ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Third-year Bachelor's thesis - International Bachelor's in Economics and Business Economics

Martin Odinot (560088)

21 of July 2023

Third-year (Final version) Bachelor's thesis

International Bachelor's in Economics and Business Economics

Thesis supervisor: Dinand Webbink

Do primary school students perform better in small classes than in large classes?

Evidence from France

0. Abstract

Class size reduction policies are often used in France. Yet, researchers did not reach a consensus with regard to the benefits of such a policy on student achievement. I collected a large amount of data (on France) and used regression analysis to study the relationship between class size and student achievement. I do not have enough evidence to conclude that changes in class size affect the performance of students in reading. Nevertheless, I found that students in small classes perform on average better than students in large classes (higher scores of 0.45% - statistically significant at a 10% confidence level), keeping everything else constant. Those are the overall results but I also did subsample analyses. I tried to see differences in test scores between students in small and large classes in disadvantaged schools and among disadvantaged students. I did not have enough evidence to conclude disparities in test scores between small and large classes for those groups. Overall, those results are disappointing. Either the coefficient of interest is low or either it is not statistically significant. Nevertheless, my models suffer from omitted variable bias and I suspect that including variables such as past test scores and migration background would have made increased the magnitude of the (positive) effect of class size on student performance. Hence, this study should not discourage French policymakers to pursue class reduction policies.

1. **Introduction**

Class size reduction is a commonly used policy to improve the learning outcomes of students. In the paper of Krueger et., al (2002), we learn that in 1996, the state of California which had the largest average class size in the USA implemented a maximum class size rule of 20 students for its third graders. One of the conclusions drawn from this event is that class size reduction policies should be made if and only if there is sufficient evidence in the scientific literature about the positive effects of a decrease in class size which was not the case at this time. Typically, California experienced a problem: as demand for teachers rose massively because of this measure, the quality of teachers decreased which was harmful to students. The scores on standardized tests of K-3 students increased after this measure but one cannot be sure that it was caused by it since many other educational policies were implemented at this time. A massive amount of researchers started to work on the data of this study and some came to the conclusion that a class size reduction was an efficient policy in terms of test score gains while others said class size had no significant effect. The *class size debate* was at its peak at this time (Krueger et al., 2002).

According to Ehrenberg et., al (2001), “Class size refers to the actual number of pupils taught by a teacher at a particular point in time”. Many papers do not talk about the effect of class size on student achievement but about the effect of pupil/teacher ratio which is a different measure: “The calculation of a pupil/teacher ratio typically includes teachers who spend all or part of their day as administrators, librarians, special education support staff, itinerant teachers, or other roles outside the classroom” (Ehrenberg et al., 2001). In this paper, I will solely deal with class size.

I chose in this study to focus on France. In terms of education, this country mainly suffers from 2 issues: the low level of its students[1] and inequalities in schooling[2]. To counter these, the President Emmanuel Macron in 2017 decided to divide the class sizes by 2 in the first and second grades classes that were in priority zones (REP+). No empirical studies have been made about this policy. However, the proportion of students with extreme difficulties in mathematics and reading in those zones went from 40% in 2017 to 30% in 2022 (Europe 1, 2022). Of course, we cannot say that this decrease is solely caused by class size reduction since many policies have been implemented. Nevertheless, many judge this result insufficient with regard to the high cost of the policy (500 million euros). Teacher unions make pressure on the government to expand this policy to other grades. Accounting to 2024, the same policy will take place for kindergartens in disadvantaged districts. It might be the case that other grades level will be concerned in the future so policy-makers should be sure that class size reduction policies are efficient. This is why I chose to study the class size effect in France.

One can philosophically adhere to these policies because they attempt to decrease the inequalities between pupils in disadvantaged districts compared to pupils in advantaged districts. However, are these class size reduction policies efficient? In other words, is the

[1] At PISA 2023, France was ranked at the 23rd place out of 79 countries (Académie de Créteil, 2023)

[2] The results of PISA (2012) showed that the impact of social origins on student performance was the largest in France compared to all other OECD countries (Government, 2021)

increase in test scores (and hence in future productivity/earnings) of students caused by the class size reduction sufficiently high to cover the cost of the policy?[3]

It seems that there is no consensus among scientists about the benefits (in terms of student performance) of class size reduction (see literature review). Typically, in France, 4 of the 7 correlational studies found no significant effect of class size on student performance (E. Caubel, 2019). One criticism that can be addressed to those studies is that there is little variation in class size in their dataset (most of the students are in classes with 24 to 26 students). This makes it hard to capture the effect of a class size decrease on student performance. In my data, the variation in class size is more important[4]. As French researchers, I will regress the student's reading scores on class size and other control variables. As most of them, I do not obtain a statistically significant relationship.

I will also compare the results of classes composed of less than 25 students (I call them small classes) with classes that have more than 25 students (I call them large classes). My main research question will be:

"Do students in small classes perform better (in reading) than students in large classes?"

Of course, many control variables will be added to this regression. However, the result I obtain cannot be interpreted causally because of endogeneity (or omitted variable bias). I arrive at the conclusion that students in small classes perform significantly better (at 10% confidence level) in their reading test than students in large classes, keeping everything else constant. Nevertheless, I cannot conclude that they perform better because they are in small classes.

In this paper, I will first derive the main methods and results of the class size literature (2). I will then present my dataset and the method I will be using (3). I will then deal with my results (4) and see if they are robust to changes in my model setting (5). I will then expose my suggestions for future studies and explain the caveats of my paper (6). Lastly, I will sum up my results (7).

[3] Class size reduction policies are very expensive. New teachers must be hired, new schools must be built…

[4] I am the first to exploit an international test (PIRLS) to study class size effect in French research. I have also the largest dataset among the French researchers - the highest one was in Piketty (2004) study: more than 9,000 students

2. Literature review

The relationship between class size and student performance is heavily debated in the scientific literature. While some studies find a positive effect (negative coefficients) of lower class size on pupil achievement, others find no effect and more rarely negative effects (positive coefficients). Identifying a causal relationship between class size and achievement is hard because of selection bias. Indeed, weaker students tend to be allocated to larger classes, which means that in general, students in large classes tend to perform better on average than students in small classes. In this section, I will first explain why schools tend to allocate weaker students to small classes with the Lazear Model (2.1). I will then explain the main methods used in the scientific literature to eliminate the selection bias, namely random experiment (2.2), quasi-experiment (2.3), fuzzy regression discontinuity design (RDD) using maximum class size rules (2.4), and multiple regression (2.5).

2.1. The optimal class size: Lazear model (1999)

Lazear (1999) made a theoretical model that predicts that the most disruptive students should be assigned by their schools to small classes. He explains that if a student is disruptive, he creates a negative externality to his classmates because the teacher interrupts the class to put the order back. In his model, students have a probability p to disrupt the class during the year. The decision of class allocation belongs to the school administration and depends on this probability. An optimal behavior would be to assign the best (the less disruptive) students to large classes because they would less enjoy a class reduction than disruptive pupils. In other words, the well-behaved students should be assigned to large classes and the badly-behaved students should be assigned to small classes (Lazear, 1999).

This theory has important implications for empirical research. If one compares the average results of students in a large class with the average results of students in a small class, he might find that the large class on average performs better than the small class. Serious research should fight against the selection bias and make sure that students in small and large classes have the same characteristics on average. Randomized experiments try to achieve this goal.

2.2. Randomized experiment – evidence from the STAR experiment (1985)

The only randomized experiment of class size allocation was the STAR (Student-Teacher Achievement Ratio) experiment operated by the state of Tennessee. The study setting has been described by Ehrenberg et., al (2001). In 1985, the state began this 12$ million dollars project. Schools from every district could participate on the only condition that they met a few requirements[5]. A hundred schools were selected. Pupils starting kindergarten as well as teachers were randomly assigned by these schools to one of the three groups: a class with 13 to 17 students, a class with 22 to 26 students, and a class with 22 to 26 students with a teaching assistant. Each student stayed in one of the groups for four years before going back to a regular class. At the end of each academic year, pupils did a standardized test (Ehrenberg et., al, 2001).

According to Krueger et., al (2002), this study revolutionized the mindsets of scientists and politicians about educational public spending - especially in class size reduction - in the United States. Indeed, in the three last decades of the XXth century, educational spending had increased "more slowly than people believed" (Krueger et., al, 2002). For them, it can be explained by the fact that at this period the "money makes no difference" theory, defended by the economist Hanushek was predominant (Krueger et al., 2002). This researcher collected results from hundreds of papers published in academic journals and showed that the number of positive effects of lower class size on achievement exceeded slightly the number of negative effects (42% against 38%) (Hanushek, 1999).

However, the STAR experiment was much more credible than all the previous papers. Mosteller (1999) said it was: "one of the greatest education experiments in education in the United States history" (Mosteller, 1999). Randomized experiments are indeed the most serious approach to doing group comparisons. If individuals are randomly assigned to groups, their characteristics are on average similar so one can be sure that the difference in outcome (score) is solely due to the treatment variable (class size).

Many researchers worked on the STAR data and found positive effects of reducing class size on student performance. Finn and Achilles (1999) sum up the results of all the studies done with the STAR data and concluded that there was "an array of benefits of small classes, including improved teaching conditions, improved student performance, and, after the experimental years, improved student learning behaviors, fewer classroom disruptions and discipline problems, and fewer student retention" (Finn and Achilles, 1999).

Still, Hanushek and other economists were not convinced by the benefits of reducing class size. In 2002, Krueger (a pro-reduction class size investment) debated with Hanushek on the benefits of class size reduction. For Krueger (2002), a good null hypothesis should not be to test if the effect of class size on student performance is zero, but rather, to test if the present value of delayed benefits (increase in future wages because of higher scores due to a decrease in class size) exceeds the costs of class size decrease. In his model, he estimates the increase in wages associated with a standard-deviation unit rise in test score (8%). Also, he incorporates the results of Finn and Achilles: splitting size by 2 conducts to a rise in performance by 20% of a standard deviation unit (Finn & Achilles, 1990). He then discounts the future benefits and compares them with the cost of the operation. He concludes that in this experiment set, the net present value is positive and high. However, Hanushek criticizes his model built with "heroic assumptions". He concludes: "The existing evidence suggests that

[5] They had to have enough students to assign them to the three groups. They also had to accept that those students took a standardized test each year. They had to accept to randomly allocate teachers and students to the three groups.

any effects of overall class size reduction will be small and very expensive" (Krueger et al., 2002).

Moreover, Krueger is convinced about the internal and external validity of the experiment (Krueger et al., 2002) which is not the case of Hanushek and Hoxby. Hoxby (1998) points out the "Hawthorne effect" argument. As schools know that a successful result could lead to policy reforms to their advantage, they have incentives to change their behavior (Hoxby, 1998). Hanushek says legitimately that there is no evidence that the schools randomly allocated their students and teachers to the three groups (Hanushek, 1999).

Hence, we saw that this experiment has been praised and criticized. If random experiments are the methods with the highest causal interpretation, they are very costly to operate. Hence, researchers often work with other designs such as quasi-experiments.

2.3. Quasi-experiments

We call "quasi-experiments", experiments that lack random assignment. One example is the California CSRP (Class Size Reduction Program) which began in 1996. The state reduced the maximum class size from 33 to 20 students per class. One would have wished to compare the test scores before and after the implementation of the policy. Unfortunately, pupils did not take any standardized test before the intervention (Ehrenberg et., al, 2001). The experiment has been analyzed by Bohrnstedt & Stecher (1999) three years after its start. They compared the scores between students in classes with less than 20 students and students in classes with more than 20 students while controlling for socio-economic variables. They came to the conclusion that students in small classes performed on average better than students in large classes, keeping everything else constant (Bohrnstedt & Stecher, 1999). However, the coefficient they found was not statistically significant. This method of estimation is closely related to the one I will use since I will compare scores of students in classes of less than 25 students with test scores of students in classes with more than 25 pupils.

Another example of quasi-experiment is the one conducted by the French Ministry of Education in 2002. 10 academies with the highest amount of priority schools (REP+) were asked to propose a hundred classes of first grade in schools where at least 50 % of third-graders were in the bottom 20% at a national test. Originally, 2000 students have been selected. There would have been 101 treatment classes (with 8 to 12 students) and 99 control classes (with the regular amount of classes – 24 or more students). Unfortunately, because of non-compliance and administrative issues, the sample consisted of only 454 students, 230 of them assigned to treatment classes and 224 assigned to control classes (Direction de l'évaluation, de la prospective et de la performance, 2005). This experiment has many caveats, mostly the small sample size and the non-random assignment of students to treatment and control classes. Still, the study has been analyzed by Bressoux and Lima in 2011: the split of class (to 24 to 12 students) leads to an increase in test scores by 2% (statistically significant) of a standard deviation unit (Bressoux & Lima, 2011).

We saw that quasi-experiments were not really convincing because of the non-random allocation of students to classes which causes selection bias to be present. If some assumptions hold, fuzzy RDD can be a credible approach to estimate a causal effect of lower class size on student performance.

2.4. Fuzzy Regression Discontinuity Design using maximum class size rule

Many countries have a maximum class size rule. Some authors have used a regression discontinuity design to see if there is a causal effect of class size on achievement. We call those RDD "fuzzy" (and not sharp) because one class size is partly (and not fully) determined by the maximum class size rule. In other words, schools are not formally obliged to respect the rule but many comply with it. The method of estimation is closely related to an instrumental variable (IV) setting. The authors use what we call a "two-staged least square" estimation. For the first stage, they regress the actual class size of individual i on the theoretical class size of individual i. This theoretical class size depends on two factors: the maximum class size rule and the number of students enrolled in grade *x* in the school. If for example the maximum class size rule is 20 and there are 45 students enrolled in the school of individual i, the function would predict that he is in a class of 15 students (there should be 3 classes of 15 in his school). In the second stage, they regress the score of the student on the predicted actual class size (based on the theoretical class size) drawn in the first stage. This coefficient can be interpreted as causal if 3 assumptions hold:

(i) Strong first stage: there should be a strong relation between the instrument (theoretical class size) and the variable of interest (actual class size).

(ii) Independence: the theoretical class size should not be correlated with the error term.

(iii) Exclusive restriction: the theoretical class size should not directly affect the score.

Angrist and Lavy were the first to adopt this method in 1999. They worked on Israel which "Maimonides rule" (a 12th century Rabbinic scholar) states that the maximum class size should be 40 students. The two authors came to the conclusion that class size reduction causes a "significant and substantial increase in test scores for fourth and fifth graders, although not for third graders" (Angrist and Lavy, 1999)

One year after, Hoxby (2000) tried to use the same method for Connecticut's third and fifth graders. There was no formal maximum size there but he observed an "implicit" maximum class size rule of 25. He operated the RDD and concluded: "the estimates indicate that class size does not have a statistically significant effect on student achievement" for third and fifth graders (Hoxby, 2000).

For Piketty (2004), the fact Hoxby could not reproduce the results of Angrist and Lavy may be due to the fact that the average class size is much bigger in Israel than in Connecticut (30 vs 20 students). He wanted to apply the RDD model to France. He exploited a panel data from 1997. 9,000 students entering in third grade were sampled. They had to fill a background questionnaire and take a standardized test in third grade. The author tried to see if there was a positive effect (negative coefficient) of class size in second grade on the score obtained in third grade. As in the USA, there is not an explicit maximum class size rule in France.

Nevertheless, there is an implicit one of 30 students. He concluded that an increase in class size of one student on average leads to a reduction in score of 0.422 (out of 20). This coefficient is statistically significant (Piketty, 2004). This model is probably the statistically strongest we have in France. Nevertheless, the study has a big issue: Piketty does not evaluate the effect of current class size on student performance but the effect of past class size on student performance.

The methods we saw until now are difficult to implement, either because they are costly or because strong assumptions need to hold to apply them. Most of the time, researchers rely on regression analysis to estimate the impact of class size on student performance.

2.5. Multiple regression method

Ehrenberg et., al (2001) in their literature review talk about the Coleman report (1966) which constitutes the "beginning of the *educational production function* literature" (Ehrenberg et., al, 2001). The sample was composed of 570,000 students in the US. Their test score has been regressed on many variables including family, community, and school characteristics. The main conclusion of the paper is that community and family background variables had much more influence on the test score than school characteristics. This study typically supported the "money makes no difference" argument of Hanushek since school infrastructures, class size… seemed to have a negligible impact on student performance. However, the study has received a large number of criticics. The paper studied the effect of the average pupil/teacher ratio of students' schools on their test scores, not the effect of the students' actual class size. Most importantly, this study was a "snapshot". Only variables at time t were included, not characteristics of students in the past (G. Ehrenberg et., al, 2001).

Piketty (2004) tried to correct for the second critic in his 2 regression models. In the first model, he regressed test score on past class size and socio-demographic characteristics of students. He came up with the following result: when a class size in the second grade increases by 1 student, the test score in the third grade on average decreases by 0.169 points (out of 20) keeping everything else constant. In the second model, he adds the test score that the student had obtained in his first grade. His coefficient of interest increases significatively: 0.205. The reason for this increase is that first-grade test score is positively correlated with class size in the second grade (see the Lazear model) and positively correlated with the test score in the third grade (Piketty, 2004).

Overall, multiple regression models suffer from endogeneity because of omitted variable bias. In other words, many observed or unobserved variables correlated with class size and test score are not included in the model which biases the coefficient of interest. There have been hundreds of papers that used multiple regression to study the relationship between class size and achievement. However, they present contradictory results. For the case of France, Caubel (2019) says in her literature review that out of the 7 studies that used this method, 4 of them concluded that there was no association between class size and achievement (Caubel, 2019). This is probably because their independent variable of interest (class size) is continuous despite the fact that there is little variation in class size in their dataset. In my dataset, there are more variations in class size. I will perform a regression with a continuous dependent variable and a regression closely related to the one of Bohrnstedt & Stecher (1999) with a binary independent variable. Let me now present my methodology and the data I will use.

3. Data and Methodology

The data I will use comes from the *International Progress In Reading Literacy* (PIRLS) database. Since 2001, every 5 years, some 4-th grade students (9 to 10 years old) across about 40 countries take the PIRLS reading test in their home language. The tests are run by the International Association for the Evaluation of Educational Achievement (IEA). The students are selected by a two-stage random sampling. Firstly, in all countries, schools are sampled based on probabilities proportional to their size. Secondly, one or more classes of the chosen school must take the test (Martin, 2012). Students as well as their principal, parents, and teacher must fill a background questionnaire. The test consists of 135 multiple-choice and open questions of reading comprehension over 10 passages (5 literacy and 5 informational) (Quéré, 2011). For a timing matter, each student does a sub-sample of the test and answers questions on 2 reading passages in one hour (Martin et., al, 2007). Their overall score is predicted by five *"plausible values"* which are calculated based on their background questionnaire and the answers to their test. The test occurs in May so 9 months after the start of the academic year. As mentioned in the Introduction, I chose to focus on France. To have the largest sample size possible, I merged the PIRLS data from 2006, 2011, and 2016. My sample consists of 13,269 students coming from 546 schools.

Many French studies opt for a linear regression of class size on student performance. As mentioned in the introduction, the caveat of such an approach is that there is little variation in class size in the researcher's datasets. Thus, it is hard to really capture the effect of class size on student performance and we often obtain disappointing results. However, as we see in Figure 1, there are variations in class size in my data. I will first plot the following *Ordinary Least Square* (OLS) linear regression:

(1)

$$\text{Score}_i = \beta 0 + \beta 1 * \text{Class}_i + \beta 2 * \text{Affluent}_i + \beta 3 * \text{Disadvantage}_i + \beta 4 * \text{Desk}_i + \beta 5 * \text{Computer}_i + \beta 6 * \text{Books}_i + \beta 7 * \text{FatherHighEduc}_i + \beta 8 * \text{MotherHighEduc}_i + \beta 9 * \text{JobFather}_i + \beta 10 * \text{JobMother}_i\ \beta 11 * \text{FrenchAtHome}_i + \beta 12 * \text{TimeRead}_i + \beta 13 * \text{Male}_i + \beta 14 * \text{Year}_i + \varepsilon_i$$[6]

Score_i[7] is our dependent variable. It represents the predicted reading score of student i. The maximum score a student can have is 700 and the minimum score is 300[8]. In Figure 2, we see that scores are almost normally distributed around the mean which is 517 (Table 1). Here, the independent variable of interest is Class_i which depicts the class size of student i. The coefficient of interest β1 can be interpreted as: when class size increases by 1 student, reading test score on average increases/decreases by β1 points, keeping everything else constant. The

[6] One of the assumptions of linear regression is the independence of observations. More formally, there should not be clusters in a data set. However, in my case, groups of students coming from the same class are included in the Data, which violates the assumption. To correct this, I used for all my regressions the cluster function in Stata.

[7] As we said, 5 plausible values have been calculated. The Score here corresponds to the first plausible value (PV1)

[8] The international average is 500

other variables are control variables which affect the class size of student i and his reading score. I will describe them below.

I also want to study the difference in reading scores between students in "large" classes and students in "small" classes. I had to choose a benchmark of class size under which would be "small classes" and above which would be "large classes". This benchmark of class size is 25, the median class size in our sample. In Figure 1, I made a histogram of class size. We see that the distribution is left-skewed, the mean (24.58) is inferior to the median class size. This is because of outliers (really small classes) to the left of the distribution. In my main model, I chose to keep those outliers because the cumulative distribution of class size from 10 to 17 is very small (1.33% of the sample). In the Robustness part, I will drop those outliers and compare the classes with 18 – 24 students to classes with 25 – 32 students. In my model, students in "small classes" have class sizes inferior to 25 students and represent 47.28 % of the sample. Students in "large classes" have class sizes larger or equal to 25 and represent 52.72 % of the sample. My two comparison groups are thus almost symmetric in terms of size.

To study the difference in test scores between students in large and small classes, I chose to use an *Ordinary Least Square* (OLS) regression model. With such a method, I can evaluate the difference in average reading scores between the two groups and control for variables that could affect the score and the class size allocation. Here is the regression I want to estimate:

(2)

$Score_i = \beta0 + \beta1 * Small\_Class_i + \beta2 * Affluent_i + \beta3 * Disadvantage_i + \beta4 * Desk_i + \beta5 * Computer_i + \beta6 * Books_i + \beta7 * FatherHighEduc_i + \beta8 * MotherHighEduc_i + \beta9 * JobFather_i + \beta10 * JobMother_i \beta11 * FrenchAtHome_i + \beta12 * TimeRead_i + \beta13 * Male_i + \beta14 * Year_i + \varepsilon_i$

$Small\_Class_i$ is a dummy variable = 1 if student i class size is less than 25 and 0 if student i class size is more or equal to 25. β1 is our coefficient of interest. It can be interpreted as: on average, students in small classes obtain β1 points less/more on their reading test than students in large classes, keeping everything else constant. To my knowledge, only one study has assigned randomly students to small and large classes to compare their performance[9]. This ensures that both treatment and control groups are similar in characteristics and thus, that the difference in score between the large and the small class can be interpreted causally. However, in my setting, students are not assigned randomly to large or small classes. I thus have to add control variables to my model to eliminate some of the selection bias. A good control should affect the class size and also the score of student i. I decided to control for 13 variables[10] that I will now describe[11].

[9] STAR experiment, 1985

[10] To choose my control variables, I read sociological papers. I had to see what school or student characteristics affect the most their performance and thus their class size allocation (see the Lazear model). Sociologists most of the time provide some empirical evidence about the impact of a variable on student performance and give

a) School Characteristics

The categorical variables Affluent$_i$ and Disadvantage$_i$ come from the School Questionnaire and can take 4 values. For the first variable, the school principal answered the question: “Approximately, what % of students in your school come from economically affluent homes?”. For the second: “Approximately, what % of students in your school come from economically disadvantaged homes?”. 4 answers were possible: 0-10 %; 11-25 %; 26-50%; More than 50%. We see that the average score of students evolves positively with the proportion of “affluent” students in their school and negatively with the proportion of disadvantaged students in their school. In the sociological literature, I chose[12] 2 theoretical arguments for this phenomenon. First, the quality of teaching is higher in the richer than in the poorest school. Teachers with less experience tend to be allocated to the most “difficult” schools (Dubar, 2002). Also, teachers in difficult schools decrease their requirements and spend more time taking care of extra-learning activities (e.g., discipline of students) than in richer schools (Dubar, 2002). Second, there is a *Pygmalion* effect in the richest schools and an inverse *Pygmalion* effect in the poorest schools (Rosenthal et al., 1968). In the richest (poorest) schools, teachers believe that their students have high (low) capabilities which make them progress (regress) (Rosenthal et al., 1968).

In Table 1, we see that students enrolled in schools with a low proportion of affluent students (less than 10 %) are almost twice more in small classes than in large classes. The phenomenon is reversed when we look at students enrolled in schools with a high proportion of affluent students (more than 50%). Thus, we see that the more affluent students are in the school, the higher the probability of being enrolled in a large class. The opposite reasoning holds for the proportion of disadvantaged students in the school.

b) Student’s parents’ characteristics

I think it is important to control for the education and the occupation of student’s parents. FatherHighEduc$_i$ and MotherHighEduc$_i$ are dummy variables = 1 if respectively student father and student mother pursued higher education (at least a bachelor’s degree) and 0 otherwise. JobFather$_i$, JobMother$_i$ are two categorical variables depicting the professional situation (at the time of their child test) of the father and mother. Both can take 11 values:

1) Has never worked outside the home for pay

some theoretical interpretation. I will here provide some sociological interpretations for differences in test scores across groups. Please note that the interpretations I will give are non-exhaustive. For example, the fact that males perform weaker than females cannot solely be explained by the fact that girls want to defy the stereotypes assigned by their classmates (Gagnon, 2005). The literature on gender theory is huge and there are many other explanations. I just tried for every variable to give some (not every) sociological interpretation about differences in test scores across groups.

[11] In Table 1, I expose partial descriptive statistics with the most important variables. In Table 9 (Appendix) you can find the full descriptive statistics.

[12] Typically, there are many more arguments.

2) Small Business Owner (fewer than 25 employees)

3) Clerk

4) Service or Sales Worker

5) Skilled Agricultural or Fishery Worker

6) Craft or Trade Workers

7) Plant or Machine Operator

8) General Laborers

9) Corporate Manager or Senior Official

10) Professional (scientists, mathematicians, computer scientists, architects, engineers, health professionals, teachers, legal professionals, social scientists)

11) Technician or Associate Professional

Parents' socio-professional characteristics surely influence their children's academic performance. Bourdieu and Passeron (1999) extensively worked on school inequalities and social reproduction. They explained that school inequalities mostly came from differences in *habitus* (set of norms and values acquired during the primary socialization) between students. Pupils coming from privileged backgrounds incorporate some norms and values that mostly respond to school expectations (in terms of language, relationship to knowledge…) and thus have higher chances to succeed than students coming from disadvantaged background (Bourdieu & Passeron, 1999). In Table 1, we see that children with high-educated parents do better on average in their reading test than students with low-educated parents (more than 30 points difference). We also observe disparities in students' scores depending on the profession of their parents. For example, children of executives (professionals, corporate managers) on average perform significantly better than children of non-executives (Table 9).

In Table 1, we see that students with low-educated parents are half in large classes and half in small classes. However, students with high-educated parents tend to be allocated to larger classes: approximately 60 % of pupils with high-educated parents are allocated in large classes against 40% in small classes. With regards to the profession of parents, we see that students that are the most allocated to large classes are children of executives (almost 60% of them are allocated to large classes). We also see that the group of children with the highest proportion of allocation to small classes are children of parents that have never been working (Table 9).

c) Student possessions

I think it is important to control for the material students have at home to work because it can affect their academic level and thus their class allocation. $Book_i$ represents the number of books student i has at his house. It can take 5 values: 0-10; 11-25; 26-100; 101-200; more than 200. We see in Table 9 that the average score of students is increasing in the number of books they have at home. Having many books at home might be a sign of parental high cultural capital which is a crucial factor in academic success (Brecko, 2004). We see class size disparities across groups in Table 9 but the results are counter-intuitive if we consider the Lazear model. Students with the least number of books at home are more often assigned to large classes and students with the largest amount of books are more often assigned to small classes.

$Desk_i$ is a dummy variable equal to 1 if student i has a desk at home and 0 otherwise. It is a good indicator of student working conditions. We see that students with no desk at home obtain on average less in their reading test (by 18 points) than students with a desk at home. In terms of class size allocation, we see that students that have a desk at home are more often in large classes than students with no desks at home (Table 9).

$Computer_i$ is a dummy variable equal to 1 if the student has a computer at home and 0 otherwise. Students with a computer at home perform on average significantly better (by 31 points) than students with no computer at home. This may be in part because the possession of a computer is positively correlated with the socio-economic condition of families. We can add another explication: with the arrival of *pronote*[13] in the 2000s, parents can be informed of the performance of their child, the comments of their teacher, and their homework to do. Maybe, access to computers can decrease the asymmetry of information between students and parents and hence increase the workload of students. In Table 9, we see that students with a computer at home are more often allocated to larger classes than children without a computer at home.

d) Student language at home

$French_i$ is a dummy variable equal to 1 if the student always speaks French at home and 0 if the student sometimes or always speaks another language at home. We see that students who speak French at home have on average higher scores than pupils who do not always speak French at home (by 20 points) (Table 9). In terms of class allocation, we see in Table 9 that students who speak French at home are more often allocated to large classes than students who do not speak French at home.

e) Student gender

$Male_i$ is a dummy variable equal to 1 if student i is a male and 0 if student i is a female. We see that girls perform on average better than males on the reading test (by 8 points) (Table 1). Sociologists try to explain why girls are in general better students than boys. One explication of this phenomenon is that girls want to defy the stereotypes assigned by their classmates and hence are motivated to work harder (Gagnon, 2005). In terms of class size allocation, there are no significant differences between the two groups.

[13] Pronote is an app where figures the results of children, their homework to do, the appreciations of their teachers…

f) Student time allocated to reading

$TimeRead_i$ is a categorical variable. It is drawn from the student questionnaire. The question is: “On average, how much time per week do you read for fun”. This categorical variable can take four values: Less than one hour a week; 1 - 5 hours a week; 6 - 10 hours a week; more than 10 hours a week. We see that average test scores are increasing in the number of hours per week spent reading. However, in terms of class size allocation, there are no differences across groups (Table 9).

g) Year

$Year_i$ is a categorical variable that can take 3 values: 2006, 2011, 2016. I think it was important to control for years because there are disparities in scores and class size allocation between them. We see in Table 1 that the average score has kept declining over years. In terms of class size, we see that more students were allocated in large classes in 2011 than in other years (+4 points).

A final remark: it is important to note that β1 will not give a causal effect but just an association. In other words, the average difference in test scores β1 (keeping everything else constant) between students in large and small classes cannot be imputed to the size of their class. This is because of endogeneity issue: to have a causal interpretation, the correlation between the error term ε_i (omitted variables that affect $Score_i$) and the variable of interest $Small\_Class_i$ should be null. Hence, we should incorporate all the variables that could affect $Score_i$ and $Small\_Class_i$ (good controls). We tried here to incorporate the most intuitive variables that could affect both score and class size. However, there are many observables and unobservable good controls that we did not incorporate. Our model thus suffers from omitted variable bias and cannot be interpreted causally.

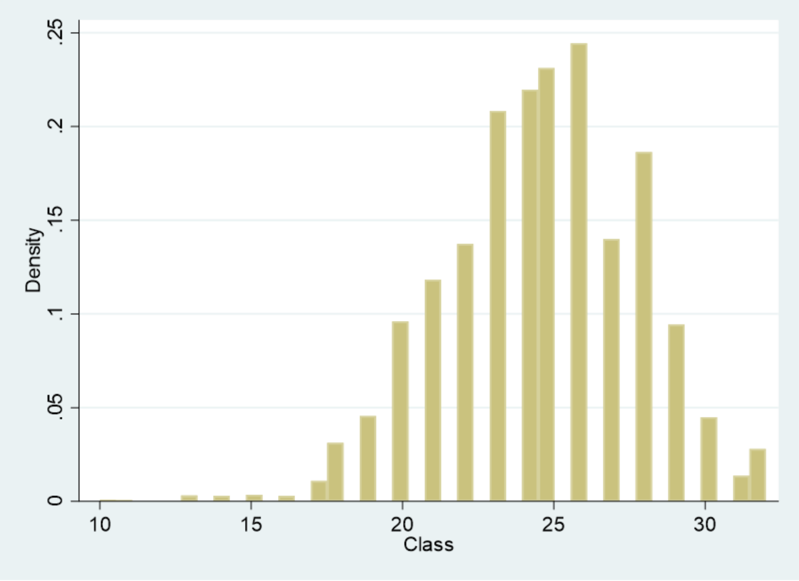
Table 1. Partial descriptive statistics

		Large Class	Small Class	Score	N
Affluent	0-10 %	38.1	61.9	497.668	4,160
	11 - 25 %	54.1	45.9	512.456	2,797
	26 - 50 %	53.9	46.1	534.778	2,738
	more than 50 %	67.9	32.1	538.601	3,574
Disadvantaged	0-10 %	65.5	34.5	534.917	4,878
	11 - 25 %	56.5	43.5	524.088	3,281
	26 - 50 %	45.3	54.7	508.944	2,260
	more than 50 %	32.3	67.7	486.464	2,850
FatherHighEduc	Yes	49.9	50.1	539.969	4,752
	No	57.7	42.3	504.821	8,517
MotherHighEduc	Yes	48.1	51.9	540.244	5,457
	No	59.4	40.6	501.457	7,821
Sex	Male	52.6	47.4	513.207	6,734
	Female	52.9	47.1	521.567	6,535
Year	2006	51.8	48.2	522.896	4,330
	2011	55.2	44.8	521.656	4,331
	2016	51.3	48.7	508.002	4,628
Total	Total	52.7	47.3	517.409	13,269

Source: Calculation made with the PIRLS 2006, 2011, 2016 databases

Notes: Notes: The first 2 columns (Small Class, Large Class) show the percentage of students among groups that are allocated to small and large classes. The first line can be read as: among pupils who are in schools with 10% of students coming from affluent homes, 38.1% are in large classes, and 61.9 % in small classes. The Score column depicts the average score per subgroup. For example, Males on average obtain a score of 513 (out of 700) on their reading test. The N column shows the sample size.

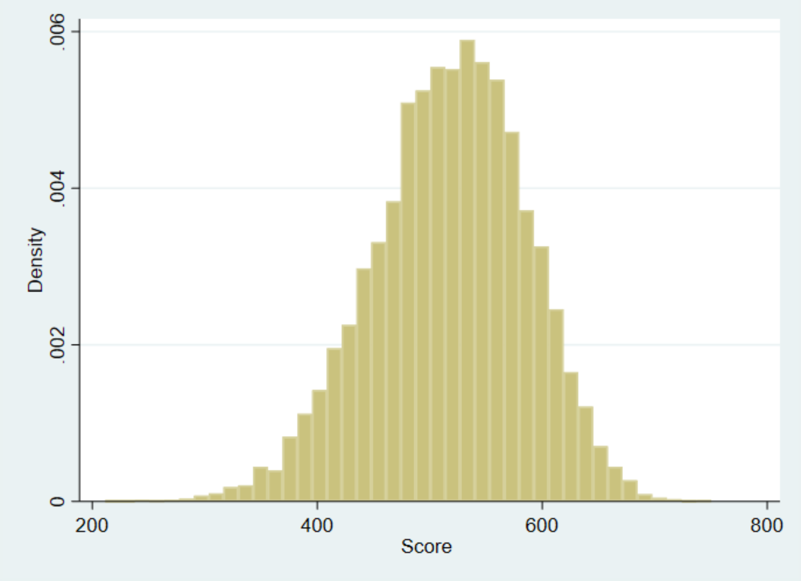
Figure 1. Histogram of class size



Source: Histogram made with the PIRLS 2006, 2011, 2016 database

Notes: The x-axis depicts the class size and the y-axis the density. We see for example that 9% of the sample have classes with 20 students.

Figure 2. Histogram of scores



Source: Histogram made with the PIRLS 2006, 2011, 2016 database

Notes: The x-axis depicts the score and the y-axis the density.

4. Results

4.1. Association between class size and student achievement

The results of the first model are in Table 2. For all the upcoming tables, Model 1 does not incorporate any controls while Model 2 incorporates all the controls. In Model 1, we see that increasing the class size by 1 student on average leads to a higher test score of 2.423. This coefficient is statistically significant at a 1% confidence level. This relation is counter-intuitive because it suggests that the more students are in the class, the higher the test score (on average). However, we saw with the Lazear model (1999) that weaker students tended to be allocated to larger classes. We thus controlled for variables that could affect the level of a student and hence its class allocation. The second model incorporates all the controls that we evocated in section 3. We see that the coefficient is not statistically significant. That means that we do not have enough evidence to say that class size affects student achievement. The 95% confidence interval of the coefficient: [-0.314; 0.353] informs us that we obtain a "precise estimated zero". Indeed, we can be 95% confident that the actual coefficient for the French population (with our model settings) is significantly close to zero.

Table 2. Association between class size and student achievement - OLS estimate

	Model 1	Model 2
Class	**2.423 *****	**0.019**
(s.e)	(0.474)	(0.222)
Controls	No	Yes
N	13,269	13,269

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable is Score and the independent variable of interest is Class (class size). In the first model, there are no controls, and in the second model, there are the 13 controls that we described in Section 3

4.2. Difference in test scores between small and large classes (scale: 25 students)

In Table 3 figure the results of our second regression. In Model 1, we see that on average, students in small classes have on average 12.488 less points on their reading exam than students in small classes. This coefficient is significant at a 1% confidence level. In Model 2, we add all the controls cited in the previous section. We see that on average, students in small classes have 1.777 more points than students in large classes, keeping everything else constant. This coefficient is significant but only at a 10% confidence level. That means that I can be 90% confident, that the actual average difference in test scores between students in small and large classes for the French population is different from 0[14].

We thus see that by adding controls, we go from a negative to a positive coefficient. However, the magnitude of this coefficient is very low since the range of the scores goes from 300 to 700. We therefore need to see if by changing the settings of our model, we still obtain a positive and significant coefficient. This is what we will now explore in our Robustness section.

Table 3. Difference in test scores between small and large classes (scale: 25 students) - OLS estimate

	Model 1	Model 2
Small_Class	**-12.488 *****	**1.777 ***
(s.e)	(3.243)	(1.067)
Controls	No	Yes
N	13,269	13,269

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable is Score and the independent variable of interest is Small_Class. In the first model, there are no controls, and in the second model, there are the 13 controls that we described in Section 3

[14] The 90% confidence interval is [0.008 ; 3.548]

5. Robustness

5.1. Dropping the outliers

In the Data part, we saw that there were really small classes in our sample (with 17 or less students). Those classes were outliers but we decided to incorporate them in our model because there were few of them. We know that two of the assumptions of an OLS regression are normality and homoskedasticity. Outliers can negatively affect those assumptions and hence decrease the statistical inference of our model. We now drop those outliers, 177 observations are deleted. We see that the coefficient of interest in Model 2 is now less than the coefficient we had in our result and that it is not significant. Hence, our model is not robust to the elimination of outliers.

Table 4. Difference in test scores between small and large classes (scale: 25 students) without outliers – OLS estimate

	Model 1	Model 2
Small_Class	**-12.767*** **	**1.567**
(s.e)	3.361	1.099
Controls	No	Yes
N	13,092	13,092

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable is Score and the independent variable of interest is Small_Class. In the first model, there are no controls, and in the second model, there are the 13 controls that we described in section 3. In this model, we sacrifice 99 outliers.

5.2. Log Score as dependent variable

We now perform a logarithmic transformation of our dependent variable. Such a transformation is generally used for two reasons: to make the dependent variable distribution more symmetric and to have a better linear relation between two variables. We saw in section 3 that scores were normally distributed. Hence, doing a logarithmic transformation may improve the linearity of our model. As in our main model, we see that we go from a negative to a positive coefficient. However, this time, the coefficient is not statistically significant. Here is how the coefficient in Model 2 can be interpreted: on average, students in small classes have test scores 0.3% higher than students in large classes, keeping everything else constant. The magnitude is more or less the same as in the previous model (1.777/700 = 0.0025), but this coefficient is not significantly different from 0. Hence, our model is not robust to a logarithmic transformation of the dependent variable.

Table 5. Difference in the logarithm of scores between students in small and large classes (scale: 25 students) - OLS estimate

	Model 1	Model 2
Small_Class	**-0.025 *****	**0.003**
(s.e)	0.006	0.002
Controls	No	Yes
N	13,269	13,269

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable is the logarithm of Score and the independent variable of interest is Small_Class. In the first model, there are no controls, and in the second model, there are the 13 controls that we described in section 3.

5.3. Comparing students in classes of 20 students and fewer with students in classes of 29

In the last OLS regressions, our variable of interest was Small_Class. We compared students in classes with less than 25 students with students in classes with more than 25 students. We chose this benchmark because the number of students above and below it was almost the same. Also, it enabled us to have the most observations possible. We now want to compare students in classes of 29 students with students in classes of 20. The symmetry between the two groups is present: In Figure 1, we see that approximately 9% of the sample is in both types of classes. However, we sacrifice our sample size, going from 13,269 to 1,354 students. Small_Class20 is a dummy variable equal to 1 if the student is in a class of 20 students and 0 if the student is in a class of 29 students. In Model 1, we see that the difference in scores between the two groups is more important in terms of magnitude than in our main model. When adding controls, this coefficient increases by 13 points but it is negative and not significant. This “disappointing” result may be due to the fact that the sample size is very small.

Table 6. Difference in scores between small (20 students) and large classes (29 students) - OLS estimate

	Model 1	Model 2
Small_Class 20	**-17.856 *****	**-5.103**
(s.e)	(4.531)	(3.899)
Controls	No	Yes
N	1,354	1,354

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable Score and the independent variable of interest is Small_Class 20. In the first model, there are no controls, and in the second model, there are the 13 controls that we described in Section 3. Here, as we only select students that are in classes of 20 or 29, we sacrifice a large part of the sample (approximately 12,000 students)

5.4. Difference in test scores between small and large classes among disadvantaged schools

One question that can be asked is whether students in disadvantaged schools benefit from smaller classes in terms of performance. As we saw in the introduction, French policies of class size reduction concern schools in disadvantaged districts (REP+). Typically, in disadvantaged schools, we saw that teachers spend a lot of time taking care of extra-knowledge areas, for example, discipline issues. Maybe in smaller classes, teachers have more time to actually teach since the distraction opportunities for pupils are smaller. For this section, I selected from the sample students in schools that have 25% or more children coming from economically disadvantaged homes. We see that our sample size drops significantly. In Model 1 (without controls), we observe that in those disadvantaged schools, students in large classes have on average higher scores than students in small classes. The coefficient is significant at a 1% confidence level but the magnitude is way smaller than in our main model. When adding controls, the relationship is inversed. Students in large classes have on average higher scores than students in small classes keeping everything else constant. The coefficient is higher than the one estimated in our main model but it is not statistically significant. Again, this is probably due to the small sample size.

Table 7. Difference in scores between students in small and large classes (scale: 25 students) in disadvantaged - OLS estimate

	Model 1	Model 2
Small_Class	**-5.555 *****	**1.987**
(s.e)	2.001	1.831
Controls	No	Yes
N	5,110	5,110

Source: Regression made with the PIRLS 2006, 2011, and 2016 database

Notes: The dependent variable Score and the independent variable of interest is Small_Class. In the first model, there are no controls, and in the second model, there are 11 out of the 13 controls that we described in Section 3. We dropped here the controls i.Affluent and i.Disadvantaged. Here, as we only select students that are in schools with 25% or more disadvantaged pupils, we sacrifice more than half of the sample.

5.5. Difference in test scores between disadvantaged students in small and large classes

It might be the case that disadvantaged students benefit from being allocated to smaller classes. Here, we call “disadvantaged students”, students whose parents did not pursue any higher education. Of course, it is a proxy and it might be the case that some of those students are actually not socially and economically disadvantaged. In Model 1, we see a negative and significant difference in test scores between disadvantaged students in small and large classes. The magnitude is less important than in the main model. In Model 2, when adding controls, we see that students in small classes have on average 1.408 more points than students in large classes, keeping everything else constant. However this time, the coefficient is not significant. This might be due to the sample size that we divided by two.

Table 8. Difference in scores between disadvantaged students in small and large classes (scale: 25 students) in disadvantaged - OLS estimate

	Model 1	Model 2
Small_Class	**-7.475 *****	**1.408**
(s.e)	1.879	1.659
Controls	No	Yes
N	6,298	6,298

Source: Regression made with the PIRLS 2006, 2011, 2016 database

Notes: The dependent variable Score and the independent variable of interest is Small_Class. In the first model, there are no controls and in the second model, there are the 11 out of the 13 controls that we described in section 3. We dropped here the controls FatherHighEduc_i *and* MotherHighEduc_i. *Here, as we only select students whose both parents did not pursue any higher education, we sacrifice about half of the sample.*

6. Discussion

6.1. Caveats of my study

In this paper, I derived a regression model to estimate the difference in achievement between students in small and large classes. However, the model suffers from many caveats. Here are two of them:

(i) Omitted variable bias

My model suffers from endogeneity which is a crucial assumption to infer a causal relationship. For a regression to be internally valid, the correlation between the error term and the variable of interest should be 0. Hence, all the variables affecting both the outcome variable (Y - Score) and the variable of interest (X – Small_Class) should be included in the model. We lack many of them but in my sense, the most crucial one is past test scores. This one might be positively correlated with the PIRLS test score and with the class size of a student. Hence, if we had included such a coefficient, the coefficient would have probably been higher. This omitted variable causes downward bias. Another crucial one is the migration background of the student which we do not have access to due to ethical reasons. We might expect students with migration backgrounds to obtain less in their reading score and to be allocated to smaller classes. Again, if we had included this variable, the coefficient would have probably been higher. This omission is hence another source of downward bias.

(iii) PIRLS test score

One of the issues with a PIRLS score as a dependent variable is that it is hard to exploit it for policy-makers. A class size-reduction policy should be implemented if the rise in productivity in the future (or higher wages) caused by the higher test score offsets the cost of operation. However, such an international test does not have any effect on the academic trajectory of a pupil and hence on his future wage. Even if one scores a 0 on such a test, he can go to the next grade if he succeeds in a national test. That is why national standardized tests are often chosen to study the effect of class size (Krueger et al., 2002).

6.2. Suggestion for future research – Use the Krueger (2002) model

What is the implication of my study for policy-makers? My study showed that being in a small class was associated with a larger test score, keeping everything else constant. However, since my model suffers from endogeneity, no causal conclusions can be drawn from it. Nevertheless, I would like to give suggestions for future research which could have an impact on policy decisions. We saw in the introduction that the French government divided class sizes by 2 in disadvantaged districts. However, we do not know if this maximum class size rule (12 students) is optimal based on actual research. For future class size policy decisions, it might be necessary to apply Krueger's (2002) cost-benefit analysis to see what would be the optimal class size. Namely, governments should reduce the class size until the marginal benefit (MB) of class size reduction (the increase in the present value of future earnings caused by an x rise in test score because of a class reduction by 1 student) equals the marginal cost (MC) of class reduction (Krueger et al., 2002).

To build the MB curve, researchers should study the relation between class size and test score[15] and the relation between test score and PV of future earnings[16]. The MB would have as a slope let's say z. Here is how z could be interpreted: A decrease in class size by 1 leads to an average increase in test score by x which causes an increase in PV of future earnings by z. We here suppose that the MB of class reduction is constant because we use linear regression. However, researchers may want to plot polynomial equations.

Based on studies about class reduction policies in France or abroad, researchers could come up with an MC curve of class reduction. That one would depict the incremental cost of reducing classes by 1. The optimal class arises when the MB of a class size reduction equals the MC.

I am conscious that this type of study preconized by Krueger (2002) is really hard to implement. However, this constitutes a strong theoretical model that researchers can use to evaluate what would be the optimal class size in France.

[15] This could be done using a randomized experiment (as in STAR). The government could select schools and ask them to randomly allocate students and teachers to two, three or more groups (classes with different sizes). Then researchers could build a regression of the impact of class size (x) on student test score (y). We could imagine a downward-sloping curve

[16] This one would be very difficult to find. We would have to obtain data on students test score at grade x, their future yearly earnings and other characteristics (IQ, family background, school background). We would have to discount the yearly earnings by a discount factor. By plotting a regression with test score on the x axis and PV of future earnings on the y axis, we can imagine that the curve would be upward sloping.

7. Conclusion

In this study, I first wanted to see if class size was correlated with student performance. I built a regression model with the 4th grade reading score as dependent variable and the class size as independent variable. I added 13 controls. I did not find a statistically significant relationship. I then wanted to see if students in small classes (less than 25 students) performed better (in reading) than students in small classes (more than 25 students). I chose to focus on France because class size reduction policies are often implemented but researchers do not agree on the benefits of class size reduction. I built an OLS regression model with 13 control variables and I used a very large dataset (more than 13,000 persons). According to my model, students in small classes perform on average significantly (at a 10% confidence level) better than students in large classes, keeping everything else constant. The magnitude of the difference is very low (approximatively, they have 0.45 % better scores if we consider the score scale: 300-700). Moreover, because of omitted variable bias, our coefficient cannot be interpreted causally. In the Robustness part, we changed the settings of our model. We tried to see the difference in test scores between small and large classes among disadvantaged schools and disadvantaged students. We also tried to study the difference in test scores between students in classes of 20 and students in classes of 29 students. Every time, we obtained no statistically significant differences in test scores.

Those “disappointing” results should not discourage the government to pursue class size reduction policies. Firstly, because we chose to focus on fourth grade and class size reduction policies typically concern lower grades (because disruptive behaviors are higher). Secondly, even if the government considers to do a reduction policy on fourth grade, the outcome might be positive. We indeed mentioned that our model suffers from endogeneity and that we would have expected the difference in test scores to increase if we had access to past test scores and to the migration background of students. Hence, our study should not be taken as an example for economists arguing that “money makes no difference”.

8. References

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, *114*(2), 533-575. https://doi.org/10.1162/003355399556061

Bohrnstedt, G.W. & Stecher, B.M. (Eds.) (1999). Class Size Reduction in California: Early Evaluation Findings, 1996–98. Palo Alto, CA: American Institutes for Research.

Bourdieu, P., & Passeron, J. (1999). *La reproduction : éléments d'une théorie du système d'enseignement*.

Brecko, B. N. B. (2004). How familly backround influences children achievement. *Educational Research Institute, Slovenija*.

Bressoux, P., & Lima, L. (2011). La place de l'évaluation dans les politiques éducatives : le cas de la taille des classes à l'école primaire en France. *Raisons Educatives*.

Caubel, E. C. (2016). *Taille des classes et réussites scolaires. Education. 2019*. Education.

Coleman, J. S. (1966). *EQUALITY OF EDUCATIONAL OPPORTUNITY.* ERIC. https://eric.ed.gov/?id=ED012275

Diggle, P. J., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2001). Analysis of longitudinal data. Dans *Psychology Press eBooks* (p. 215-242). https://doi.org/10.4324/9781410605542-11

Dubar, C. (2002). Van Zanten (Agnès). - L'école de la périphérie. *Revue française de pédagogie*, *140*(1), 143-145. https://www.persee.fr/doc/rfp_0556-7807_2002_num_140_1_2906_t1_0143_0000_5

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest*, *2*(1), 1-30. https://doi.org/10.1111/1529-1006.003

Finn, J.P. & Achilles, C.M. (1999). Tennessee's class size study: Findings, implications and misconceptions. Educational Evaluation and Policy Analysis, 21, 97–110.

Gagnon, C. (2005). La dynamique de la réussite scolaire des filles au primaire : les motivations et les enjeux des rapports sociaux de sexe. *Recherches féministes*, *11*(1), 19-45. https://doi.org/10.7202/057965ar

Goldstein, H. & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with special reference to study design. British Educational Research Journal, 24, 255–268.

Hanushek, E. A. (1999). *The evidence on class size | Eric A. Hanushek*. Standford. http://hanushek.stanford.edu/publications/evidence-class-size

Hoxby, C. M. (1998). *The effects of class size and composition on student achievement : New evidence from natural Population Variation*. https://doi.org/10.3386/w6869

Hoxby, C. M. (2000). *The effects of class size on student achievement : New evidence from population Variation on JSTOR*. https://www.jstor.org/stable/2586924

Krueger, A. B., Hanushek, E. A., & Rice, J. K. (2002). *The Class Size Debate*. Economic Policy Institute. https://www.epi.org/publication/books_classsizedebate/

Krueger, A. B., & Schanzenbach, D. W. (1997). The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results : Evidence from Project STAR. *Princeton University, Industrial Relation Sections Working Papers*. https://doi.org/10.2139/ssrn.223492

Lazear, E. P. (1999). *Educational production*. NBER Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=194830

Martin, M.O., Mullis, I.V.S. & Kennedy, A.M. (2007). Pirls 2006, Technical report. Chesnut Hill, Boston College

Martin, M.O., Mullis, I.V.S., Foy, P., & Arora, A. (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M.O. Martin & I.V.S Mullis (Eds.), Methods and Procedures in TIMSS and PIRLS 2011 (pp.1– 11). (s. d.). *TIMSS & PIRLS International Study Center, Boston College*

Mosteller, F. (1995). The Tennessee study of class size in the early school grades: The future of children

Piketty, T. (2004). *L'impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises : une estimation à partir du panel primaire 1997*. EHESS, Paris-Jourdan. http://piketty.pse.ens.fr/files/Piketty2004b.pdf

Quéré, M. (2010). PIRLS 2011 - Étude internationale sur la lecture des élèves au CM1 Évolution des performances à dix ans. *Ministère de l'Education Nationale*. https://www.lireetfairelire.org/sites/default/files/depp-ni-2012-21-pirls-2011-etude-internationale-lecture-eleves-cm1_236680.pdf

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16-20. https://doi.org/10.1007/bf02322211

Rothstein, R. I., & Miles, K. H. (1995). Where's the Money Gone ? Changes in the Level and Composition of Education Spending. *Economic Policy Institute*. https://eric.ed.gov/?id=ED396422

L'expérimentation d'une réduction des effectifs en cours préparatoires / François Alluin, Marc Colmant, Olivier Cosnefroy, Chi-Lan Do, Jean-Claude Emin, Fabienne Gibert, Jacqueline Levasseur, Jean-François Levy, Aude Mulliez et Catherine Régnier. (s. d.). https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/11477/l-experimentation-d-une-reduction-des-effectifs-en-cours-preparatoires-francois-alluin-marc-colmant-?_lg=fr-FR

Mathématiques : L'évaluation internationale PISA 2022. (s. d.). Académie de Créteil. https://www.ac-creteil.fr/mathematiques-l-evaluation-internationale-pisa-2022-121922#:~:text=La%20France%20dans%20le%20dernier%20classement%20PISA,-Le%20dernier%20classement&text=En%20math%C3%A9matiques%2C%20les%20r%C3%A9sultats%20de,24e%20place%20dans%20cette%20mati%C3%A8re.

Laplaud, L. S. É. P. L. (2022, 2 septembre). Dédoublement des classes : qu'en est-il, cinq ans plus tard ? *Europe 1*. https://www.europe1.fr/societe/dedoublement-des-classes-quen-est-il-cinq-ans-plus-tard-4131652

PISA 2012 : meilleure réussite et moins d'inégalités en résolution de problèmes. (s. d.). *Ministère de l'Education Nationale et de la Jeunesse*. https://www.education.gouv.fr/pisa-2012-meilleure-reussite-et-moins-d-inegalites-en-resolution-de-problemes-3818

9. Appendix

9.1. Full table of descriptive statistics

Table 9. Full descriptive statistics

		Large Class	Small Class	Score	N
Affluent	0-10 %	38.1	61.9	497.668	4,160
	11 - 25 %	54.1	45.9	512.456	2,797
	26 - 50 %	53.9	46.1	534.778	2,738
	more than 50 %	67.9	32.1	538.601	3,574
Disadvantaged	0-10 %	65.5	34.5	534.917	4,878
	11 - 25 %	56.5	43.5	524.088	3,281
	26 - 50 %	45.3	54.7	508.944	2,260
	more than 50 %	32.3	67.7	486.464	2,850
Desk	Yes	53.5	46.5	519.936	11,404
	No	47.9	52.1	501.787	1,865
Computer	Yes	53.4	46.6	519.268	12,461
	No	42.6	57.4	488.742	888
Books at home	0 – 10	43.7	56.3	463.687	1,363
	11-25	46.3	53.7	494.725	2,803
	26 -100	54.5	45.5	524.072	4,554
	101 – 200	57.4	42.6	535.928	2,425
	more than 200	58.8	41.2	546.106	2,124
FatherHighEduc	Yes	49.9	50.1	539.969	4,752
	No	57.7	42.3	504.821	8,517
MotherHighEduc	Yes	48.1	51.9	540.244	5,457

	No	59.4	40.6	501.457	7,821
JobFather	Never worked outside for pay	45.2	54.8	488.778	365
	Small Business owner	56.8	43.2	522.928	1,591
	Clerk	53.1	46.9	517.693	739
	Service or Sales Worker	51.7	48.3	512.658	1,251
	Skilled Agricultural or Fichery work	45.1	54.9	502.426	603
	Craft or Trade Workers	47.4	52.6	501.625	1,671
	Plant or Machinery Operator	45.9	54.1	503.727	1,804
	General Laborers	47.3	52.7	487.624	638
	Corporate Managers or Senior Official	59.6	40.4	534.246	1,825
	Professional	59.4	40.6	547.726	1,742
	Technician or Associate Professional	54.6	45.4	521.729	1,040
JobMother	Never worked outside for pay	42.8	57.2	486.039	796
	Small Business owner	54.5	45.5	517.635	1,037
	Clerk	54.3	45.7	525.725	2,255
	Service or Sales Worker	51.2	48.8	511.729	2,113
	Skilled Agricultural or Fichery work	47.2	52.8	496.004	406
	Craft or Trade Workers	50.5	49.5	498.073	489
	Plant or Machinery Operator	47.8	52.2	496.736	735
	General Laborers	46.5	53.5	498.745	1,289

	Corporate Managers or Senior Official	57.4	42.6	527.376	1,182
	Professional	59.2	40.8	546.736	2,023
	Technician or Associate Professional	54.7	45.3	523.928	935
FrenchAtHome	Always	48.7	51.3	522.491	9,709
	Sometime/Never	54.2	45.8	503.548	3,560
TimeRead	Less than one hour a week	51.2	48.8	498.028	4,010
	1 - 5 hours a week	53.1	46.9	526.726	5,753
	6 - 10 hours a week	54.1	45.9	526.912	2,273
	More than 10 hours a week	53.2	46.8	527.546	1,233
Sex	Male	52.5	47.5	513.207	6,734
	Female	52.9	47.1	521.567	6,535
Year	2006	51.8	48.2	522.896	4,330
	2011	55.2	44.8	521.656	4,331
	2016	51.3	48.7	508.002	4,628
Total	Total	52.7	47.2	517.409	13,269

Source: Calculation made with the PIRLS 2006, 2011, 2016 databases

Notes: The first 2 columns (Small Class, Large Class) show the percentage of students among groups that are allocated to small and large classes. The first line can be read as: among pupils who are in schools with 10% of students coming from disadvantaged homes, 38.1% are in large classes, and 61.9 % in small classes. The Score column depicts the average score per subgroup. For example, Males on average obtain a score of 513 (out of 700) on their reading test. The N column shows the sample size.