercepts differ significantly from each other. [23]

[22] We check if our results are sensitive to the number of groups used in estimating enrollment. We run the enrollment regression for 3,5 and 10 equally sized groups. These results are presented in Table A.11 and A.13 in Appendix H.

[23] One might ask why we choose three groups. With a long enough panel, we would be able to estimate the AR1 process at the school level. Since our panel is short, we group schools according to the average of cohort size. The estimates when we use 3, 5 and 10 groups of roughly equal size are presented in Appendix H. Notice that the more the groups, the estimated slope is slightly flatter. The estimates of the structural parameters corresponding to these estimates of the AR1 process are presented in Table A.10, Table A.12 and Table A.14 in Appendix H. It is reassuring to note that the structure parameters are relatively unaffected by the choices made in estimating the AR1 process.

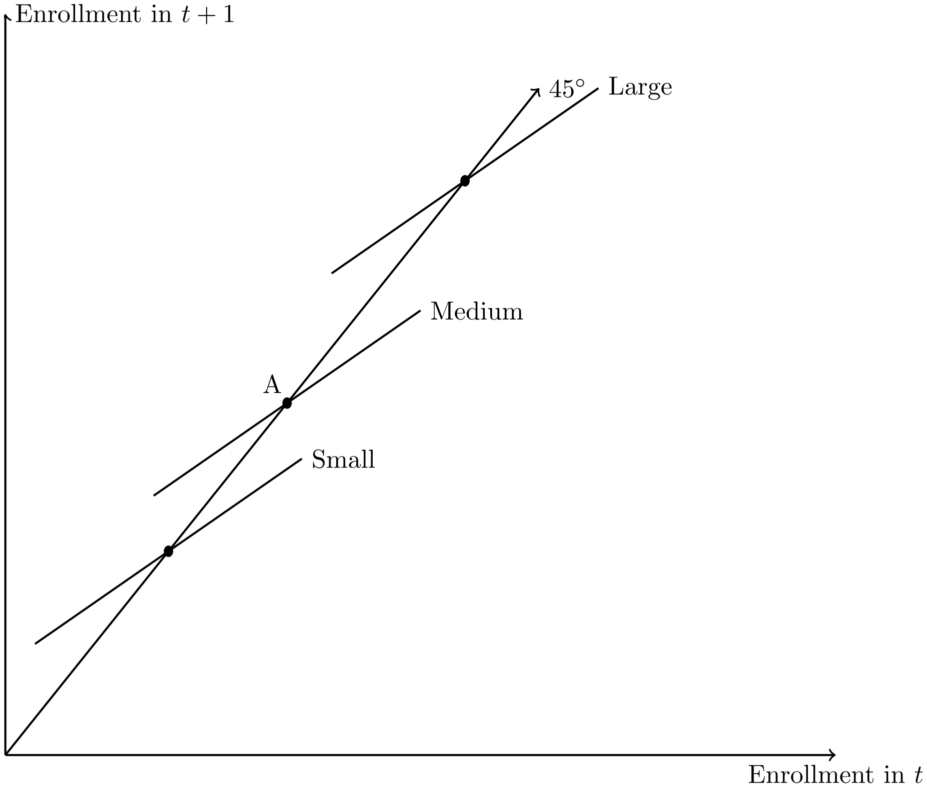

Figure 7: Enrollment Process

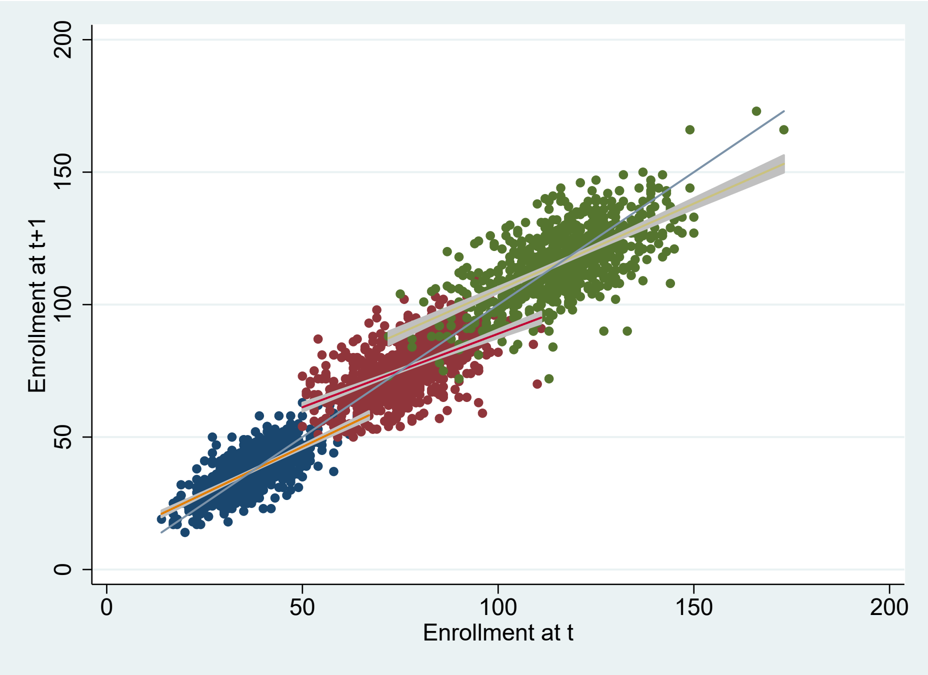

Figure 8: Enrollment Process (Simulated)

Table 8: Estimation of Enrollment Process

	<= 53	53 < *enrol* < 99	>= 99
γ_1	0.71	0.60	0.63
sd	(0.01)	(0.01)	(0.02)
γ_0	10.27	30.45	43.08
sd	(0.49)	(0.94)	(2.42)
N	264	535	244

(1) The standard errors are presented in parentheses.

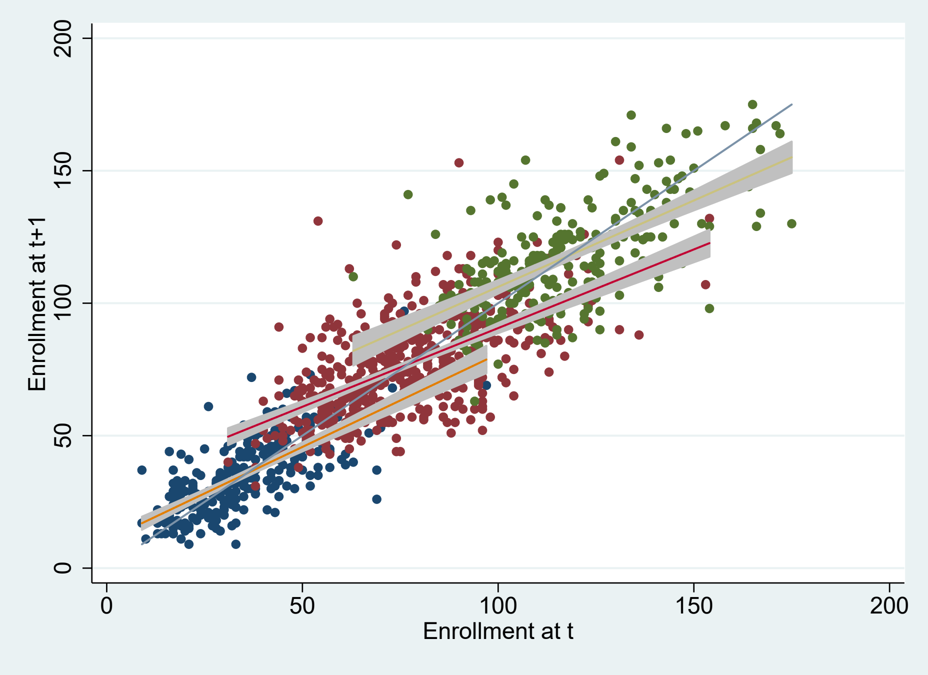

Figure 9: Enrollment Process (Data)

6.3 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.

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

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

since $\mathbb{E}_{\boldsymbol{\varepsilon}_t} \max_n f(n, \varepsilon_t) = \sum_n p(n) \mathbf{E}_{\boldsymbol{\varepsilon}_t}[f(n, \boldsymbol{\varepsilon}_t | n$ 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 that each is 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

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

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

Table 9: Estimation of the Structural Dynamic Model

Average Cohort Size	c	H	F	σ
All	171.30	99.06	156.49	140.24
sd	(51.59)	(42.90)	(42.41)	(19.82)
Euro	€20,572	€11,896	€18,793	

(1) The standard errors are presented in parentheses.

The estimates are presented in Table 9. 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.

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 (Stylianidou et al., 2004). H and F are the adjustment costs per class. To get the adjustment cost in Euros, we divide H 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 9. It costs €11,896 to add a new class and €18,793 to drop a class. The cost of dropping a class is much higher than the cost of adding a class. This is reasonable in Greece as firing a teacher is hard for public schools. The optimal class size in the absence of adjustment costs is 27.[24]

6.4 Counterfactual Exercises

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 presented in Table 10. 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 calculate the mean effects using the last 10 periods as the data is by then invariant to choice of starting point. The simulations show that class size and GPA change

[24] We also estimate the structural parameters when we use the enrollment regressions for 3,5 and 10 equally sized groups. These results are presented in Table A.12 and Table A.14 in Appendix H. As is evident, our estimates of the structural parameters do not change much.

for the different school groups as in Table 10. Reducing firing cost to zero raises average class size by about 4 students and reduces GPA by about a point (recall the scale was from 1-20) and by more for smaller schools than for larger schools. Since class size is larger, fewer teachers are hired, and thus, variable cost is lower. The total costs, variable and adjustment, fall relative to the status quo by 14 to 19%. Finally we calculate the welfare change. The total welfare increases by 1 to 3%.

Online platforms and internet have made hiring easier in a number of ways. Schools can post openings online and teachers can apply to multiple posts more easily than before. If we think of this as reducing hiring costs, we could ask what the effect might be of further improvements in the matching technology. These costs will never vanish, but could be reduced significantly. In the next counterfactual, we reduce hiring costs by half and evaluate the effects on class size and GPA. In both these exercises we are looking only at the partial equilibrium effects of these changes. These results are presented in Table 10. Reducing hiring cost decreases class size by 1 to 2 students as more teachers are hired and raises GPA slightly, and more so for smaller schools. The total cost goes up as schools hires more teachers (so that variable costs rise). As shown in the last column, total welfare, which includes the flow utility, adjustment costs and the continuation value, rises, though by under 1%.

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 by 4-7 students and GPA falls by more than a full point in all cases. The impact in terms of class size and GPA is larger for small schools than larger schools. The total cost goes up by 12 to 24% and welfare decreases. Note that costs overall rise by less than 50% as there are adjustments on the hiring and firing margins. It is well understood that quality teacher 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 class well before the cap is reached when the enrollment is more volatile and the adjustment cost is large. The effects of such caps are larger 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 more than 12 students, while large schools have class size falling by 7-8 students. Welfare falls by 16% for small schools and 6% for large ones while costs rise by 65% for small schools and 38% 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 8 for small schools and 2-3 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 this. In addition, we find that even caps which seem non binding have very significant impacts, especially for small schools.

Table 10: Counterfactuals

	Δ Average Class Size (#)	Δ Average GPA (points)	Δ Cost (%)	Δ Welfare (%)
		$F' = 0$		
Small	4.12	-0.90	0.81	1.03
Medium	4.08	-0.95	0.86	1.02
Large	3.59	-0.84	0.86	1.01
		$H' = 0.5H$		
Small	-1.75	0.35	1.06	1.01
Medium	-1.15	0.24	1.05	1.00
Large	-1.22	0.26	1.05	1.00
		$c' = 1.5c$		
Small	6.13	-1.35	1.12	0.90
Medium	5.16	-1.21	1.24	0.90
Large	4.71	-1.10	1.24	0.91
		Class Cap = 25		
Small	-12.19	1.57	1.65	0.84
Medium	-8.47	1.37	1.46	0.91
Large	-7.52	1.35	1.38	0.94
		Class Cap = 30		
Small	-9.36	1.58	1.37	0.91
Medium	-5.91	1.10	1.24	0.96
Large	-4.72	0.95	1.20	0.97
		Class Cap = 35		
Small	-8.10	1.46	1.23	0.95
Medium	-3.49	0.70	1.15	0.98
Large	-2.50	0.54	1.08	0.99

7 Conclusions

Our work shows a clear hump shaped relationship between class size and GPA. Moreover, the hump shape remains even when we allow for quantile effects. In addition, there is some evidence that class size matters more for weaker students. 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 linear 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 finds 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 are reasonable and suggest that reducing firing costs actually hurts achievement. Teachers' unions may not be as pernicious as might be thought. Reducing hiring cost, which might be done at low cost given the web has reduced search costs, decreases class size and thus, improves students' achievement since most class sizes are above the turning point of 15. 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.

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.

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.

Ghanem, Dalia, "Testing identifying assumptions in nonseparable panel data models," *Journal of Econometrics*, 2017, *197* (2), 202–217.

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.

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.

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.

Lee, Sokbae, "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 2007, *141* (2), 1131–1158.

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, "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.

Stylianidou, F, G Bagakis, and D Stamovlasis, "Attracting, developing and retaining effective teacher," 2004.

Todd, Petra E. and Kenneth I. Wolpin, "On the specification and estimation of the production function for cognitive achievement*," *The Economic Journal*, 2003, *113* (485), 3–33.

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

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 lightly 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

B Parametric Estimation Restricting to Small Schools

One of the concerns is that the sample of small class is not big enough. Figure A.1 plots the distribution of class size in the data.

Figure A.2 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.

We use the average enrollment to define the size of schools. A school is a small school if the average enrollment over years is less than 30, 50 or 70. We try different cutoffs to show the robustness. Table A.2 presents the estimation results.

C Testing the time-homogeneity assumption

Assumption 1 (time-homogeneity)**.**

$$\varepsilon_{jt} \mid \boldsymbol{CS}_j,\ \alpha_j \sim F(. \mid \boldsymbol{CS}_{j,}\alpha_j).$$

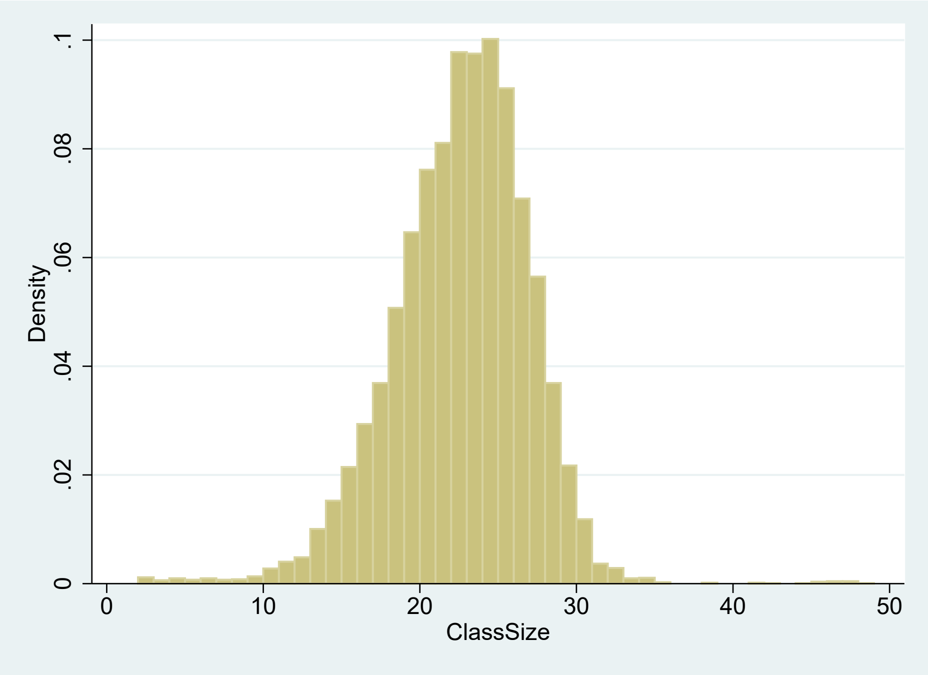

Figure A.1: The Distribution of Class Size in Each Bin

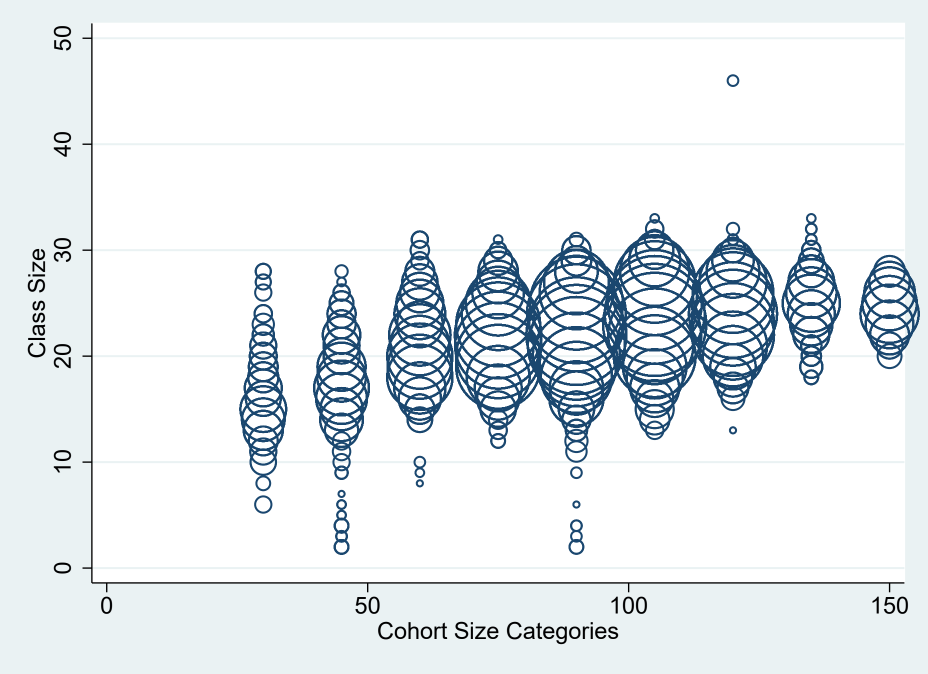

Figure A.2: The Relationship between Cohort Size and Class Size

Table A.2: Parametric Estimation of Linear Class Size with IVs Restricting to Small Schools

	(1)		(2)		(3)	
	$\leq$ 30		$\leq$ 50		$\leq$ 70	
			Second Stage			
lnClassSize	39.1		32.2		42.6	
	(29.8)		(17.0)*		(22.7)*	
lnClassSizeSQ	-7.21		-6.05		-7.94	
	(5.3)		(3.2)*		(4.2)*	
Female	1.35		1.09		0.94	
	(0.3)***		(0.1)***		(0.08)***	
Age	-1.56		-2.13		-2.32	
	(0.5)***		(0.3)***		(0.2)***	
AgeSQ	0.025		0.035		0.038	
	(0.009)***		(0.005)***		(0.004)***	
sd of ln GPA	-5.79		-3.08		-3.82	
	(2.7)**		(1.3)**		(1.0)***	
Kleibergen-Paap Statistic	84.3		57.4		65.9	
p-value	0.000		0.000		0.000	
School FE	YES		YES		YES	
R-sq	0.095		0.062		0.025	
N	667		4678		10426	
			First Stage			
	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ
lnU	1.06	5.88	0.61	3.03	0.60	3.10
	(0.1)***	(0.7)***	(0.09)***	(0.5)***	(0.09)***	(0.4)***
lnUSQ	-0.78	-3.88	-0.21	-0.94	-0.20	-1.00
	(0.3)**	(2.0)*	(0.07)***	(0.3)***	(0.04)***	(0.2)***

[1] Standard deviations are clustered at class level. *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Ghanem (2017) derived testable equality restrictions for the time-homogeneity Assumption 1. She therefore proposed a statistical test based on Kolmogorov-Smirnov and Cramer-von-Mises statistics. Below, we explain the intuition of the test. For the sake of simplicity, suppose we only have two periods. As mentioned in Chernozhukov et al. (2013), the time-homogeneity assumption is equivalent to $(\epsilon_{jt}, \alpha_j)|CS_{j.} =^d (\epsilon_{j1}, \alpha_j)|CS_{j.}$, for all t. Then this assumption implies that the conditional distribution of the second period average GPA for a school j is the same as its conditional distribution of the first period average GPA given its history of class size choices $CS_{j.} = (x, x')$. Indeed, we have:

$$
\begin{aligned}
&(\epsilon_{j2}, \alpha_j)|CS_{j.} = (x, x') =^d (\epsilon_{j1}, \alpha_j)|CS_{j.} = (x, x') \\
&\Rightarrow g(x', \alpha_j, \epsilon_{j2})|CS_{j.} = (x, x') =^d g(x, \alpha_j, \epsilon_{j1})|CS_{j.} = (x, x') \\
&\Rightarrow g(CS_{j2}, \alpha_j, \epsilon_{j2})|CS_{j.} = (x, x') =^d g(CS_{j1}, \alpha_j, \epsilon_{j1})|CS_{j.} = (x, x') \\
&\Rightarrow GPA_{j2}|CS_{j.} = (x, x') =^d GPA_{j1}|CS_{j.} = (x, x'),
\end{aligned}
$$

where $=^d$ means equal in distribution.

Testing procedure: bootstrap

Let T_N be a test statistic. The following summarizes the steps of the test.

1. Compute the statistic T_N for the original data $\{(GPA_{1.}, CS_{1.}), \ldots, (GPA_{N.}, CS_{N.})\}$.
2. Resample N observations $\{(GPA^*_{1.}, CS^*_{1.}), \ldots, (GPA^*_{N.}, CS^*_{N.})\}$ with replacement from the original data. Compute T^b_N, the centered statistic for the bth bootstrap sample.
3. Repeat points 1. and 2. B times.
4. Calculate the p-values of the tests with $p = \frac{1}{B}\sum_{b=1}^{B} 1\{T^b_N > T_N\}$. Reject if p-value is smaller than some significance level α.

For the implementation of the test, we set $B = 500$. We use the Kolmogorov-Smirnov and Cramer-von-Mises statistics (See (Ghanem, 2017) for details on the formulas). All p-values are higher than 10%, suggesting that the identifying Assumption 1 is not rejected at any 1%, 5% nor 10% significance levels. The p-values for the standard parallel trend assumption are 0.99 and 0.90 for the Kolmogorov-Smirnov and Cramer-von-Mises statistics, respectively.

D Nonparametric Estimates

Table A.3: Nonparametric results

s_0 s_1	12		13		14		15		16		17	
	δ_{12}	δ_{23}	δ_{12}	δ_{23}	δ_{12}	δ_{23}	δ_{12}	δ_{23}	δ_{12}	δ_{23}	δ_{12}	δ_{23}
21	4.99	-0.24	3.66	-0.25	2.35	-0.24	1.58	-0.24	1.04	-0.21	0.69	-0.23
	$(1.05)^{*}$	$(0.07)^{*}$	$(0.86)^{*}$	$(0.07)^{*}$	$(0.68)^{*}$	$(0.07)^{*}$	$(0.53)^{*}$	$(0.07)^{*}$	$(0.44)^{*}$	$(0.07)^{*}$	$(0.32)^{\dagger}$	$(0.07)^{*}$
22	4.94	-0.18	3.61	-0.19	2.31	-0.19	1.55	-0.18	1.02	-0.16	0.67	-0.17
	$(1.06)^{*}$	$(0.07)^{*}$	$(0.87)^{*}$	$(0.07)^{*}$	$(0.69)^{*}$	$(0.07)^{*}$	$(0.53)^{*}$	$(0.07)^{*}$	$(0.44)^{*}$	$(0.07)^{\dagger}$	$(0.32)^{\dagger}$	$(0.07)^{*}$
23	4.93	-0.19	3.59	-0.19	2.30	-0.19	1.54	-0.18	1.01	-0.17	0.65	-0.18
	$(1.05)^{*}$	$(0.07)^{*}$	$(0.86)^{*}$	$(0.07)^{*}$	$(0.68)^{*}$	$(0.07)^{*}$	$(0.53)^{*}$	$(0.07)^{*}$	$(0.44)^{\dagger}$	$(0.07)^{*}$	$(0.32)^{\dagger}$	$(0.07)^{*}$
24	4.89	-0.19	3.55	-0.20	2.27	-0.20	1.52	-0.19	0.99	-0.18	0.64	-0.18
	$(1.05)^{*}$	$(0.08)^{*}$	$(0.86)^{*}$	$(0.08)^{*}$	$(0.68)^{*}$	$(0.08)^{*}$	$(0.53)^{*}$	$(0.07)^{*}$	$(0.44)^{\dagger}$	$(0.07)^{*}$	$(0.31)^{\dagger}$	$(0.08)^{\dagger}$

(1) †, * indicate significance at the 5% and 1% levels, respectively.

E Full First Stage Estimation

Table A.4 reports the full first stage estimation for Table 3. Table A.5 reports the full first stage estimation for Table 4.

Table A.4: Full First Stage Estimation for Table 3

	(1)	(2)
	First Stage	
	lnClassSize	lnClassSize
lnU	0.25	0.24
	(0.01)***	(0.01)***
Female	0.0022	0.0023
	(0.001)**	(0.001)**
Age	-0.022	-0.022
	(0.007)***	(0.007)***
AgeSQ	0.00033	0.00034
	(0.0002)*	(0.0002)*
sd of ln GPA		0.039
		(0.04)
School FE	YES	YES
R-sq	0.449	0.449
N	81845	81845

(1) †, * indicate significance at the 5% and 1% levels, respectively.

F The Restricted Sample when Teacher Data is Available

Table A.6 presents the summary statistics for the sample with teacher data available, as well as the difference between the restricted sample and the full sample.

Table A.7 presents the analogs as in Table 4.

G The Decision of Class Number

We look at the change in number of classes with respect to the change in cohort size. The baseline regresses the change in class number to the change in cohort size in the 10th grade over years. We compare this to the change in number of classes with respect to the change in cohort size from 10th to 11th grade and 11th to 12th grade.

Table A.5: Full First Stage Estimation for Table 4

	(1)	(2)	(3)	(4)
		First Stage		
	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ
lnU	0.45	2.22	0.45	2.22
	(0.06)***	(0.3)***	(0.06)***	(0.3)***
lnUSQ	-0.065	-0.24	-0.064	-0.24
	(0.02)***	(0.10)**	(0.02)***	(0.10)**
Female	0.0023	0.013	0.0023	0.013
	(0.001)**	(0.006)**	(0.001)**	(0.006)**
Age	-0.023	-0.16	-0.024	-0.16
	(0.006)***	(0.03)***	(0.006)***	(0.03)***
AgeSQ	0.00036	0.0026	0.00037	0.0026
	(0.0002)**	(0.0008)***	(0.0002)**	(0.0008)***
sd of ln GPA			0.037	0.19
			(0.04)	(0.2)
School FE	YES	YES	YES	YES
R-sq	0.453	0.457	0.453	0.457
N	81845	81845	81845	81845

(1) †, * indicate significance at the 5% and 1% levels, respectively.

Table A.8 shows that schools' decision on class number is more sensitive to the coming cohort size in 10th grade, while the class number remains relatively stable when the same cohort go to 11th and 12th grades.

H Structural Estimation with Finer Groups

Table A.9, Table A.11 and Table A.13 present the estimation of enrollment when schools are divided into 3, 5 or 10 categories evenly based on the average cohort size over time. Table A.10, Table A.12 and Table A.14 present the structural estimation of c, H, F, respectively.

Table A.6: Summary Statistics of the Sample with Teacher Data

	(1)	(2)	(3)
	Full	Sample with Teacher Data Available	(1) - (2)
		Individual Level Data	
gpa_notrack	11.79	11.88	-0.0916
	(3.79)	(4.11)	(-1.07)
female	0.54	0.55	-0.00600
	(0.50)	(0.50)	(-0.54)
Age	15.97	16.03	-0.0579***
	(0.60)	(0.51)	(-4.11)
N	81845	2031	
		Class Level Data	
ClassSize	22.62	19.75	2.872***
	(4.15)	(3.39)	(6.96)
N	3641	103	3744
		School Level Data	
CohortSize	76.17	52.15	24.01***
	(33.90)	(33.26)	(4.35)
ClassNo	3.37	2.64	0.728***
	(1.24)	(1.46)	(3.57)
N	1082	39	

(1) †, * indicate significance at the 5% and 1% levels, respectively.

Table A.7: Parametric Estimation of Linear Class Size with IVs for the Sample with Teacher Data Available

	(1)		(2)	
	Second Stage			
lnClassSize	68.7		41.5	
	(41.3)*		(39.8)	
lnClassSizeSQ	-11.6		-7.04	
	(6.8)*		(6.6)	
Female	0.96		0.94	
	(0.2)***		(0.2)***	
Age	7.10		7.09	
	(2.9)**		(2.9)**	
AgeSQ	-0.25		-0.25	
	(0.09)***		(0.09)***	
sd of ln GPA			-7.50	
			(1.6)***	
Kleibergen-Paap Statistic	3.1		3.1	
p-value	0.080		0.077	
School FE	YES		YES	
R-sq	0.108		0.122	
N	2031		2031	
	First Stage			
	lnClassSize	lnClassSizeSQ	lnClassSize	lnClassSizeSQ
lnU	0.059	0.070	0.059	0.076
	(0.3)	(1.9)	(0.3)	(1.8)
lnUSQ	0.15	0.99	0.15	0.98
	(0.10)	(0.6)*	(0.09)	(0.5)*

(1) †, * indicate significance at the 5% and 1% levels, respectively.

Table A.8: The Change in Class Number with Respect to the Change in Cohort Size

	(1) dClassNo	(2) dClassNo
dlnCohortSize	1.53	
	(0.06)***	
10th to 11th grade × dlnCohortSize	-0.44	
	(0.1)***	
11th to 12th grade × dlnCohortSize	-0.52	
	(0.1)***	
dCohortSize		0.031
		(0.0009)***
10th to 11th grade × dlnCohortSize		-0.0100
		(0.002)***
11th to 12th grade × dlnCohortSize		-0.012
		(0.002)***
10th to 11th grade	0.040	0.053
	(0.02)*	(0.02)***
11th to 12th grade	-0.057	-0.058
	(0.02)***	(0.02)***
R-sq	0.255	0.350
N	3247	3247

(1) †, * indicate significance at the 5% and 1% levels, respectively.

Table A.9: Estimation of Enrollment Process with Three Even Categories of School Size

	$<= 53$	$53 < enrol < 99$	$>= 99$
γ_1	0.71	0.60	0.63
sd	(0.01)	(0.01)	(0.02)
γ_0	10.27	30.45	43.08
sd	(0.49)	(0.94)	(2.42)
N	264	535	244

(1) The standard errors are presented in parentheses.

Table A.10: Estimation of the Structural Dynamic Model with Three Even Categories of School Size

	c	H	F	σ
	156.90	96.92	129.97	130.19
sd	(40.96)	(34.43)	(33.70)	(16.86)
Euro	€20,572	€12,707	€17042	

(1) The standard errors are presented in parentheses.

Table A.11: Estimation of Enrollment Process with Five Categories of School Size

Enrollment Quantile	0-20	20-40	40-60	60-80	80-100
γ_1	0.62	0.46	0.45	0.29	0.63
sd	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
γ_0	11.64	31.69	41.27	65.56	44.81
sd	(0.64)	(1.53)	(2.26)	(1.87)	(2.61)

[1] The standard errors are presented in parentheses.

Table A.12: Estimation of the Structural Dynamic Model with Five Categories of School Size

	c	H	F	σ
	167.88	80.10	133.19	120.79
sd	(48.21)	(36.40)	(34.49)	(16.58)
Euro	€20,572	€9,816	€16,322	

[1] The standard errors are presented in parentheses.

Table A.13: Estimation of Enrollment Process with Ten Categories of School Size

Enrollment Quantile	0-10	10-20	20-30	30-40	40-50
γ_1	0.43	0.42	0.46	0.21	0.32
sd	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
γ_0	13.80	21.69	28.58	50.27	48.90
sd	(1.09)	(1.43)	(2.03)	(2.72)	(2.88)
Enrollment Quantile	50-60	60-70	70-80	80-90	90-100
γ_1	0.51	0.11	0.34	0.26	0.53
sd	(0.05)	(0.04)	(0.02)	(0.05)	(0.04)
γ_0	38.12	78.16	63.87	78.96	62.93
sd	(3.88)	(3.71)	(2.29)	(5.56)	(5.30)

[1] The standard errors are presented in parentheses.

Table A.14: Estimation of the Structural Dynamic Model with Ten Categories of School Size

	c	H	F	σ
	176.68	55.81	120.34	103.55
sd	(46.14)	(28.87)	(26.76)	(13.79)
Euro	€20,572	€6,498	€14,012	

[1] The standard errors are presented in parentheses.