Why Do Smaller Classes Fail to Improve Student Learning? Quasi-Experimental Evidence from Surveys of Teachers, Parents and Students

Damon Clark* Pierre Deschamps† José Montalbán‡

September 26, 2025

Abstract

Across many settings, quasi-experimental studies find negligible effects of smaller classes on student learning. This is puzzling: teachers and parents favor smaller classes, and economists and education researchers have proposed mechanisms through which they should improve student outcomes - reduced classroom disruption (Lazear, 2001), greater use of individualized instruction (e.g., one-on-one help with math), and greater effectiveness of this instruction due to teachers' increased capacity to provide it (i.e., fewer students to help). We address this puzzle using administrative and survey data from Madrid, a large school district in which class size varies quasi-experimentally. We confirm that class size has negligible effects on student learning and then ask: (i) do smaller classes activate the proposed mechanisms? and (ii) do offsetting behavioral responses - such as teachers focusing on student enjoyment, or reductions in student effort or parent involvement - help explain the results? We find that class size has only modest effects on classroom disruption and the use of individualized instruction, and negligible effects on the effectiveness of individualized instruction. While we detect some evidence of offsetting behavioral responses, we calculate that even in their absence, smaller classes would have limited effects on student learning.

JEL: I21,I26,J24
Keywords: class size

*UC Irvine and NBER. Email: clarkd1@uci.edu
†SOFI, Stockholm University. Email: pierre.deschamps@sofi.su.se
‡SOFI, Stockholm University. Email: jose.montalban@sofi.su.se

1 Introduction

Across many settings, quasi-experimental studies find negligible effects of smaller classes on student learning. For example, in a recent meta-analysis of estimates collected from 66 studies, Opatrny et al. (2025) conclude that "(t)hese studies collectively suggest a negligible class size effect" (p.31).[1] This is surprising: Teachers and parents favor smaller classes; crucially, policy-makers around the world continue to use class size reduction as a tool for improving student outcomes.[2] These findings thus call attention to the question of why class size might fail to improve student learning. Evidence on this point could both advance our knowledge of the education production function and help policymakers to implement class size reductions more effectively - for example, by identifying complementary instructional practices. For instance, while some commentators suggest that the benefits of smaller classes will only be realized when teachers use particular types of instruction (Ehrenberg et al., 2001), this hypothesis remains untested.

Economists and education researchers suggest that smaller classes should improve student outcomes via three main mechanisms: reduced classroom disruption (Lazear, 2001), greater use of individualized instruction (e.g., one-on-one help with math), and greater effectiveness of this instruction due to teachers' increased capacity to provide it (i.e., fewer students to help).[3] Remarkably however, we have almost no evidence on whether smaller classes activate these mechanisms.[4] This has led some to speculate that smaller classes might fail to improve student outcomes because they do not activate these mechanisms (Ehrenberg et al., 2001; Opatrny et al., 2025). A separate possibility, discussed by Leuven and Oosterbeek (2018), is that smaller classes might trigger other offsetting responses that counteract their benefits. For example, teachers might use their extra capacity to make lessons more enjoyable, students might reduce their effort and parents might reduce their investments in their child's education (e.g., via decreased involvement). Prior work finds some evidence of parent responses (Fredriksson et al., 2016), but the evidence base remains thin, and teacher and

[1] Opatrny et al. (2025) note that Project STAR is an exception to this pattern, but view it as an anomaly when set against "dozens of quasi-experimental studies" (p.32) that suggest negligible effects. Leuven and Oosterbeek (2018) also review the recent literature and Rockoff (2009) provides a useful historical perspective on this literature. See Adusumilli et al. (2024) for a recent analysis of Project STAR.

[2] To take a recent example, the state of New York enacted a reform in 2022 that limits classes to no more than 20 students in Kindergarten through 3rd grade, 23 students in fourth through eighth grade and 25 students in ninth through twelfth grade. Districts have until 2028 to comply with these limits. In New York City, the reform is expected to reduce class size by around 4-7 students per grade at a cost of $1.3bn-$1.6bn per year (Chingos and Meltzer, 2023; Guarda and Subraminian, 2023)

[3] See for instance, Blatchford and Mortimore (1994) or Evertson and Randolph (1989).

[4] Betts and Shkolnik (1999) examine class size effects on teaching behavior, but focus on high schools (most studies focus on elementary schools) and do not leverage quasi-exogenous variation in class size.

student responses have been largely unexplored.

In this paper we seek to understand why smaller classes fail to improve student learning. We provide what we believe is the first rigorous assessment of how class size affects the key mechanisms thought to drive its effects: classroom disruption and the use - and effectiveness - of individualized instruction. For classroom disruption specifically, we offer the first empirical test of the Lazear (2001) model of education production. We also examine potential offsetting responses from teachers, parents and students. We gather all this evidence from a large school district where, consistent with other studies, we find that class size has at best small effects on student outcomes.

Our paper focuses on Madrid, a region with about 7 million people, where primary schools follow a maximum class size rule, which we use to isolate quasi-random variation in class size (Angrist and Lavy, 1999). Our key innovation is to combine administrative data on class size and student outcomes with large-scale surveys of students, parents, and teachers. The teacher and student surveys provide rich measures of the classroom learning environment, including the extent of classroom disruption and various indicators of individualized instruction. The student surveys include various proxies for student effort (e.g., homework, effort on tests) and student wellbeing (which we use to assess whether classes are more enjoyable). The student and parent surveys provide measures of parental investments, including parental involvement and the use of private tutors.

We begin by estimating the effects of class size on measures of student learning- standardized tests in math, Spanish and English taken by students in grade 3 and 6. Pooling students across the two grades and focusing on a student's average score across the three subjects, we estimate that a 5-student reduction in class size increases scores by around 0.01 standard deviations (SD). These estimates are fairly precise and we can reject effects larger than 0.04SD.[5] Splitting the sample, we find little evidence that effects vary across grade, subject, type of student (e.g., lower-SES versus higher-SES), type of teacher (e.g., more-experienced versus less-experienced) or type of school (e.g., in more-disadvantaged versus less-disadvantaged areas).

To understand these modest impacts, we turn to the mechanisms thought to drive the effects of class size on student learning. We begin with classroom disruption, which we measure using a teacher survey question about the extent to which "disruption of order in class" is a problem and a student survey question about whether there is a "good working

[5]These estimates are much smaller than those typically associated with Project STAR (Schanzenbach, 2020) and those reported by (Fredriksson et al., 2013), but are in line with more recent estimates from several countries (e.g., Angrist et al., 2017; Leuven and Løkken, 2020), including updated estimates of the effects of class size in Israel (Angrist et al., 2019).

atmosphere" in their classes. By measuring disruption from both the teacher and student perspective, we help to ensure that teacher-based estimates are not driven by experimenter demand effects (for example). Whether we use the teacher or the student measure, our results support the key qualitative predictions of the Lazear (2001) model. First, we find that a 5-student reduction in class size generates a 4 percentage point reduction in the probability that teachers report that disruption is a moderate or severe problem (off a base of 40%) and a 2.5pp increase in the probability that students report that there is a good working atmosphere in the class. Second, for both the teacher and student measures we find larger effects among schools characterized by higher baseline levels of disruption. However, these estimates are modest in magnitude (e.g., when compared to the effects of boys on disruption) and, based on estimates from the literature on the effects of disruption, we would not expect them to have large impacts on student outcomes.[6]

We next examine individualized instruction, the second key mechanism through which smaller classes are thought to improve student learning. Again, we measure this concept using questions from both the teacher and student surveys. Information derived from the teacher survey includes whether teachers agree with statements such as "I prepare a different task for those students who have learning difficulties and/or those who progress faster". Information derived from the student survey includes whether students agree with statements such as "They (the teachers) let me demonstrate what I have learned". Overall, across our various measures, we find little support for the claim that smaller classes promote more individualized instruction. Focusing specifically on small-group instruction - a type of individualized instruction highlighted in prior class size research (e.g., Ehrenberg et al. (2001)) - we find that smaller classes increase its use, but only marginally.[7] For example, we find that a 5-student reduction in class size increases the probability that teachers report using small-group methods by only 4 percentage points (off a base of about 70%) and increases the probability that students report working in small groups by around only 1.6 pp (off a base of 35%).[8]

Our finding that class size has little effect on individualized instruction mitigates potential "bad control" concerns and allows us to test the hypothesis that class size reductions only

[6]Studies of the effects on student achievement of the proportion of girls in a class contend that effects likely operate through reduced classroom disruption and find that a 10pp increase in the fraction of girls increases test scores by 0.01-0.03SD - effects in line with our class size estimates (Hoxby, 2000; Lavy and Schlosser, 2011).

[7]Small-group teaching is individualized in that it typically involves teachers moving between groups and helping individual students.

[8]The available evidence suggests small group instruction generates only modest or no improvements in student achievement (Fryer Jr, 2017; Bonesrønning et al., 2022)

improve learning when teachers use certain instructional practices (Ehrenberg et al., 2001; Blatchford, 2012). However, we do not find much difference in the effect of class size for teachers reporting that they use individualized teaching practices and those that do not, suggesting that even when teachers regularly use small group methods, smaller classes fail to improve student learning.

These findings on classroom disruption and individualized instruction help explain why we find limited effects of class size on student learning. Another possibility is that class size might affect the learning environment in ways that are not captured by our measures, while simultaneously triggering offsetting responses by teachers, students and parents. We investigate this possibility but find limited evidence for it. For teachers we analyze students' satisfaction with teachers - a proxy for whether teachers make lessons more enjoyable. We find no evidence that satisfaction increases in smaller classes. For students, we examine effort measures including time devoted to homework (as reported by both students and parents) and self-reported effort on standardized tests. We find that smaller classes lead students to reduce effort, but effects are small. For example, a 5-student reduction in class size increases homework time by only 34 minutes per month. For parents, we analyze various investments in their child's education, including involvement and use of private tutors. Again, we find that smaller classes reduce these investments, but effects are small.

We conclude that at least in this setting, small classes have weak effects on the mechanisms through which they are expected to improve student learning. To help fix ideas about how this might be possible, we discuss various hypotheses. First, in relation to classroom disruption, we extend the Lazear (2001) model to allow disruption to depend on class size and teacher effort (which could in turn depend on class size). We show that smaller classes could lead teachers to reduce effort, thereby negating their effects on disruption.[9] Second, in relation to individualized instruction, and why this is not productive in smaller classes, the class sizes in our sample (20-30 students) may still be too large for these methods to show clear advantages. This lack of productivity advantage could help to explain why teachers do not make greater use of these practices when class sizes are smaller. Other possibilities are that teachers may lack the necessary training or the necessary incentives. Our findings and this discussion of hypotheses focus attention on precisely how smaller classes are expected to improve student outcomes and whether auxiliary policies (e.g., teacher training) could effectively complement class size reduction policies.

Our paper is organized as follows. In the next section we sketch a simple conceptual

[9]This is a specific example of a more general point made by Leuven and Oosterbeek (2018), that teachers exposed to smaller classes could exert less effort.

framework. This lays out the theory for why smaller classes could improve student outcomes and provide a roadmap for our investigation of why smaller classes generally fail to improve them. In Section 3 we describe the Madrid setting and in Section 4 we describe our data and our empirical strategy, which closely follows the prior literature (e.g., Angrist and Lavy (1999). This section also establishes the validity of this empirical strategy, for example by showing that our instrument for class size is uncorrelated with measures of teacher effectiveness or classroom student composition. In Section 5 we analyze the impacts of class size on student outcomes. In Section 6 we analyze the impacts on the key mechanisms of interest - classroom disruption and individualized teaching. Section 7 examines the evidence for offsetting responses to smaller classes. Section 8 discusses our findings and concludes.

2 Conceptual Framework

In this section, we draw on existing literature to highlight the mechanisms through which class size is most likely to affect student outcomes. We use a simple model to formalize these mechanisms, discuss other forces that might limit class size effects, and briefly preview our empirical strategy for investigating these relationships.

2.1 Mechanisms

A vast literature has studied the effects of class size on student outcomes, with many studies enumerating potential mechanisms. While some research has examined the effects of class size on these mechanisms, this evidence is almost entirely correlational.

Three potential mechanisms have received the most attention in this literature. First, the possibility that smaller classes reduce the extent of classroom disruption. As discussed by Lazear (2001), disruption could be driven by student misbehavior but also via other channels (e.g., students asking questions that are not understood by other students in the class). Second, the possibility that instructional activities are more effective in smaller classes - and especially activities that involve interaction between teachers and individual students (e.g., one-on-one help, small-group work). We refer to these activities as "individualized instruction". Third, the possibility that small classes lead teachers to devote more time to individualized instruction, presumably because these activities become more effective when classes are smaller. These mechanisms also feature prominently in class policy debates.

2.2 Model

To formalize these mechanisms, we can write down a simple model of education production in the spirit of Lazear (2001). First, we assume that student learning over a period of time can be described by the equation

$$L = L_u(1 - d) \tag{1}$$

where L_u is the learning that would have occurred if the class was never disrupted and d is the fraction of time it is disrupted. Following Lazear (2001), we assume that:

$$(1 - d) = b^n \tag{2}$$

which depends on the probability that a student does not disrupt the class during any given interval of time (b) and the number of students in the class (N).

Second, we assume that learning is produced by two types of activities: individualized instruction as defined above and non-individualized instruction that does not involve interactions between teachers and individual students (e.g., whole-class teaching, individual study). Third, we assume that the effectiveness of individualized instruction depends on class size whereas the effectiveness of non-individualized instruction does not. Intuitively, if a teacher has students work in small groups and circulates to help individual students, then as class size expands, the average time per student spent on one-on-one decreases. In contrast, non-individualized instruction seems easier to scale. Finally, we assume a simple functional form for the effectiveness of individualized instruction and we assume that these two types of activities are combined in a CES production function. This gives us:

$$L_u = \left[(A(1 - \pi))^\rho + \left(\frac{B}{N}\pi\right)^\rho\right]^{\frac{1}{\rho}} \tag{3}$$

where A and B are parameters that govern the productivities of the two activities, π is the fraction of time devoted to individualized instruction and $\frac{1}{1-\rho}$ is the elasticity of substitution between these actitivies.

If we assume that teachers choose π to maximize student learning, then we can show that

$$\pi^*(N) = \frac{1}{1 + \left(\frac{AN}{B}\right)^{\frac{1}{\rho-1}}} \tag{4}$$

It follows that smaller classes lead to increased individualized instruction provided the two sets of activities are substitutes (i.e, the elasticity of substitution exceeds one).

Finally, differentiating with respect to class size yields

$$\frac{dL}{dN} = [\underbrace{\frac{\partial L_u}{\partial N}}_{\text{more effective instruction}} + \underbrace{\frac{\partial L_u}{\partial p} \cdot \frac{dp^*}{dN}}_{\text{more individualized instruction}}](1-d) - \underbrace{L_u \frac{\partial d}{\partial N}}_{\text{less disruption}}$$

This expression maps to the three channels through which class size could improve student learning. First, small classes are less likely to be disrupted (i.e., $\frac{\partial d}{\partial N} < 0$). Second, small classes mean that individualized instruction becomes more effective, such that $\frac{\partial L_u}{\partial N} < 0$. Third, depending on the substitutability of individualized and non-individualized instruction, teachers could reallocate time towards more individualized instruction, such that $\frac{\partial L_u}{\partial p} \cdot \frac{dp^*}{dN} < 0$.[10]

A corollary of the second channel is that, other things equal, the teachers that tend to do the most individualized instruction (conditional on class size) will enjoy the largest benefits of smaller classes. Indeed, if we extended the model to allow the probability of disruption to depend on the type of teaching activity used, and if we assumed that non-individualized instruction was never disrupted, then teachers that use non-individualized instruction will not enjoy any benefits of smaller classes. This was a point made by Ehrenberg et al. (2001) and also Blatchford (2012).

2.3 Discussion

While our formal model illustrates the three key mechanisms through which class size may affect student learning, the actual relationship between class size and these mechanisms remains an empirical question. First, the functions relating learning efficacy and disruption to class size likely differ from our stylized assumptions. Most notably, these relationships might be relatively flat within the range of class sizes (20-30 students) observed in our data. Second, smaller classes could trigger behavioral responses that offset the benefits captured by the model. For example, teachers might exert less effort or decrease the instructional time devoted to core academics, students might reduce their academic effort and parents might invest less in their child's education (e.g., by reducing their involvement).

[10] In that case we would expect $\frac{\partial L_u}{\partial p} > 0$, meaning that more individualized instruction raises learning. Otherwise, as class size fell, the returns to individualized instruction would not increase, and teachers would not increase individualized instruction.

2.4 Looking ahead

Guided by this general framework, our analysis proceeds in four steps, each of which leverages the quasi-experimental variation in class size in our data. First, we estimate the effect of class size on student learning and confirm that these effects are small in our context. Second, we estimate the effect of class size on classroom disruption. Third, we estimate the effect of class size on proxies for teachers' use of individualized instruction, and examine how the effectiveness of this instruction varies with class size. Finally, we estimate the effects of class size on various offsetting behavioral responses by teachers, students and parents - factors that could help explain why the overall effects are small.

3 Institutions and Data

3.1 Institutions

Compulsory education in Spain spans ages 6 to 16. This period covers six years of primary-stage education (*1º de Primaria* through *6º de Primaria*) and four years of secondary-stage education (*1º de ESO* through *4º de ESO*). The state school system serves about 85% of compulsory school-aged children, with the remaining 15% attending private schools. Within the state system, public schools enroll about 60% of students and semi-public schools (*Centros Concertados*) enroll the remainder. Public schools are government funded and managed by civil servants, while semi-public schools are privately operated but receive most of their funding from public sources. Most state schools specialize in either primary- or secondary-stage education.

Like most class size research, our study focuses on primary schools, specifically state primary schools.[11] Student assignment to these schools occurs through a centralized choice system similar to that used in many other countries (Neilson, 2024). Parents submit enrollment applications with rank-ordered school preferences, and the Region of Madrid assigns students using an algorithm that considers these preferences, school capacities, and priorities (e.g., sibling attendance). The algorithm is based on a version of the so-called Boston mechanism (Abdulkadiroğlu and Sönmez, 2003). Schools also offer free universal access to three years of non-compulsory infant-stage education (*Educación Infantil de segundo ciclo* for

[11] At the secondary school level, there is substantial variation in class size both across and also within subjects (e.g., because students are more likely to be tracked). As a result, the average size of a student's class in a particular subject is only weakly correlated with our measure of predicted class size (results available upon request). Private school classes tend to be small and hence the maximum class size rule is largely irrelevant.

children aged 3-6), with most families applying to enter those earlier grades. Once enrolled, students maintain priority over new applicants.

Within each school, students in Primary Education at each grade level (*curso*) are divided into classes (*grupos clase*), with each class assigned a single teacher (*maestro tutor*) responsible for teaching all core subjects. The number of classes - and thus teachers - is determined by enrollment and the maximum class size rule. Initially, the maximum class size rule was set at 25 students (this rule is laid out in the article 157.1(a) of the national *Organic Law LOE 2/2006*). However, it was allowed to exceed this by up to 10 percent to account for multiple factors such as late enrollment or family re-locations (see Article 87 of the national *Organic Law LOE 2/2006*). A 2012 reform expanded this allowance to 20% for all our period of analysis, such that the effective maximum size for primary school classes during our sample period was 30 students (see Article 2 of *Royal Decree 14/2012)*. For example, a grade with 29 students requires one teacher and forms one class, while a grade with 31 students requires two teachers and two classes. In the first case average class size would be 29, in the second it would be 15.5. Later in this section we show that compliance with this rule is generally high.

3.2 Data

Our paper uses administrative data linked to surveys of students, teachers, principals and parents. The administrative data include an annual schools database that contains geographic coordinates and indicators for school type (e.g., public or semi-public). Crucially, the administrative data also include information related to the number of students enrolled, class size and test scores. We discuss these data in more detail below, and then describe the survey data that we link to.

3.2.1 Class size

We have access to school enrollment records from 2004 to 2019, which contain information on Spring enrollment and the number of classes taught in each grade. We use these records to construct average class size in each grade and year.[12]

[12]These data are held by the Deputy Directorate General for Evaluation and Analysis (*Subdirección General de Evaluación y Análisis*). The data are collected in February and March; schools enter them into an integrated data management application known as *Raices*. In cases where schools fail to provide the required information or errors are identified, the Deputy Directorate General for Evaluation and Analysis directly contacts the schools to rectify these data discrepancies. The data are made publicly available via descriptive reports published on the Region of Madrid Statistics website. See `https://estadisticas.educa.madrid.org/`.

As discussed below, our empirical analysis will focus on class size in grades 3 in 6 over the period 2016-2019, the years for which we have access to standardized test scores. For these grades and years, Figure 1 uses our administrative data to plot the empirical relationship between average class size and enrollment for grades 3 and 6. The unit of observation is a grade-school-year (e.g., 3rd grade at School A in 2016); each point represents the average class size at each enrollment level. We also superimpose the relationship that we would see if schools fully complied with the maximum class size rule (the solid line). As expected, the empirical relationship closely follows the "saw-toothed" pattern implied by the maximum class size rule. For example, as enrollment increases towards 30, average class size increases almost one-for-one; when school enrollment exceeds 30 and a second class has to be added, average class size drops sharply. This pattern repeats when enrollment crosses 60 (when a third class must be added) and again at 90 (when a fourth class must be added). At each of these thresholds, there are small discrepancies between actual average class size and the average size implied by a strict application of the maximum class size rule. Nevertheless, Panels A and B of Figure 1 suggest that compliance with the rule is generally high.

3.2.2 Test score data

Over the period of our study (2016-2019), Spanish law required that every student in Spain was tested at the end of primary education (6th grade) and secondary education (10th grade). Regional governments were responsible for the final design, implementation and grading of the tests, but were subject to national regulations designed to ensure common standards across the country. To comply with this new regulation, the regional government of Madrid introduced standardized testing (LOMCE) during the 2015-2016 school year. With limited exceptions for some students in special education programs, all students in grades three, six and ten are tested in mathematics, Spanish and English at the end of the academic year.[13] The test results are intended to provide teachers, parents, and students with additional information on students' relative performance and have no academic consequences.

We have test records for students in grades 3 and 6 from 2016 to 2019. Each record contains a student identifier, as well as information on school, grade and class (*grupos clase*). For each student we average their scores across mathematics, Spanish and English to construct our main test score outcome measure (we check robustness to using individual subject scores). These scores are standardized at the grade-cohort level to have a mean of zero and

[13]The mathematics test consists of multiple-choice questions, while the English and Spanish tests have two parts: linguistic competence (multiple-choice test) and written communication (written test). Additionally, students are tested in science in grade 6 and in "social and civic competence" (which mainly covers history) in grade 10.

a standard deviation of one.

3.2.3 Survey data

We have data from surveys of students, parents, teachers and principals.

Student Survey. In each year from 2016 to 2019, surveys were completed by all tested students in grade 6. The survey instrument was the same in each year and contained questions related to the home environment, homework and students' attitudes towards their teachers and their school. Students completed a pen-and-paper survey in 2016 and online surveys from 2017 onwards. The response rate was over 95% in 2016 and around 70% for the later surveys.

Parent Survey. In each year from 2016 to 2019, surveys were completed by parents of all tested students in grades 3 and 6. Parents completed a single survey for each child in the tested grades. The survey instrument was the same in each year and contained questions related to their parents' socioeconomic background, their child's homework and their satisfaction with their child's school. Like students, parents completed a pen-and-paper survey in 2016 and online surveys from 2017. The response rate was around 70% in 2016 and lower in subsequent years (about 45, 36 and 30 percent).

Teacher Survey. In each year from 2017 to 2019, surveys were completed by the teacher of each tested classroom. As discussed in the next section, this means that a single teacher is likely to complete multiple surveys (e.g., one for mathematics, one for Spanish and one for English). The teacher surveys contain classroom-specific information (e.g., on the teaching methods used). They also contain more general information on teacher characteristics (e.g., age, gender, years of experience, type of contract) and teacher attitudes.

Principal Survey. In each year from 2016 to 2019, surveys were completed by the principals of all schools with students in tested grades. These surveys contained general information on the school environment as well as principal characteristics and attitudes.

3.2.4 Linking Survey to Administrative Data

For all years except 2018, we can link student and parent surveys to the student identifiers in the test data. We can also link teacher surveys to the class identifiers in the test data. Since the 2018 student and parent surveys contain neither student nor class identifiers, the linking process is more complicated. Even without explicit identifiers, we can link the 2018 student

surveys using information available in both the survey and test datasets.[14] However, absent further sample restrictions, the teacher surveys for 2018 can be matched to the school but not the grade. This means that we cannot determine whether the teacher teaches in grade 3 or grade 6. We thus exclude the 2018 data from our teacher analysis.

3.2.5 Descriptive Statistics

We present some descriptive statistics on the schools, students and teachers in our dataset in Table 1, for grades 3 and 6 separately. Our dataset is relatively large - we have over 220 000 student-grade observations, 4,500 classes and 1,200 schools by grade. These large sample sizes allow us to give precise estimates of the effects of class size on test scores and other potential outcomes. The average number of students enrolled at the school-grade level is 55.57, and the average number of classes within school-grade is 2.2. In our dataset, about 10% of students are non-Spanish, and over 75% of fathers hold Spanish nationality. As in other contexts, the vast majority of teachers are female. Teachers have an average of more than 14 years of experience, with nearly all employed full-time, and about 64% serving as civil servants.

4 Empirical Strategy

We use our linked administrative and survey datasets to examine the effects of class size on various outcomes derived from the test score data, from student and parent surveys and from teacher surveys. In each case, the core empirical strategy is the same: to use class size variation driven by maximum class size rules to isolate the causal effects of class size on these outcomes. However, there are specific issues associated with each type of outcome and so we discuss them separately.

4.1 Test outcomes

A naive approach to estimate the effect of class size on student achievement could be to simply regress observed class size on standardized test scores. However, as discussed by Lazear (2001), we expect class size to be endogenous, biasing OLS estimates.[15]

[14] In 2018, the test data include covariates also present in the student survey -gender, absenteeism, well-being, and digital technology use- which we employ to match students one-to-one across the survey and test datasets.

[15] When the outcome is student achievement, Lazear (2001) argued that the bias would likely be towards zero, since schools would tend to group less-disruptive (and higher-ability) students into larger classes.

Building on the influential contribution of Angrist and Lavy (1999), we address this bias by using the maximum class size rules in effect in Madrid during our study period (c.f., Figure 1). There are at least two ways to use the variation generated by maximum class size rules to estimate the causal effects of class size on test scores. First, one could use regression discontinuity (RD) methods to exploit the discontinuities in the relationship between average class size and enrollment at the various thresholds (i.e., 30, 60, 90, 120). Second, after controlling for a flexible function of enrollment, one could instrument average class size with the average class size that would follow from a strict application of the maximum class size rule. Intuitively, this "rule-based" approach exploits the discontinuities and kinks at each of these thresholds.[16] Both approaches have been used in prior work, and we would expect them to generate similar estimates. We employ the rule-based approach, previously used by Angrist et al. (2017, 2019), in order to maximize statistical power.

To implement the rule-based approach we first present the following first-stage equation:

$$CS_{gst} = \pi_0 + \pi_1 Z_{gst} + f(r_{gst})\alpha + X'_{gst}\delta + \eta_{igst} \tag{5}$$

where $Z_{gst} = \frac{r_{gst}}{floor(\frac{r_{gst}-1}{30})+1}$ is the average class that follows from strict application of the maximum class size rule. The variable CS_{gst} denotes the average class size in grade g at school s and year t (as discussed above), r_{gst} is enrollment, our running variable, X_{gst} is a vector of control variables, and ϵ_{igst} is a mean-zero error term. Following Angrist et al. (2017, 2019) we present alternative functions of the running variable, r_{gst}, including linear and quadratic functions of enrollment and a piecewise linear trend.

We then specify the following outcome equation:

$$Y_{igst} = \beta_0 + \beta_1 \widehat{CS_{gst}} + f(r_{gst})\lambda + X'_{gst}\gamma + \epsilon_{igst} \tag{6}$$

where Y_{igst} represents a test outcome associated with individual i in grade g (3 or 6) at school s in year t, and $\widehat{CS_{gst}}$ is the instrumented class size variable. The parameter β_1 represents the causal effect of a one-student increase in class size on outcome Y_{igst}, and is estimated through 2SLS. In our main specifications the vector X_{igst} will include school type, district fixed effects, grade fixed effects, and year fixed effects.[17] We cluster standard errors at the school level (i.e., allow for arbitrary correlation between the error terms of students attending the same school).

[16] Since the slope of the relationship between average class size and enrollment changes at these thresholds, a regression kink design could also be used to identify class effects (Card et al., 2015).

[17] Some estimates will exploit the link to student and parent surveys and include student background characteristics derived from those (e.g., parental education).

Although this approach is fairly standard, it is worth noting that our endogenous class size variable CS_{gst} refers to class size in the particular grade being tested (3 or 6).[18] Since in our setting, enrollment and therefore class size is positively correlated across grades within a cohort, our estimates are also likely picking up the cumulative effects of being in a small class in earlier grades. We choose not to address this issue for three reasons. First, while this suggests that our test score estimates likely overstate the effects of class size on test scores, our estimates turn out to be very small. Second, we do not have student panel data and hence cannot construct a student-level measure of the average class size experienced over the first three grades (for grade 3) or the first six grades (for grade 6). We can construct these measures at the cohort level (i.e., we know the average class size experienced by a student enrolled in the school for the first three grades), but we do not have any data on when students joined the school and hence could not follow Leuven and Løkken (2020) and exploit information on exposure to these cohort-level average class sizes. Third, many of our other outcomes (discussed below) are defined in terms of the particular class (e.g., teaching methods used), and hence class size in the testing grade is the relevant concept. Using that measure for our test score outcomes makes for a consistent presentation across each of the outcomes studied.

It is also worth discussing that we construct our instrument using enrollment in the tested grade (i.e., 3 or 6), not the entry grade. One could argue that the instrument would be more convincing if it was based on entry grade enrollment, as it often is when the endogenous class size variable refers to average class size across a number of grades. We have experimented with this approach. As expected, the first stage is weaker but the estimates are qualitatively similar. Since we find no evidence to suggest that parents move in and out of schools on the basis of class size in Madrid, and since our instruments based on test grade enrollment passes the usual checks (discussed below), we use this instrument in our main specifications.[19]

4.1.1 First-stage

With Panels A and B of Figure 1 in mind, we would expect to find a strong relationship between the average class size observed in our data and the average class size that would follow

[18] Note also that our endogenous class size variable CS_{gst} is at the cohort rather than the student level. At first glance this might suggest that our estimates could be biased by (non-classical) measurement error in this variable (i.e., grade-level class size is student-level class size plus some measurement error). However, because the instrument is also defined at the level of the cohort, it will be uncorrelated with the measurement error and hence our first-stage estimates will be consistent (but less precise than they would be were a student-level measure available.)

[19] We are currently working to obtain the data necessary to implement the approach of Leuven and Løkken (2020).

from a strict application of the maximum class size rule. The "first-stage" estimates reported in Table 2 confirm this. The estimates in columns (1) and (4) indicate that a one-student increase in predicted average class size leads to a 0.77-student increase in observed average class size in third grade and a 0.8-student increase in sixth grade, respectively. Estimates are similar whether we control linearly for enrollment (columns 1 and 4), second-order polynomial in enrollment (columns (2) and (5)) or a piecewise linear trend (columns (3) and (6)). In the rest of paper, we follow Angrist et al. (2017) and control for both linearly for enrollment and the second-order polynomial for the main results. Since these first-stage estimates are relatively precise (the implied F-statistics are over 200), our 2SLS estimates should be free from the finite-sample bias associated with "weak instruments".

4.1.2 Validity

We next discuss four potential threats to our identification strategy.

Threat #1: Schools manipulate enrollment. One threat to empirical strategies based on maximum class size rules is schools manipulating enrollment relative to the threshold. For example, Urquiola and Verhoogen (2009) provides striking evidence of this in the Chilean context. Since school funding in Madrid depends partly on the number of classrooms and students, the most obvious concern is that schools would manipulate enrollment to be just above the threshold, where average class size is minimized. As a first check on the possibility that schools manipulate enrollment in the tested grades, Figures A1 and A2 plot histograms of enrollment around the key enrollment thresholds for grades 3 and 6. Both panels show that the density of schools increases as it approaches the historical maximum class size of 25 students, reflecting capacity constraints in schools typically designed for only one or two classes per grade (see Table 1). There is little to suggest that schools are manipulating enrollment to be on one side or the other of these thresholds.[20] In addition, schools could potentially adjust subsequent enrollment (e.g., via grade retention) to target the cutoff in later grades. As a check on the possibility that schools manipulate enrollment in the tested grades, we use the retention registry to estimate class size effects on grade retention. Table A4 shows that the impacts of class size on grade retention are negligible and statistically indistinguishable from zero.

[20]Note that there are relatively few schools around the third and fourth cutoffs, such that there are some "holes" in those distributions.

Threat #2: Parents move in and out of schools based on the cutoff. Another threat to the empirical strategy is families moving schools to take advantage of smaller classes. Since families can choose schools in Madrid, this may also be a relevant concern. As noted above, both of these threats could be amplified by our decision to construct the instrument using enrollment in the tested grade. We provide two tests that suggest little evidence to this source of manipulation. First, as noted above, Figures A1 and A2 show little evidence of schools manipulating enrollment. As a second check on the possibility that schools manipulate enrollment, we report a series of "balance tests". We implement these using the rule-based IV approach that we use to estimate class size effects on outcomes. Specifically, we estimate a version of equation (5), with student characteristics, such as gender, parents' years of education or occupation, as the dependent variable. We also use as an outcome a student's test score predicted by background characteristics reported in the student and parent surveys. If popular schools could manipulate enrollment more successfully (e.g., due to high demand) and attract higher-scoring students, our estimates would be biased towards finding positive effects of class size reductions. Table A1 presents the results of one such test. The coefficients on the instrument are small and statistically indistinguishable from zero, providing no evidence of any enrollment information. Separate balance tests on a series of background characteristics point to the same conclusion, as do RD-based balance checks.

Threat #3: Schools assign less-effective teachers to smaller classes. A third threat to the empirical strategy is schools assigning less or more effective teachers to smaller classes, which would lead to a failure of our exclusion restriction. This important hypothesis is advanced by Gilraine (2020): that maximum class size rules could induce variation in both class size and teacher quality. This is because schools must add a class if enrollment increases from below to above a threshold (e.g., from 30 to 31 students), or remove one if it shrinks from above to below. If adding a class means hiring a new teacher, and removing a class means releasing an existing teacher, and if new teachers are less effective than released teachers, then smaller classes will be associated with less effective teachers. To test this, Gilraine (2020) reports separate estimates of the effects of adding and removing a class. Consistent with the hypothesis, the effects of class size reduction implied by the "add-a-class estimates" are found to be smaller than those implied by the "remove-a-class estimates". To address this concern in the Madrid context, we implement the approach devised by Gilraine (2020) and generate separate estimates of the effects of adding and removing a class. These estimates suggest that adding and removing a class have similar implications for the effects

of class size reduction on test scores (i.e., both imply relatively small effects).[21]

In addition, we conduct a similar exercise as the previous "balance tests" to examine whether teachers' predetermined characteristics, which could serve as proxies for effectiveness, are systematically correlated with our instrument. Table A2 presents the results. For most regressors, such as teacher gender, type of contract, or employment type, the coefficients on the instrument are small and statistically indistinguishable from zero. We find that the coefficients for teacher age and years of experience are statistically significant for Grade 6. However, the magnitudes are very small: a 5-student increase in class size corresponds to only a six-month increase in teacher age and experience, respectively, over a baseline mean of 43 years and more than 16 years of experience. We conclude that in the Madrid context, maximum class size rules induce variation in class size but not teacher quality.

Threat #4: Schools use more classes as an opportunity to track. A fourth threat to our empirical strategy is that schools might use additional classes as an opportunity to track students (e.g., class A for "strong" academic students and class B for "weak" students), which would violate our exclusion restriction. However, formal tracking that separates students into different pathways based on academic level is not permitted in Primary Education in Spain, making it unlikely that this occurs in practice.[22] Nevertheless, we conduct an exercise analogous to the previous "balance tests" to examine whether the class-level variance in student characteristics is systematically correlated with our instrument. Table A5 provides little evidence supporting this hypothesis.

4.2 Outcomes based on student and parent surveys

Outcomes derived from the student and parent surveys are analyzed in the same way as test outcomes: using one observation per student and implementing the IV procedure described by equations (5)-(6). One issue specific to those outcomes though is that a fraction of student and parent surveys were not completed, such that analyses based on these surveys could be biased by selection into survey completion. However, applying the strategy described above to

[21] Results on the approach devised by Gilraine (2020) are available upon request.

[22] In Primary Education in Spain, schools are not allowed to: (i) Create permanent tracks or parallel classes with different levels of academic demand (see *Royal Decree 157/2022*, curriculum framework for Primary Education; and *BOCM 23/2023* of Madrid, Article 8 on flexible, temporary groupings only). (ii) Systematically segregate students by academic performance, as this is contrary to the principle of equity and the compensatory role of public education (see *LOMLOE 3/2020*, preamble and general principles of inclusion and equity). (iii) Anticipate academic selection through differentiated pathways, which is only possible later in Secondary education (*ESO*) with specific programs such as *diversificación curricular* or *FP Básica* starting in third-fourth year (see *LOMLOE 3/2020, Title I, Chapter II on Secondary Education*).

variables indicating whether students and parents completed the surveys, we find no evidence of class size effects on completion rates (see Table A6). We therefore conclude that this is not an important concern.[23]

4.3 Outcomes based on teacher surveys

Completion rates of teachers could also be a concern for the same reason as above. However, when we apply the same strategy described above to variables indicating whether teachers completed the surveys, we find no evidence of class size effects on completion rates (see Table A7). Completion rates for teacher surveys are nearly universal. As noted above, the same teacher completes multiple surveys. For example, for the typical setup in which students in a grade are divided into classes (*grupos clase*) and assigned a single teacher (*maestro tutor*) for all three subjects, we would expect that teacher to complete three surveys for the three different subjects (but all referring to the same class of students).

We could deal with these data in two ways. First, by making the teacher survey the unit of observation, such that most teachers will contribute three observations. Second, by constructing teacher-average responses and making the teacher the unit of observation. In both cases the class size variable and the instrument would be the same (i.e., those defined in equations (5)-(6)), and hence we would expect the two approaches to generate similar estimates (we have to checked to confirm this is the case). We chose to make the teacher survey the unit of observation since some of our analyses explore whether survey responses vary by subject (e.g., relating to teaching methods). As with all of our other analyses, the standard errors are clustered at the school level.

5 Impacts on Student Test Scores

To begin our analysis of the effects of class size on test scores, we plot the relationship between test scores and school enrollment. Panel A and B of Figure B1 show the relationships for grades three and six respectively. Test scores are averaged across math, English, and Spanish. The unit of observation is grade-school-year, and the points on the graph are averages across 1-student enrollment bins. To emphasize the main patterns, we overlay linear fits in the enrollment ranges of 0-30, 31-60, 61-90, and 91-120.[24]

[23] See Appendix A for the analysis of survey completion.

[24] These are based on regressions of average test scores on school enrollment at the grade-school-year level. These regressions are not weighted by enrollment, partly because there is limited variation in enrollment within these ranges.

If smaller classes led to higher average test scores, we would expect average scores to "jump up" at enrollment thresholds of 30, 60, and 90. There is little evidence of that pattern for either grade. Instead, the only consistent pattern is that test scores increase with enrollment. This is likely because larger schools (with higher enrollment) tend to be located in areas of Madrid (typically the suburbs), where, on average, families have higher socioeconomic status and students have higher test scores.[25]

Table 3 reports the OLS and our 2SLS estimates of the effects of class size on average test scores (i.e., the same outcome shown in Panels A and B of Figure 1). Column (1) shows the corresponding OLS estimates, which display a negative correlation between class size and test scores, consistent with the previous observation that larger schools (and hence, class sizes) tend to be located in higher socioeconomic areas with students achieving higher test scores. Columns (2)–(4) report the 2SLS estimates of the effects of class size on average test scores. The estimates are negative, suggesting that class size reductions improve test scores. However, these estimates never reach statistical significance and the effect sizes are small. For instance, the estimates in column (2) implies that a five-student decrease in class size would increase test scores in both grades by less than 0.0125SD. These estimates align with Panels A and B of Figure B1, which suggest that any effect of class size on test scores is, at best, small. Pooling grades 3 and 6 to maximize statistical power, we can reject that a 5-student reduction in class size would improve average test scores by more than 0.04SD.

It is interesting that we find similarly negligible and non-significant results across both grade levels.The effects for Grade 6 are slightly larger, though they remain negligible in both magnitude and statistical significance. This is consistent with the hypothesis of cumulative positive effects of smaller class sizes on test scores over years of exposure. Table C5 shows that we also find similar results when estimating separate effects for different subjects. Since instructional practices may vary by grade and by subject, and class size effects may depend on instructional practices, this is a first piece of evidence against the hypothesis that class size effects depend strongly on instructional practices. We revisit this point below.

We also investigate whether class size effects vary by school type - for example, by the proportion of lower-SES students enrolled, or by the underlying disruption in the school, as reported by principals. One hypothesis advanced by Krueger and Whitmore (2001) suggests that class size reductions benefit black students more than white students because black students might "attend schools that have a disproportionately high number of disruptive students, or students with special needs, which distracts their teachers from instructional

[25] Angrist and Lavy (1999) document a similar pattern and also attribute it to the relationship between school size and location.

time" (p.20). With this in mind, we estimate separate effects for low-SES and high-SES schools, defined by average predicted test scores at the school level. Consistent with the student-level results, the effects are similar across the two types of schools (see columns 7 and 8 of Table C5).

We present two additional tests in columns 9 to 12 of Table C5). First, we estimate separate effects for students in schools with high underlying disruption levels, as described by principals. Consistent with theory, we find point estimates for class size that are three times larger in schools with above median disruption - but the magnitudes of these effects remain low. Similarly, we split the sample by teacher experience, taking the median of the sum of teacher experience in Spanish, English and math as a cut-off point. We find that teacher with higher levels of experience seem better equipped to deal with larger class sizes, but again effect sizes remain modest and statistically insignificant. [26]

6 Impacts on Key Classroom Dynamics

Why do smaller classes lead to small, if any, improvements in test scores? One possibility is that changes in class size do not meaningfully change key elements of the classroom environment. In this subsection we evaluate this hypothesis.

6.1 Classroom disruption

The mechanism most often used to rationalize why smaller classes would improve test scores is classroom disruption. Lazear (2001) formalized this mechanism and assumed that classroom disruption is a function of class size and students' propensity to be disruptive. However, as noted by Ehrenberg et al. (2001), there is scant "causal" evidence on the relationship between class size and disruption, so this relationship could be much weaker than is often thought.

[26] We also present results using machine learning to explore heterogeneity in treatment effects. We use causal forests to explore which variables show the strongest reaction to our treatment, the maximum class size assignment rule. The standard method relies on common support of the explanatory variables, which is violated by design in our case: one of our key controls, enrollment is entirely co-determined with our treatment. We therefore tweak the standard method by first residualizing our instrument (maximum class size rule), treatment variable (actual class size rule) and dependent variable (test scores) with the set of controls we impose in all regressions (enrollment, enrollment squared etc.). We then run an instrumental forest using the SES and disruption level of the school, and all the explanatory variables at the individual level as potential drivers of heterogenous effects separately for each grade. We impose a maximum node size of 50. Overall we see very low treatment heterogeneity, as seen in Figure C2. Plotting the predicted average treatment effect for our key explanatory variables in Figures C3 shows no systematic pattern in heterogeneity across grades, or subgroups of particular variables with large treatment effects. Overall, our analysis suggests that there seems to be little treatment effect heterogeneity in our data.

As discussed above, our teacher and student survey data yield proxies to measure classroom disruption. The teacher questionnaire asks teachers to what extent "disruption of order in class" is a problem, and students about whether there is a "good working atmosphere" in class.[27] We use these two measures to estimate the effect of class size on classroom disruption. Table 4 reports 2SLS estimates based on the same model used for test scores (i.e., 2SLS applied to equations 6 and 5). The teachers' estimate shows a 5-student reduction in class size would reduce the probability that disruption is a moderate or serious problem by 4 percentage points in grades three and six, on a base of around 40%. The grade six students' estimates suggest a 5-student reduction in class size would reduce the probability that there is a good working atmosphere in class by 2.5 percentage points in on a base of around 23%. Although the questions posed to teachers and students differ and are therefore not directly comparable, both sets of estimates indicate that a 5-student reduction in class size would reduce classroom disruption by roughly 10% relative to the baseline mean.

We might expect these effects to be larger in some schools than others. For instance, there are stronger effects of class size on disruption in schools with higher levels of underlying disruption in the model presented by Lazear (2001). We construct a school-level index of underlying disruption using the principal survey. This index combines nine items posed to the school principal regarding disruption, including classroom disorder, physical aggression, bullying among students, and disrespect toward teachers. These items refer to the school as a whole, rather than to a specific grade or classroom.[28] We test whether class size has stronger effects on disruption in schools with higher levels of underlying disruption by splitting schools into those classified as having below- and above-median levels of underlying disruption. Note that the school-level index of underlying disruption is uncorrelated with our instrument (see Table A1), and therefore, splitting the sample based on this metric does not lead to endogeneity concerns. The estimates in columns (2-3) and (5-6) of Table 4 provide strong support for Lazear (2001)'s hypothesis: in schools with less underlying disruption, there are no obvious effects of class size on disruption. In schools with more underlying disruption, the

[27] The wording of the questions were (original question in Spanish): (i) For teachers: *To what extent is each of the following situations a problem in your school? Disruption of order in class*, where teachers can answer one of the following: Not a problem, Slightly a problem, Moderately a problem or Seriously a problem. (ii) For students: *In my class there is a good work environment*, where students can answer one of the following: Strongly Disagree, Disagree, Agree, or Strongly Agree.

[28] The school principal specific question reads: *To what extent is each of the following aspects a problem in your school?* There are nine items: arriving late to the school, truancy (unjustified absences), classroom disorder, vandalism and theft, inappropriate or vulgar language, bullying or insults among students or other types of harassment, physical aggression between students, disrespect towards teachers or school staff, and bullying or insults towards teachers or school staff. The school principal can answer one of the following: Not a problem, Slight problem, Moderate problem, or Serious problem.

effects of class size on disruption are much larger. This hypothesis is supported by responses from both the teacher and student surveys. Results are robust to alternative specifications (see Table B1).

To benchmark the magnitude of class size effects on disruption, we conduct a back-of-the-envelope calculation of the implied impact on test scores. Estimating the magnitude of our disruption effects requires an estimate of the effect of disruption on test scores. To this end, we draw on prior literature which uses cohort-to-cohort variation in the share of girls in a grade to estimate the effects of this share on test scores, and argues that most of these effects operate through classroom disruption (Hoxby, 2000; Lavy and Schlosser, 2011). Under the stronger assumption that all of the effects operate through disruption, this means that share girls is a valid instrument for disruption in a regression of test scores on disruption (that also controls for school fixed effects as well as the baseline controls from equations 6 and 5).

Table D7 shows the "first stage" estimates. They suggest that a 10pp increase in the share of girls reduces disruption by around 1.7-1.8pp. In other words a 5-student reduction in class size in our dataset has the same effect on disruption as a 15-21pp increase in the share of girls in a class. The "reduced form" of that procedure, i.e. directly regressing the share of girls in a class on test scores, suggests that a 10pp increase in the share girls increases test scores by 0.0127SD.[29] Taking the "second stage" of that procedure, we find that a 1pp reduction in disruption increases test scores by 0.007SD. Although this likely an upper bound on the effect of disruption on test scores (because the assumed exclusion restriction is strong - girls could improve test scores via other channels such as working hard in class), it sheds interesting light on our estimates of the effects of disruption on test scores. In particular, it suggests that we would expect the reduction in disruption from a 5 student-decrease in class size to yield an increase in test scores of 0.0175-0.028SD. This estimate exceeds our own (0.0125SD), but the difference is modest and could plausibly be offset by the behavioral responses we examine below.

6.2 Individualized Instruction

Another key mechanism thought to mediate the effects of smaller classes is instructional practices. Instructional practices comprise a range of approaches, but two dimensions of these practices are thought to be promoted by smaller classes. First, smaller classes are often said to promote the use of small-group teaching, because teachers are better able to monitor a smaller number of groups (Ehrenberg et al., 2001). Second, smaller classes are thought to promote more individualized instruction, as teachers get to know students better and can

[29]This is very similar to the Hoxby (2000) estimate but lower than the Lavy and Schlosser (2011) estimate.

tailor instruction to their specific needs.

We use teacher and student surveys to examine the effects of class size on individualized instruction. From the teacher survey, we rely on three questions asking how frequently certain classroom activities occur during the academic year. We proxy individualized instruction with the following items: (i) "I prepare different tasks for students with learning difficulties and/or for those who progress more quickly"; (ii) "I check my students' exercise books or homework"; and (iii) "Students work in small groups to find a joint solution to a problem or task." From the grade 6 student questionnaire, we use two items. The first asks how students typically work with their teachers in class, including the statement "We work in small groups". The second asks whether students agree with various statements about their teachers, including "They let me demonstrate what I have learned".[30]

Table 5 reports 2SLS estimates of the effects of class size on individualized instruction from both teacher and student perspectives. From the student perspective, we find no effect of class size on the use of small-group teaching or on whether teachers allow students to demonstrate what they have learned. From the teacher perspective, we find no evidence that class size affects whether teachers prepare different tasks for students of different abilities. We emphasize that this is only one indicator, as teachers may individualize instruction in other ways (e.g., by providing additional attention and feedback to struggling students). Finally, we find that a five-student reduction in class size decreases the probability that teachers check students' exercise books or homework by 1.1 percentage points, relative to a base of about 93%. Our estimates are consistent with smaller classes promoting small-group teaching. For example, they suggest that a 5-student reduction in class size would increase the use of small-group teaching by 4pp on a base of 70%.

Overall, we find little evidence that class size has a meaningful impact on individualized instruction. Results are robust to alternative specifications (see Table B2). An important caveat for the two statistically significant outcome measures is that baseline means are already high. Teachers in grades three and six report checking homework more than 90% of the time and using small-group teaching about 70% of the time. Thus, a five-student reduction in class size would raise these frequencies by only 1.2 and 5.7 percent with respect to the baseline mean, respectively. Moreover, the effect on small-group teaching is not corroborated by the

[30]The teacher specific question reads: *How often does each of the following situations occur in the classroom throughout the academic year?* The teacher can answer one of the following: Never or almost never, Occasionally, Frequently, or In all or almost all. The first student specific question reads: *In general, how do you work in class with your teachers?* The student can answer one of the following: Never or almost never, Sometimes Almost always, or Always. The second student specific question reads: *In general, to what extent do you agree with these statements about your teachers?* The student can answer one of the following: Strongly Disagree, Disagree, Agree, or Strongly Agree.

student perspective, which reinforces the conclusion that class size has little effects on this instructional practice.

6.3 Class Size Effects on Individualized Instruction Settings

Our finding that class size has little effect on individualized instruction mitigates potential "bad control" concerns and allows us to test the hypothesis that class size reductions improve learning only when teachers use certain instructional practices (Ehrenberg et al., 2001; Blatchford, 2012).[31] We find little support for this hypothesis. Table 6 reports the class size effects on student test scores when splitting the sample by low versus high frequency of the different individualized instructional practices reported by teachers. To construct the high- versus low-frequency dummy, we first sum the number of math, Spanish, and English teachers who report high frequency use of a given instructional practice. We then define the dummy as equal to one if this sum is above the median and zero otherwise. Interacting class size with differentiated instruction, homework checking, or small-group work yields no significant effect on student test scores. Results are similar when using interactions rather than split samples, and remain non-significant and negligible when the analysis is implemented for the different individualized instructional practices reported by students.

7 Student, Family and Teacher Responses

Another reason why smaller classes may not lead to larger test score improvements is that they may induce compensating behavioral responses from teachers, students and their families. For example, students assigned to large classes may respond by completing more homework; their families may become more involved in their education. If students and families respond in this way, then our estimates of the effects of smaller classes on test scores should be interpreted as the total effect of smaller classes and these responses. If these effects offset one another, our estimates will understate the effect of smaller classes holding these responses

[31] Since we find only small effects of class size on the use of small-group practices, a natural concern is that these estimates may be biased by sample selection. Our preliminary analysis (to be added) suggests that, because the effects of class size on these variables are very small (or non-existent), any such bias is likely to be trivial. We also investigated an alternative strategy to test the effectiveness of individualized instruction as a mediator of class size effects on test scores. Teachers in Spain have access to teacher training, and in particular to training in individualized instruction. We show that this training is predictive of adoption of individualized instruction in the classroom, and is not predicted by our class size rule. Using an IV, we show that class size does not have larger effects on test scores for teachers that adopt individualized instructional practices. These results are available upon request.

fixed.[32] As noted by Leuven and Oosterbeek (2018), these responses can operate in either direction, with class size potentially reinforcing or partially offsetting the effect. For example, families whose children are assigned to larger classes may respond by reducing educational support, such as checking homework or providing other assistance.

We assess family responses to smaller classes in three ways. First, we estimate the effect of smaller classes on the amount of time students spend completing homework. Both the student and parent surveys ask about homework, allowing us to construct multiple measures of homework completion time in grade six (recall that students are not surveyed in grade three).[33] Both measures yield similar means: parents report that their children spend an average of 8.1 hours per week on homework, while students report 7.6 hours. Columns (3) and (4) of Table 7 show that students in smaller classes spend less time on homework. Looking at the parent-based measure, the estimates suggest that a 5-student reduction in class size would reduce homework time by about 0.14 hours per week in grade three and six. The student measure yields similar estimates but not statistically significant effects on homework time in grade six.

As a second measure of family responses, we create a parental involvement index using questions relating to the frequency with which parents plan for homework, help with homework, check that homework has been completed, and discuss the work that students have done in class. As seen in Column (2) of Table 7, smaller classes are associated with less parental involvement. For example, across both grades, a 5-student reduction in class size is estimated to lead to a 0.022SD reduction in involvement.

As a third measure of family responses, we estimate the effect of class size on whether a student has a private tutor. Around 14.5% of grade-six students have a private tutor or attend an academy. The estimates in Column (5) of Table 7 suggest that students in smaller classes are less likely to have a private tutor or attend an academy. Specifically, these estimates suggest that a 5-student reduction in class size reduces the probability that grade-six students have a private tutor by about 1.1 pp.

We analyze teacher responses to smaller classes and use student satisfaction with their teachers as a proxy for whether lessons become more enjoyable. Column (1) of Table 7 shows no evidence that students in smaller classes report greater well-being with their teachers. We also analyze student responses to two additional effort measures beyond time spent on homework. A supplementary student survey, administered at the end of each subject test

[32]Note that while we refer to these as student and family responses, they may also reflect teacher responses (e.g., students completing more homework because teachers assign more). Also, teachers could respond in ways that we do not capture (e.g., by "teaching to the test" in larger classes).

[33]The two measures are positively correlated at grade six (correlation coefficient 0.2).

in 2018 and 2019, asked students about the effort they exerted both before and during the test.[34] Columns (6) and (7) of Table 7 show that smaller classes lead students to reduce effort. For instance, a 5-student reduction in class size reduces the effort prior to the test by about 0.085SD, and effort during the test by 0.019SD. Results are robust to alternative specifications (see Table B3 and Table B4).

In summary, across all measures we find consistent evidence that students and their families respond to smaller classes. The magnitude of these responses is relatively small, suggesting they are unlikely to account for the class size effects on test scores. To benchmark the potential offsetting role of these responses, we include each measure in our main regression. Table D8 reports the class size effects when controlling for behavioral responses. The impact of class size on student test scores remains statistically insignificant when controlling for all these responses individually. Overall, the evidence suggests that the negligible impact of class size on student test scores is not driven by behavioral responses, which are themselves of relatively small magnitude.

8 Discussion and Conclusion

Many studies have estimated the effects of smaller classes on student outcomes. Few, if any, have estimated the effects of smaller classes on the classroom mechanisms said to improve these outcomes (e.g., reduced disruption, increased personalization and more-appropriate teaching methods). Using large samples of linked administrative and survey data from the region of Madrid, and exploiting a maximum class size rule that generates exogenous variation in primary school class size, we begin to fill this gap. In particular, we estimate the effects of smaller classes on these key dimensions of the learning environment. We find that class size reductions do activate these mechanisms. In particular, we find that they reduce levels of student disruption and increase the use of small-group teaching. However, we find that the associated magnitudes are small. On other words, the estimated effects on the classroom learning environment would not lead us to expect changes in class size to have large impact on behavioral outcomes.

A key question is why the classroom learning environment appears largely invariant to class size. In future drafts of the paper we plan to address this question. To that end we will focus on teachers and evaluate several explanations for why they might not change

[34] The two questions read as: (i) *How much did you study for this test?*. Responses to this question on study prior to the test take four categorical values, ranging from "not at all" (lowest level), "a little," "somewhat," to "a considerable degree" (highest level). (ii) *On a scale from 0 (lowest) - 10 (highest): How hard did you try on the test?*.

classroom practices in response to changes in class size. These explanations will build on the conceptual framework presented earlier in the paper and center on the role played by teacher training, teacher incentives, and teacher beliefs about the returns to changing their practices. We will then take these hypotheses to the data, constructing tests based on variation in teacher experience and the incentives provided by local school competition. In the meantime, it seems safe to conclude that our findings ought to focus on attention on exactly how smaller classes are expected to change the classroom learning environment, and whether policies (e.g., teacher training) could help to promote these changes and secure the benefits of smaller classes for student outcomes.

References

Abdulkadiroğlu, A. and Sönmez, T. (2003). School choice: A mechanism design approach. American Economic Review, 93(3):729–747.

Adusumilli, K., Agostinelli, F., and Borghesan, E. (2024). Heterogeneity and endogenous compliance: Implications for scaling class size interventions. Technical report, National Bureau of Economic Research.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. The Quarterly Journal of Economics, 132(2):871–919.

Angrist, J. D., Lavy, V., Leder-Luis, J., and Shany, A. (2019). Maimonides rule redux. American Economic Review: Insights, 1(3):309–324.

Angrist, J. D. and Lavy, V. C. (1999). Using maimonides rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics, 114(2):533–575.

Betts, J. R. and Shkolnik, J. L. (1999). The behavioral effects of variations in class size: The case of math teachers. Educational Evaluation and Policy Analysis, 21(2):193–213.

Blatchford, P. (2012). Three generations of research on class-size effects.

Blatchford, P. and Mortimore, P. (1994). The issue of class size for young children in schools: what can we learn from research? Oxford review of education, 20(4):411–428.

Bonesrønning, H., Finseraas, H., Hardoy, I., Iversen, J. M. V., Nyhus, O. H., Opheim, V., Salvanes, K. V., Sandsør, A. M. J., and Schøne, P. (2022). Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment. Journal of Public Economics, 216:104765.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. Econometrica, 83(6):2453–2483.

Chingos, M. and Meltzer, A. (2023). Class size reductions may be inequitably distributed under a new mandate in new york city.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., and Willms, J. D. (2001). Class size and student achievement. Psychological science in the public interest, 2(1):1–30.

Evertson, C. M. and Randolph, C. H. (1989). Teaching practices and class size: A new look at an old issue. Peabody Journal of Education, 67(1):85–105.

Fredriksson, P., Öckert, B., and Oosterbeek, H. (2013). Long-term effects of class size. The Quarterly journal of economics, 128(1):249–285.

Fredriksson, P., Öckert, B., and Oosterbeek, H. (2016). Parental responses to public investments in children: Evidence from a maximum class size rule. Journal of Human Resources, 51(4):832–868.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In Handbook of economic field experiments, volume 2, pages 95–322. Elsevier.

Gilraine, M. (2020). A method for disentangling multiple treatments from a regression discontinuity design. Journal of Labor Economics, 38(4):1267–1311.

Guarda, T. and Subraminian, S. (2023). How would the new limits to class sizes affect new york city schools? executive summary. New York: Independent Budget Office.

Hoxby, C. M. (2000). Peer effects in the classroom: Learning from gender and race variation.

Krueger, A. B. and Whitmore, D. (2001). Would smaller classes help close the black-white achievement gap?

Lavy, V. and Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. American Economic Journal: Applied Economics, 3(2):1–33.

Lazear, E. P. (2001). Educational production. The Quarterly Journal of Economics, 116(3):777–803.

Leuven, E. and Løkken, S. A. (2020). Long-term impacts of class size in compulsory school. Journal of Human Resources, 55(1):309–348.

Leuven, E. and Oosterbeek, H. (2018). Class size and student outcomes in europe. Analytischer Bericht, EENEE, page 33.

Neilson, C. (2024). The rise of coordinated choice and assignment systems in education markets around the world. Background paper to the World Development Report. 2024.

Opatrny, M., Havránek, T., Havránková, Z., and Scasny, M. (2025). Publication bias and model uncertainty in measuring the effect of class size on achievement. Accepted at Journal of Labor Economics.

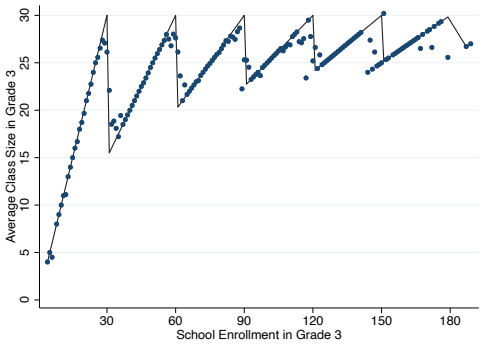
Rockoff, J. (2009). Field experiments in class size from the early twentieth century. Journal of Economic Perspectives, 23(4):211–230.

Schanzenbach, D. W. (2020). The economics of class size. In The economics of education, pages 321–331. Elsevier.

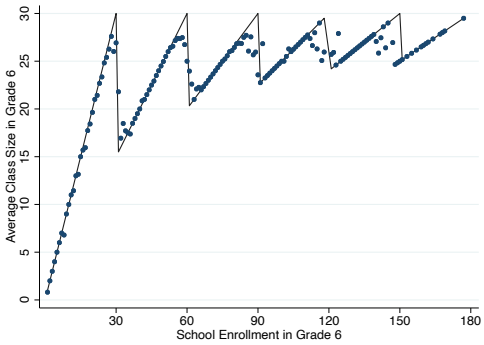
Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. American Economic Review, 99(1):179–215.

Figure 1: Maximum class size rules: Class Size in 2016-2019 for Grades 3 and 6 in the Region of Madrid.

(a) Grade 3



(b) Grade 6



Note: This figure plots unweighted average class size conditional on school enrollment in a sample containing averages for each school and year. The underlying data include an average of 1,200 schools each year between 2016-2019.

Table 1: Descriptive statistics: Schools, students and teachers

	Grade 3	**Grade 6**	**Grades 3 and 6**
Panel A: School Characteristics			
Average Class Size (records)	24.61	23.77	24.20
	(3.418)	(3.866)	(3.671)
Total number of classes (records)	2.235	2.185	2.210
	(1.029)	(0.988)	(1.009)
Number of enrolled students	55.57	52.59	54.09
	(27.83)	(26.67)	(27.30)
Number of classes	4677	4595	9272
Number of schools	1223	1198	1224
Panel B: Individual Characteristics			
Female	0.478	0.482	0.480
	(0.500)	(0.500)	(0.500)
Student born in Spain	0.880	0.890	0.885
	(0.325)	(0.312)	(0.319)
Student's father born in Spain	0.769	0.791	0.779
	(0.421)	(0.406)	(0.415)
Father years of education	13.65	13.51	13.58
	(3.699)	(3.684)	(3.693)
Mother years of education	14.08	13.88	13.99
	(3.536)	(3.512)	(3.526)
Father unemployed	0.101	0.101	0.101
	(0.302)	(0.301)	(0.301)
Mother unemployed	0.147	0.145	0.146
	(0.355)	(0.352)	(0.353)
Mother occupation: Administrative staff	0.236	0.247	0.241
	(0.425)	(0.431)	(0.428)
Father occupation: Administrative staff	0.0665	0.0698	0.0680
	(0.249)	(0.255)	(0.252)
Number of students	238009	221372	459381
Panel C: Teacher Characteristics			
Female teacher	0.783	0.694	0.735
	(0.412)	(0.461)	(0.442)
Teacher age	41.67	43.36	42.59
	(10.01)	(10.16)	(10.13)
Teacher experience (years)	14.53	16.42	15.56
	(9.758)	(10.42)	(10.17)
Full-time employment	0.964	0.966	0.965
	(0.187)	(0.181)	(0.184)
Civil servant	0.638	0.641	0.640
	(0.481)	(0.480)	(0.480)
Number of teachers	11563	13912	25475

Notes: Standard deviations in parentheses. Summary statistics on teachers, schools, classes and students from 2016-2019

Table 2: First-stage Estimates

	Grade 3			Grade 6			Grade 3 and 6		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Class size rule (instrument)	0.778***	0.766***	0.766***	0.798***	0.783***	0.785***	0.794***	0.780***	0.781***
	(0.021)	(0.022)	(0.022)	(0.020)	(0.020)	(0.021)	(0.016)	(0.016)	(0.016)
Enrollment	0.002	0.022***		0.004**	0.026***		0.003**	0.024***	
	(0.002)	(0.005)		(0.002)	(0.006)		(0.001)	(0.004)	
Enrollment squared/100		-0.011***			-0.013***			-0.013***	
		(0.003)			(0.003)			(0.003)	
Piecewise linear trend			0.011***			0.016***			0.013***
			(0.004)			(0.004)			(0.003)
R^2	0.59	0.60	0.59	0.63	0.63	0.63	0.61	0.62	0.62
F-test	241.26	238.55	236.86	364.41	336.34	366.42	409.06	392.64	405.15
Observations	237,720	237,720	237,347	221,089	221,089	221,089	458,809	458,809	458,436

Notes: The table shows the first-stage estimates of instrumenting actual class size by the maximum class size rule. Maximum class size rule is computed using administrative enrollment data. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Class Size Effects on Average Test Scores

	OLS	**IV/2SLS**		
	(1)	(2)	(3)	(4)
Panel A: Grade 3				
Class size	-0.0021	-0.0012	-0.0002	-0.0023
	(0.0033)	(0.0042)	(0.0041)	(0.0043)
Enrollment	X	X	X	
Enrollment squared/100	X	X		
Piecewise linear trend				X
Observations	237,720	237,720	237,720	237,347
Panel B: Grade 6				
Class size	-0.0053*	-0.0016	-0.0009	-0.0040
	(0.0028)	(0.0039)	(0.0037)	(0.0039)
Enrollment	X	X	X	
Enrollment squared/100	X	X		
Piecewise linear trend				X
Observations	221,089	221,089	221,089	221,089
Panel C: Grades 3 and 6				
Class size	-0.0047*	-0.0025	-0.0015	-0.0041
	(0.0025)	(0.0031)	(0.0030)	(0.0031)
Enrollment	X	X	X	
Enrollment squared/100	X	X		
Piecewise linear trend				X
Observations	458,809	458,809	458,809	458,436

Notes: Columns 1 shows results from the (endogenous) class size variable on test scores. Columns 2, 3 and 4 show the effect of class size when instrumented by the maximum class size rule. Average standardized test scores are computed as the sum of the standardized Language, English and Math test scores. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample. Clustered standard errors at the school level are shown in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 4: Class Size Effects on Disruption

	Teachers			Students		
	Average Effects	**Baseline School Disruption**		**Average Effects**	**Baseline School Disruption**	
		Below Median	Above Median		Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	0.0079***	0.0027	0.0156***	0.0050***	0.0038**	0.0066***
	(0.0027)	(0.0034)	(0.0041)	(0.0011)	(0.0015)	(0.0016)
Baseline Mean	0.404	0.352	0.468	0.228	0.222	0.235
Observations	24,851	13,702	11,149	168,115	93,295	74,820

Notes: Columns 1–3 report effects of class size on teacher-reported disruption. Columns 4–6 report effects of class size on student-reported disruption. Both these variables are transformed from a 4 point Likert scale into a binary variable. Baseline school disruption is measured as below/above median disruption level. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level when regressing the teacher-reported disruption variable. Student-reported disruption variables are not available for Grade 3 since students do not take the survey. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 5: Class Size Effects on Individualized Instruction

	Teachers			Students	
	Differentiated Instruction	**Checks Homework**	**Small Groups**	**Teachers allow us to Demonstrate Learning**	**Small Groups**
	(1)	(2)	(3)	(4)	(5)
Class size	-0.0004	-0.0023**	-0.0081***	-0.0002	-0.0032
	(0.0018)	(0.0011)	(0.0025)	(0.0008)	(0.0021)
Baseline Mean	0.826	0.932	0.698	0.824	0.350
Observations	24,650	24,611	24,652	169,358	169,485

Notes: Columns 1–3 report effects of class size on teacher-reported individualized instruction measures. Columns 4 and 5 report effects of class size on student-reported individualized instruction measures. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level when regressing the teacher-reported individualized instruction variable. Student-reported individualized instruction variables are not available for Grade 3 since students do not take the survey. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Class Size Effects in Individualized Instruction Settings

	Teachers						**Students**			
	Differentiated Instruction		**Checks Homework**		**Small Groups**		**Teachers allow us to Demonstrate Learning**		**Small Groups**	
	Low use (1)	**High use** (2)	**Low use** (3)	**High use** (4)	**Low use** (5)	**High use** (6)	**Low use** (7)	**High use** (8)	**Low use** (9)	**High use** (10)
Class size	0.0044	0.0001	-0.0103	0.0035	-0.0083	0.0065	-0.0016	-0.0023	-0.0039	-0.0005
	(0.0073)	(0.0063)	(0.0110)	(0.0055)	(0.0073)	(0.0062)	(0.0049)	(0.0039)	(0.0041)	(0.0047)
Baseline Mean	-0.066	-0.092	-0.121	-0.079	-0.140	-0.061	-0.045	-0.030	-0.009	-0.076
Observations	38,241	72,756	14,846	95,657	31,547	79,543	28,632	133,329	105,611	56,467

Notes: Columns 1–6 report effects of class size on teacher-reported individualized instruction measures. Columns 7 and 8 report effects of class size on student-reported individualized instruction measures. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level when regressing the teacher-reported individualized instruction variable. Student-reported individualized instruction variables are not available for Grade 3 since students do not take the survey. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 7: Class Size Effects on Teachers, Parents and Students Behavioral Responses

	Teachers	**Parents**		**Students**			
	Student's Wellbeing with Teachers	**Parental Involvement Anderson (2008) Index**	**Hours Homework per week**	**Hours Homework per week**	**Academy or Private Tutor**	**Self-reported Effort**	
						Before test (Anderson (2008) Index)	**During test (Std.)**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Class size	0.0001	0.0045**	0.0282*	0.0197	0.0022***	0.0169***	0.0038
	(0.0028)	(0.0019)	(0.0162)	(0.0161)	(0.0007)	(0.0052)	(0.0030)
Baseline Mean	-0.000	0.001	8.086	7.623	0.145	-0.000	-0.013
Observations	340,194	586,164	587,214	337,610	237,872	227,220	227,258

Notes: The table shows the behavioral response of teachers, parents and students to larger class sizes. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level when regressing the teacher-reported individualized instruction variable. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Appendix to:

Why Do Smaller Classes Fail to Improve Student Learning? Quasi-Experimental Evidence from Surveys of Teachers, Parents and Students

Damon Clark Pierre Deschamps José Montalbán

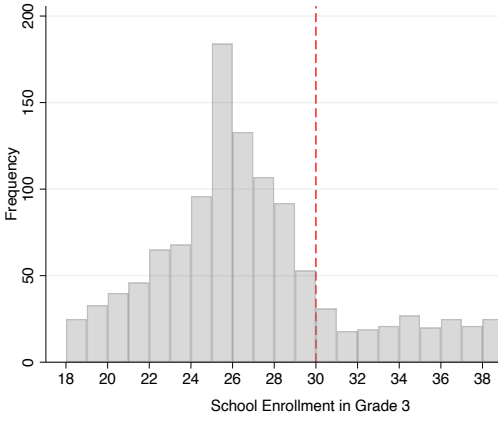
September 26, 2025

A Internal Validity

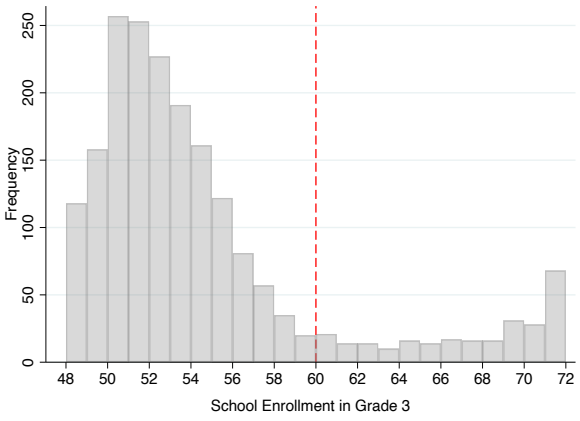
A.1 Manipulation

Figure A1: Distribution of Enrollment in Grade 3 at Cutoff 1,2,3,4,5 and 6 in 2016-2019

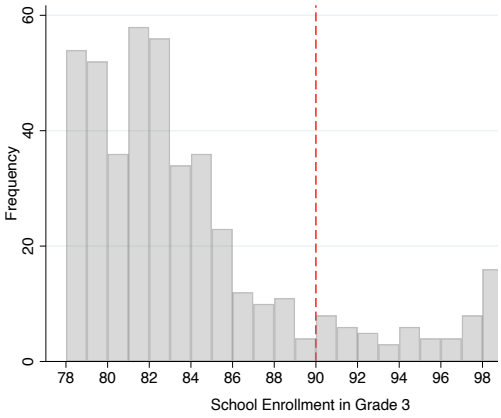
(a) Cutoff 1 (30 students)



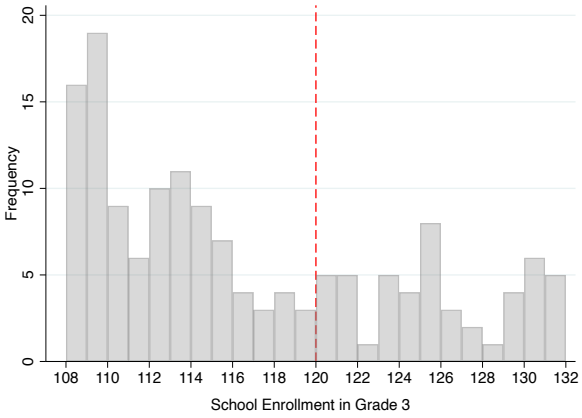
(b) Cutoff 2 (60 students)



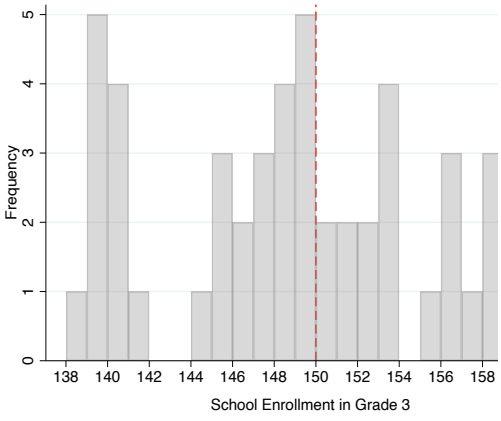
(c) Cutoff 3 (90 students)



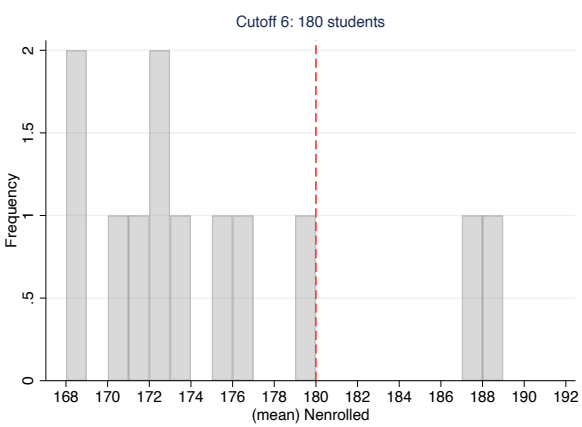
(d) Cutoff 4 (120 students)



(e) Cutoff 5 (150 students)



(f) Cutoff 6 (180 students)

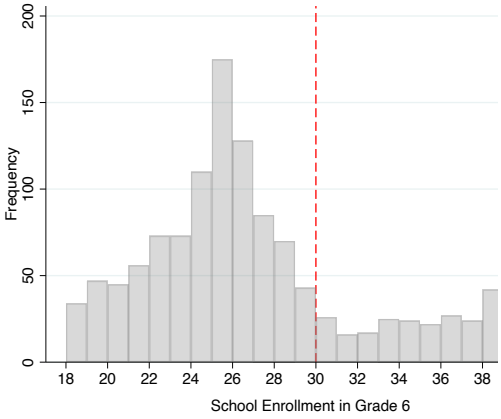


Note:The figures show histograms of school enrollment around different class size cutoff for students in Grade 6.

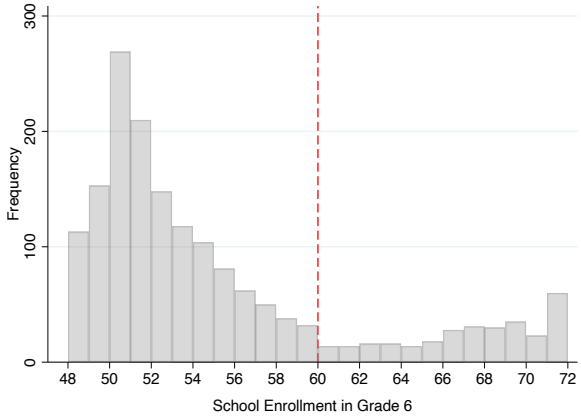
.

Figure A2: Distribution of Enrollment in Grade 6 at Cutoff 1,2,3,4,5 and 6 in 2016-2019

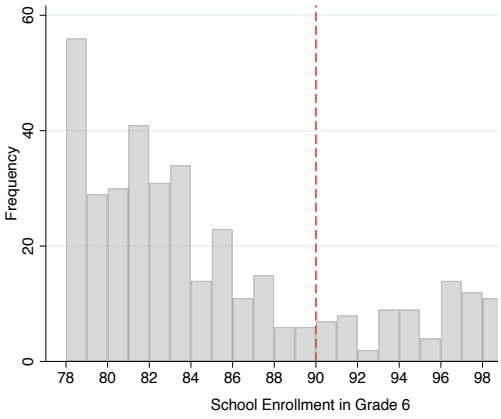
(a) Cutoff 1 (30 students)



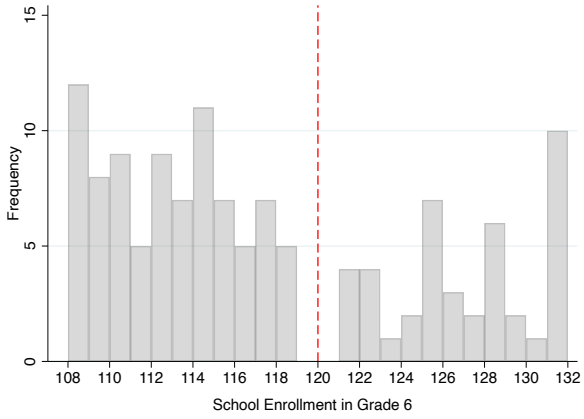
(b) Cutoff 2 (60 students)



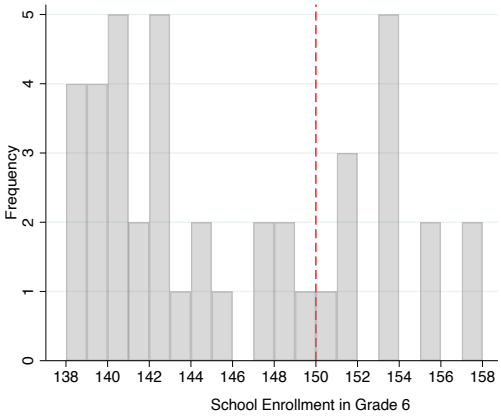
(c) Cutoff 3 (90 students)



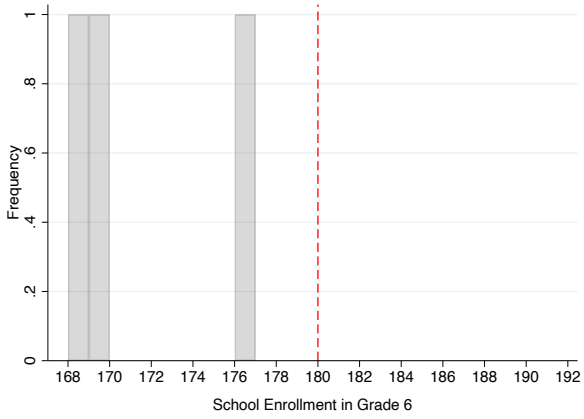
(d) Cutoff 4 (120 students)



(e) Cutoff 5 (150 students)



(f) Cutoff 6 (180 students)



Note:The figures show histograms of school enrollment around different class size cutoff for students in Grade 6.

Table A1: Maximum Class Size Rule and School and Student Covariate Balance.

	Grade 3		Grade 6		Grade 3 and 6	
	Baseline mean (1)	Class size rule (2)	Baseline mean (3)	Class size rule (4)	Baseline mean (5)	Class size rule (6)
Predicted Student Summary Index (Std.)	-0.000	-0.003	-0.000	0.001	-0.000	-0.001
		(0.004)		(0.004)		(0.003)
School SES (Std.)	0.249	-0.009	0.245	-0.011*	0.247	-0.010*
		(0.006)		(0.006)		(0.005)
School Underlying Disruption (Std.)	-0.157	0.003	-0.156	-0.004	-0.156	-0.001
		(0.007)		(0.006)		(0.005)
Female	0.478	0.001	0.482	0.001	0.480	0.001
		(0.001)		(0.001)		(0.001)
Spanish	0.880	0.001	0.890	0.001	0.885	0.001
		(0.001)		(0.001)		(0.001)
Father Spanish	0.769	-0.000	0.791	0.002	0.779	0.001
		(0.002)		(0.001)		(0.001)
Started daycare before 2 years-old	0.524	-0.004***	0.469	-0.003***	0.501	-0.004***
		(0.001)		(0.001)		(0.001)
N Digital Devices	5.576	0.002	6.480	0.006	5.992	0.006
		(0.005)		(0.006)		(0.005)
Less than 50 books at home	0.303	0.001	0.262	-0.001	0.284	0.000
		(0.001)		(0.001)		(0.001)
Mother Years Education	14.113	-0.001	13.889	-0.009	14.010	-0.006
		(0.012)		(0.011)		(0.010)
Father Years Education	13.661	-0.001	13.494	-0.010	13.585	-0.006
		(0.013)		(0.012)		(0.011)
Mother Fulltime	0.581	-0.000	0.582	-0.001	0.581	-0.001
		(0.001)		(0.001)		(0.001)
Mother Unemployed	0.143	-0.000	0.142	0.000	0.143	0.000
		(0.001)		(0.001)		(0.000)
Father Fulltime	0.745	0.000	0.727	-0.000	0.737	0.000
		(0.001)		(0.001)		(0.001)
Father Unemployed	0.088	-0.000	0.087	0.000	0.088	0.000
		(0.001)		(0.000)		(0.000)
Mother Occupation Elementary	0.108	-0.000	0.120	0.001	0.114	0.000
		(0.001)		(0.001)		(0.001)
Father Occupation Elementary	0.058	-0.000	0.057	0.000	0.057	-0.000
		(0.000)		(0.000)		(0.000)
Mother Occupation Admin	0.242	-0.000	0.249	0.001	0.245	0.000
		(0.001)		(0.001)		(0.001)
Father Occupation Admin	0.073	-0.000	0.076	0.000	0.074	-0.000
		(0.000)		(0.000)		(0.000)
Mother Occupation Tech	0.083	-0.000	0.077	-0.001**	0.080	-0.001**
		(0.000)		(0.000)		(0.000)
Father Occupation Tech	0.099	-0.000	0.095	-0.001	0.097	-0.000
		(0.000)		(0.000)		(0.000)
Observations		238,013		221,372		459,385

Notes: The table shows whether students and family characteristics are affected by instrumented class size. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A2: Maximum Class Size Rule and Teacher Covariate Balance.

	Grade 3		Grade 6		Grade 3 and 6	
	Baseline mean (1)	Class size rule (2)	Baseline mean (3)	Class size rule (4)	Baseline mean (5)	Class size rule (6)
Teacher Female	0.783	-0.003 (0.002)	0.694	0.003 (0.003)	0.735	0.000 (0.002)
Teacher Age	41.670	0.038 (0.059)	43.362	0.124** (0.054)	42.593	0.094** (0.041)
Teacher experience (years)	14.530	0.087 (0.059)	16.417	0.127** (0.056)	15.561	0.116*** (0.042)
Teacher experience at current school (years)	7.949	0.055 (0.047)	9.397	0.102** (0.048)	8.741	0.086** (0.036)
Full-time employment	0.964	0.000 (0.001)	0.966	-0.000 (0.001)	0.965	0.000 (0.001)
Civil servant, permanent position	0.429	0.002 (0.002)	0.461	-0.000 (0.002)	0.446	0.001 (0.002)
Civil servant trainee	0.011	0.000 (0.000)	0.010	0.001 (0.001)	0.011	0.001 (0.000)
Civil servant, temporary position	0.198	-0.002 (0.002)	0.170	-0.001 (0.002)	0.183	-0.001 (0.002)
External hire, permanent contract	0.317	-0.001 (0.001)	0.323	0.002 (0.001)	0.320	0.001 (0.001)
External hire, temporary contract	0.045	0.001 (0.001)	0.036	-0.001 (0.001)	0.040	-0.001 (0.001)
Observations		11,572		13,925		25,497

Notes: The table shows whether students and family characteristics are affected by instrumented class size. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample Teacher contract categories are mutually exclusive indicator variables. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A3: Maximum Class Size Rule and Teacher Covariate Balance for Grades 3 and 6.

	Grade 3		Grade 6		Grade 3 and 6	
	Baseline mean (1)	Class size rule (2)	Baseline mean (3)	Class size rule (4)	Baseline mean (5)	Class size rule (6)
Panel A: Math Teachers						
Teacher Female	0.769	-0.003	0.651	0.001	0.713	-0.001
		(0.003)		(0.004)		(0.002)
Teacher experience (years)	16.067	0.118	18.491	0.140*	17.212	0.142**
		(0.079)		(0.085)		(0.060)
Teacher experience at current school (years)	8.817	0.067	10.699	0.133*	9.706	0.108**
		(0.064)		(0.073)		(0.051)
Full-time employment	0.970	0.000	0.971	0.001	0.971	0.000
		(0.001)		(0.001)		(0.001)
Civil servant, permanent position	0.445	0.004	0.474	0.002	0.459	0.003
		(0.003)		(0.003)		(0.002)
Observations Math		3,947		3,539		7,486
Panel B: Spanish Teachers						
Teacher Female	0.778	-0.006*	0.700	0.003	0.741	-0.001
		(0.003)		(0.003)		(0.002)
Teacher experience (years)	15.728	0.102	17.763	0.112	16.701	0.108*
		(0.078)		(0.080)		(0.057)
Teacher experience at current school (years)	8.575	0.074	10.040	0.081	9.277	0.079
		(0.063)		(0.066)		(0.048)
Full-time employment	0.966	0.001	0.970	-0.000	0.968	0.000
		(0.002)		(0.001)		(0.001)
Civil servant, permanent position	0.437	-0.001	0.468	0.000	0.451	-0.000
		(0.003)		(0.003)		(0.002)
Observations Spanish		3,864		3,536		7,400
Panel C: English Teachers						
Teacher Female	0.803	-0.001	0.748	0.006*	0.777	0.002
		(0.003)		(0.004)		(0.002)
Teacher experience (years)	11.687	0.044	13.220	0.096	12.410	0.076
		(0.063)		(0.066)		(0.047)
Teacher experience at current school (years)	6.397	0.028	7.322	0.069	6.834	0.054
		(0.050)		(0.054)		(0.039)
Full-time employment	0.953	0.000	0.959	-0.001	0.956	-0.000
		(0.002)		(0.002)		(0.001)
Civil servant, permanent position	0.405	0.004	0.421	-0.000	0.413	0.001
		(0.003)		(0.003)		(0.002)
Observations English		3,761		3,355		7,116
Panel D: Science Teachers						
Teacher Female			0.679	0.001	0.679	0.001
				(0.004)		(0.004)
Teacher experience (years)			16.030	0.177**	16.030	0.177**
				(0.078)		(0.078)
Teacher experience at current school (years)			9.427	0.133**	9.427	0.133**
				(0.064)		(0.064)
Full-time employment			0.964	0.000	0.964	0.000
				(0.001)		(0.001)
Civil servant, permanent position			0.479	-0.003	0.479	-0.003
				(0.003)		(0.003)
Observations Science				3,495		3,495
Total Observations		11,572		13,925		25,497

Notes: This table shows teacher covariate balance across treatment and control groups by subject. Each coefficient comes from a separate regression on the variables listed in the rows. Maximum class size rule is computed using school enrollment. These regressions include year fixed-effects and school administrative districts fixed-effects. Science is only taught in Grade 6. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A4: Class Size Effects on Share of Repeaters

Dependent Variable	**Share of Repeaters**		
	Grade 3	**Grade 6**	**Grades 3 and 6**
	(1)	(2)	(3)
Class size rule (instrument)	0.0001	-0.0001	0.0000
	(0.0004)	(0.0005)	(0.0003)
Baseline Mean	0.034	0.030	0.032
Observations	7,180	6,926	14,106

Notes: This table shows whether the the share of repeaters in a class depends on the instrument. We keep only one observation per classroom-year Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A5: Class Size Effects on Standard Deviation of SES at the Class Level

Dependent Variable	**Within-class variance of SES**		
	Grade 3	**Grade 6**	**Grades 3 and 6**
	(1)	(2)	(3)
Class size rule (instrument)	-0.0011	-0.0018	-0.0013
	(0.0015)	(0.0015)	(0.0011)
Baseline Mean	0.802	0.809	0.805
Observations	7,633	7,157	14,790

Notes: This table shows whether the variance of student-level SES varies depending on class size. Student-level SES is computed by using the predicted test-score based on student and family pre-determined characteristics We keep only one observation per classroom-year Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

A.2 Non-response rate

Table A6: Maximum Class Size Rule and Non-response Student and Family Covariates.

	Grade 3		Grade 6		Grade 3 and 6	
	Baseline mean (1)	Class size rule (2)	Baseline mean (3)	Class size rule (4)	Baseline mean (5)	Class size rule (6)
Missing data on Female	0.000	0.000	0.000	0.000	0.000	0.000
		(0.000)		(0.000)		(0.000)
Missing data on Spanish	0.261	0.000	0.260	-0.000	0.261	-0.000
		(0.000)		(0.000)		(0.000)
Missing data on Father Spanish	0.486	-0.001	0.534	-0.001	0.509	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Started daycare before 2 years-old	0.319	0.000	0.525	-0.001	0.418	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on N Digital Devices	0.507	-0.000	0.547	-0.002*	0.526	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Less than 50 books at home	0.500	0.000	0.543	-0.002*	0.521	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Mother Years Education	0.508	-0.000	0.551	-0.002**	0.529	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Father Years Education	0.530	-0.001	0.570	-0.002**	0.549	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Mother Fulltime	0.524	0.000	0.568	-0.001	0.545	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Mother Unemployed	0.524	0.000	0.568	-0.001	0.545	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Father Fulltime	0.525	0.000	0.565	-0.001	0.544	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Father Unemployed	0.525	0.000	0.565	-0.001	0.544	-0.000
		(0.001)		(0.001)		(0.001)
Missing data on Mother Occupation Elementary	0.519	-0.000	0.560	-0.002*	0.539	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Father Occupation Elementary	0.538	-0.001	0.576	-0.002*	0.556	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Mother Occupation Admin	0.519	-0.000	0.560	-0.002*	0.539	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Father Occupation Admin	0.538	-0.001	0.576	-0.002*	0.556	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Mother Occupation Tech	0.519	-0.000	0.560	-0.002*	0.539	-0.001
		(0.001)		(0.001)		(0.001)
Missing data on Father Occupation Tech	0.538	-0.001	0.576	-0.002*	0.556	-0.001
		(0.001)		(0.001)		(0.001)
Observations		238,013		221,372		459,385

Notes: The table shows whether there is differential non-response depending on instrumented class size. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A7: Maximum Class Size Rule and Non-response Teacher Covariates.

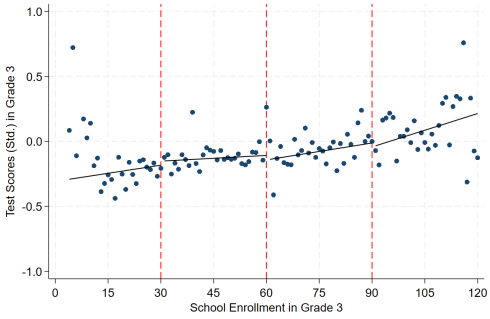
	Grade 3		Grade 6		Grade 3 and 6	
	Baseline mean (1)	Class size rule (2)	Baseline mean (3)	Class size rule (4)	Baseline mean (5)	Class size rule (6)
Missing: Female	0.0028	0.0001	0.0052	-0.0001	0.0041	-0.0000
		(0.0002)		(0.0004)		(0.0003)
Missing: Age	0.0054	-0.0001	0.0074	0.0000	0.0065	0.0000
		(0.0004)		(0.0005)		(0.0003)
Missing: Experience	0.0041	0.0004	0.0037	0.0002	0.0038	0.0003*
		(0.0003)		(0.0002)		(0.0002)
Missing: Experience at school	0.0074	0.0004	0.0045	0.0002	0.0058	0.0003
		(0.0004)		(0.0003)		(0.0002)
Missing: Full-time	0.0035	0.0003	0.0042	0.0001	0.0039	0.0002
		(0.0002)		(0.0002)		(0.0001)
Missing: Part-time	0.0035	0.0003	0.0042	0.0001	0.0039	0.0002
		(0.0002)		(0.0002)		(0.0001)
Missing: Permanent civil servant	0.0029	0.0003	0.0044	-0.0003	0.0037	-0.0000
		(0.0002)		(0.0003)		(0.0002)
Missing: Civil servant trainee	0.0029	0.0003	0.0044	-0.0003	0.0037	-0.0000
		(0.0002)		(0.0003)		(0.0002)
Missing: Temporary civil servant	0.0029	0.0003	0.0044	-0.0003	0.0037	-0.0000
		(0.0002)		(0.0003)		(0.0002)
Missing: External permanent	0.0029	0.0003	0.0044	-0.0003	0.0037	-0.0000
		(0.0002)		(0.0003)		(0.0002)
Missing: External temporary	0.0029	0.0003	0.0044	-0.0003	0.0037	-0.0000
		(0.0002)		(0.0003)		(0.0002)
Observations		11,572		13,925		25,497

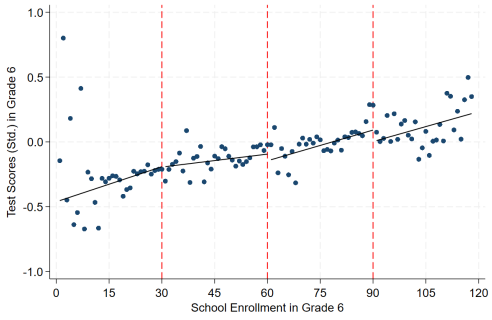
Notes: The table shows whether there is differential non-response for teacher surveys depending on instrumented class size. Each regression controls for year, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

B Alternative specifications

Figure B1: Average Test Scores by Enrollment in 2016-2019 for Grades 3 and 6.

(a) Grade 3





(b) Grade 6

Note: This figure plots the unweighted average standardized test scores on school enrollment. The y-axis variable is standardized test scores. The x-axis variable is school enrollment.

Table B1: Class Size Effects on Disruption- alternative specifications

	No squared enrollment control						Piecewise linear trend					
	Teachers			Students			Teachers			Students		
	Average Effects	Baseline School Disruption		Average Effects	Baseline School Disruption		Average Effects	Baseline School Disruption		Average Effects	Baseline School Disruption	
		Below Median	Above Median		Below Median	Above Median		Below Median	Above Median		Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Class size	0.0061**	0.0021	0.0125***	0.0046***	0.0037***	0.0063***	0.0081***	0.0020	0.0167***	0.0054***	0.0043***	0.0068***
	(0.0025)	(0.0031)	(0.0039)	(0.0010)	(0.0014)	(0.0015)	(0.0029)	(0.0038)	(0.0042)	(0.0011)	(0.0016)	(0.0016)
Baseline Mean	0.404	0.352	0.468	0.228	0.222	0.235	0.404	0.352	0.467	0.228	0.222	0.235
Observations	24,851	13,702	11,149	168,115	93,295	74,820	24,815	13,686	11,129	168,115	93,295	74,820

Notes: Columns 1–3 and 7–9 report effects of class size on teacher-reported disruption. The first six columns show specifications without the squared enrollment control. The last six columns show specifications with a piecewise linear trend control. Columns 4–6 and 10–12 report effects of class size on student-reported disruption. Each coefficient comes from a separate regression on the variables listed in the rows. Baseline school disruption is measured as below/above median disruption level. Each regression controls for year, district and school type (public or semi-public) fixed-effects We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B2: Class Size Effects on Individualized Instruction - alternative specifications

	No squared enrollment control					Piecewise linear trend				
	Teachers			Students		Teachers			Students	
	Differentiated Instruction	Check Homeworks	Small Groups	Teachers allow to Demonstrate Learning	Small Groups	Differentiated Instruction	Check Homeworks	Small Groups	Teachers allow to Demonstrate Learning	Small Groups
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Class size	-0.0013	-0.0024**	-0.0061**	-0.0000	-0.0020	-0.0005	-0.0021**	-0.0076***	0.0003	-0.0025
	(0.0016)	(0.0010)	(0.0024)	(0.0007)	(0.0021)	(0.0018)	(0.0011)	(0.0026)	(0.0008)	(0.0022)
Baseline Mean	0.826	0.932	0.698	0.824	0.350	0.826	0.933	0.697	0.824	0.350
Observations	24,650	24,611	24,652	169,358	169,485	24,614	24,575	24,616	169,358	169,485

Notes: Columns 1–3 and 6–8 report effects of class size on teacher-reported individualized instruction measures. Columns 4–5 and 9–10 report effects of class size on student-reported individualized instruction measures. The first five columns show specifications without the squared enrollment control. The last five columns show specifications with a piecewise linear trend control. Each regression controls for year, district and school type (public or semi-public) fixed-effects We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B3: Class Size Effects on Teachers, Parents and Students Behavioral Responses- no squared enrollment

	Teachers	Parents		Students			
	Student's Wellbeing with Teachers	Parental Involvement Anderson (2008) Index	Hours Homework per week	Hours Homework per week	Academy or Private Tutor	Self-reported Effort	
						Before test (Anderson (2008) Index)	During test (Std.)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Class size	-0.0002	0.0076***	0.0440***	0.0270*	0.0021***	0.0131***	0.0023
	(0.0027)	(0.0020)	(0.0160)	(0.0154)	(0.0007)	(0.0050)	(0.0029)
Baseline Mean	-0.000	0.001	8.086	7.623	0.145	-0.000	-0.013
Observations	170,097	293,082	293,607	168,805	118,936	113,610	113,629

Notes: This table shows the behavioral responses of teachers, parents and students to instrumented class size. The specifications in this table do not control for squared enrollment. Baseline school disruption is measured as below/above median disruption level. Each regression controls for year, district and school type (public or semi-public) fixed-effects We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

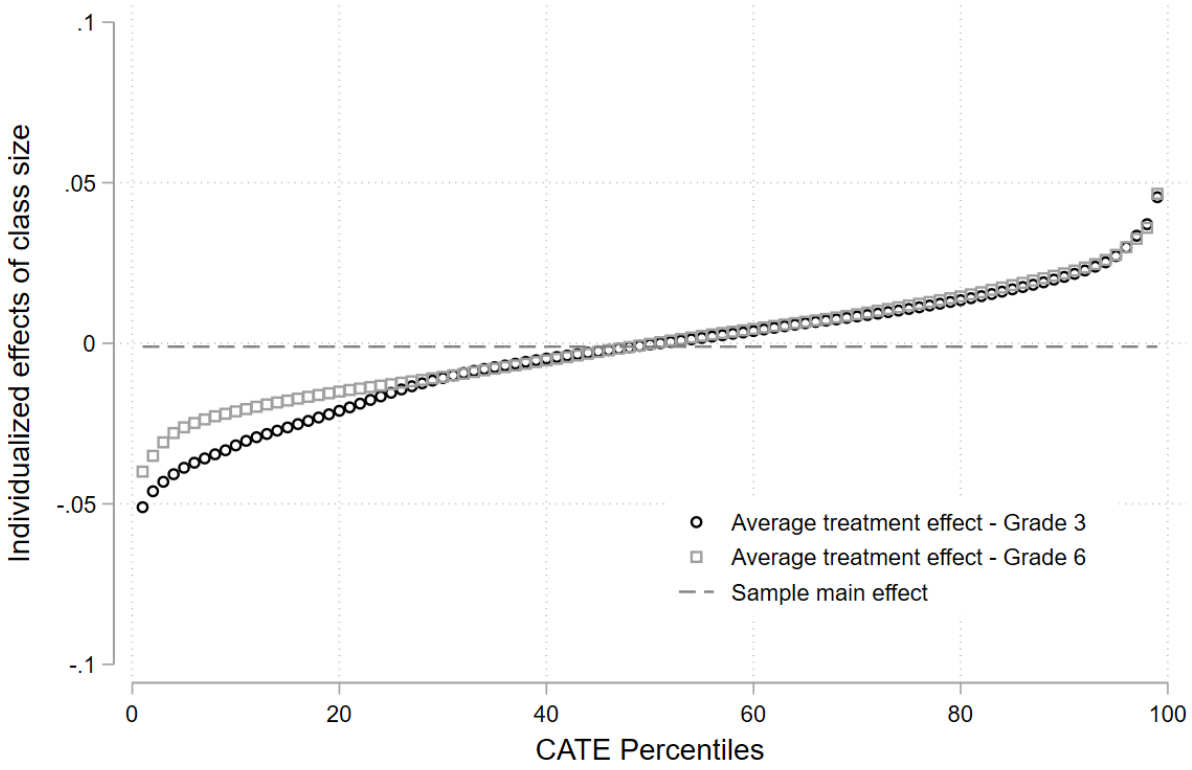
Table B4: Class Size Effects on Teachers, Parents and Students Behavioral Responses for Grades 3 and 6 - piecewise linear trend

	Teachers	Parents		Students			
	Student's Wellbeing with Teachers	Parental Involvement Anderson (2008) Index	Hours Homework per week	Hours Homework per week	Academy or Private Tutor	Self-reported Effort	
						Before test (Anderson (2008) Index)	During test (Std.)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Class size	0.0012	0.0068***	0.0408**	0.0235	0.0023***	0.0173***	0.0029
	(0.0028)	(0.0019)	(0.0165)	(0.0162)	(0.0007)	(0.0053)	(0.0030)
Baseline Mean	-0.000	0.001	8.089	7.623	0.145	-0.000	-0.013
Observations	170,097	292,819	293,354	168,805	118,936	113,610	113,629

Notes: This table shows the behavioral responses of teachers, parents and students to instrumented class size. The specifications in this table control for a piecewise linear trend instead of enrollment. Baseline school disruption is measured as below/above median disruption level. Each regression controls for year, district and school type (public or semi-public) fixed-effects We additionally control for grade level in the pooled sample Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

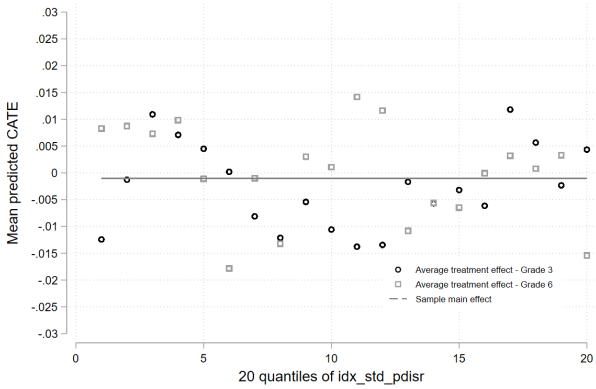
C Heterogeneous effects

Figure C2: Distribution of individual treatment effects, estimated through an instrumental forest

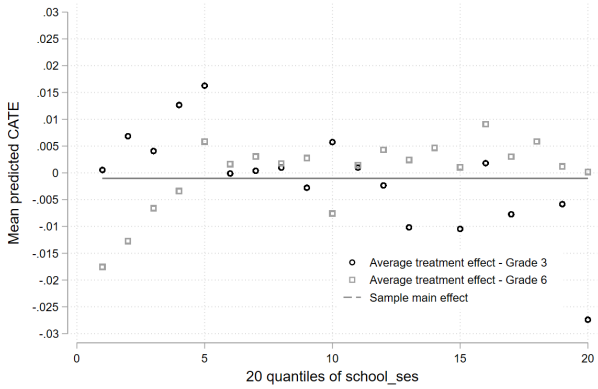


Note: This figure plots the distribution of individualized treatment effects of class size on test scores, etimated through an instrumental forest. We use the school level SES and disruption index, and information on student gender, number of digital devices at home, age at which they started preschool, and number of books at home, as well as data on parental education, employment status, nationality and occupation.

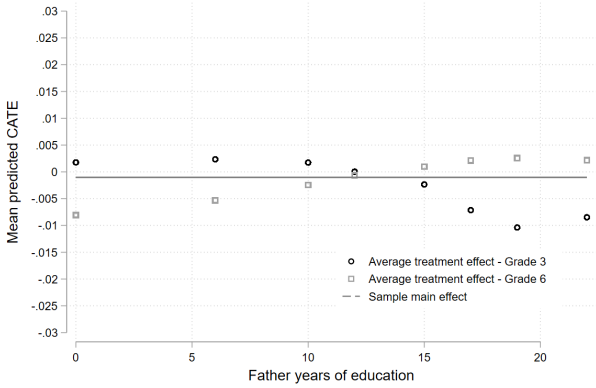
Figure C3: Treatment effect heterogeneity



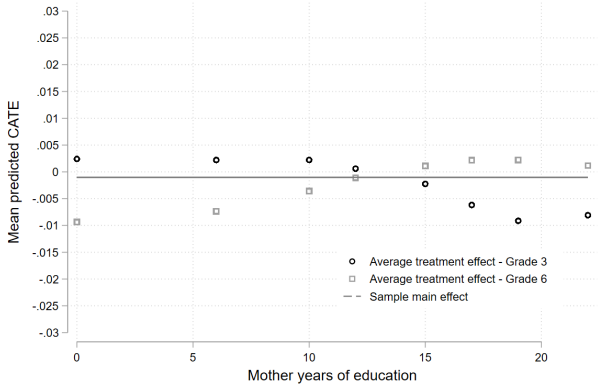
(a) Disruption ventiles



(b) SES ventiles



(c) Father education



(d) Father full-time

Note: This figure plots the distribution of individualized treatment effects of class size on test scores, estimated through an instrumental forest. We use the school level SES and disruption index, and information on student gender, number of digital devices at home, age at which they started preschool, and number of books at home, as well as data on parental education, employment status, nationality and occupation.

Table C5: Class Size Effects on Average Test Scores - Heterogeneity

Dependent Variable	**Test scores**										
	Students		**Subject**			**School SES**		**School Disruption**		**Teacher experience**	
	Male	**Female**	**Language**	**Math**	**Science**	**Below median**	**Above median**	**Below median**	**Above median**	**Below median**	**Above median**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Class size	-0.0021	-0.0031	0.0005	-0.0025	-0.0010	-0.0008	0.0018	-0.0011	-0.0035	-0.0005	0.0030
	(0.0032)	(0.0033)	(0.0030)	(0.0033)	(0.0037)	(0.0039)	(0.0043)	(0.0039)	(0.0048)	(0.0054)	(0.0052)
Baseline Mean	-0.129	0.021	-0.038	-0.043	-0.029	-0.332	0.126	0.045	-0.186	-0.091	-0.032
Observations	238,565	220,240	452,210	451,115	217,744	183,108	275,701	255,653	203,061	95,774	97,711

Notes: This table shows heterogeneity in the effect of instrumented class size on test score by student gender, subjects, school characteristics and teacher experience Grades are pooled together Each regression controls for year, grade level, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table C6: Class Size Effects on Disruption as Reported by Students - Heterogeneity

Dependent Variable	**Disruption - student measure**							
	Students		**School SES**		**School Disruption**		**Teacher experience**	
	Male	**Female**	**Below median**	**Above median**	**Below median**	**Above median**	**Below median**	**Above median**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class size	0.0058***	0.0040***	0.0054***	0.0045**	0.0038**	0.0066***	0.0061*	0.0070***
	(0.0012)	(0.0013)	(0.0014)	(0.0018)	(0.0015)	(0.0016)	(0.0033)	(0.0026)
Baseline Mean	0.242	0.213	0.236	0.222	0.222	0.235	0.240	0.226
Observations	87,063	81,052	70,373	97,675	93,295	74,820	20,542	24,292

Notes: This table shows heterogeneity in the effect of instrumented class size on disruption by student gender, subjects, school characteristics and teacher experience Grades are pooled together Each regression controls for year, grade level, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

D Benchmarking mechanisms

Table D7: Effect of Share of Boys in Class on Class Disruption and Test Scores

	Disruption		**Test Scores**
	Teachers	**Students**	
	(1)	(2)	(3)
Share of Boys in Class	0.185**	0.170***	-0.127***
	(0.082)	(0.030)	(0.043)
Baseline Mean	0.404	0.227	-0.060
Observations	24,607	166,795	453,029

Notes: This table shows how the share of boys in class affects disruption and test scores. Grades are pooled together, except for column 2 since students in Grade 3 did not answer this survey question. Each regression controls for year, grade level, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

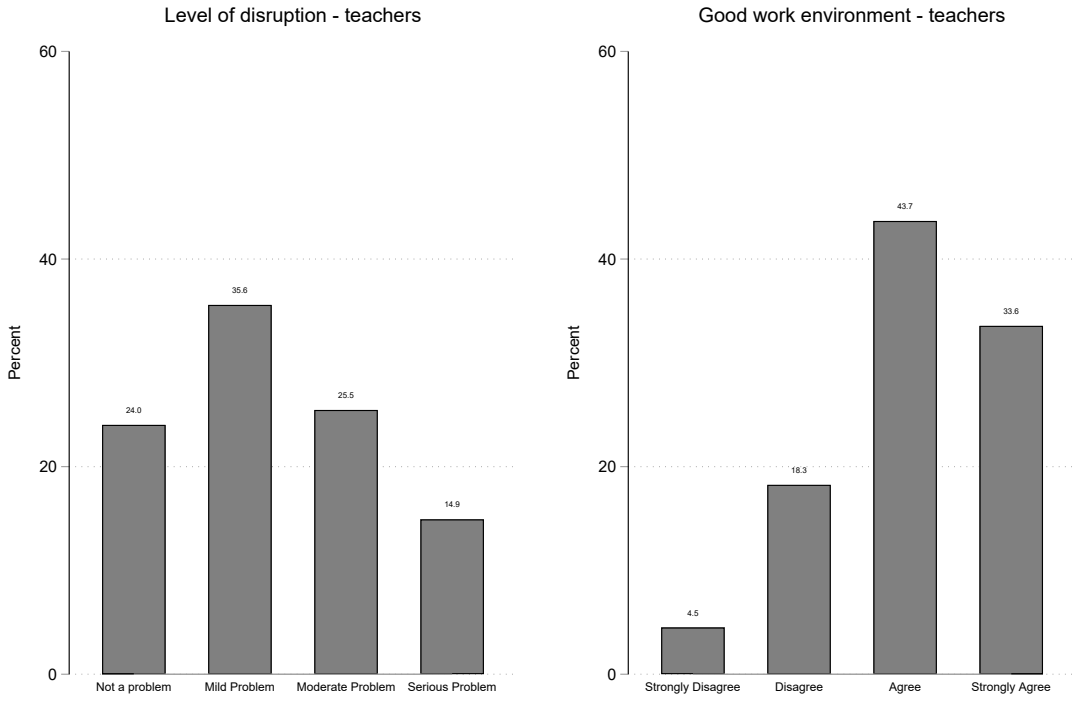
Table D8: Class size Effects when Controlling for Behavioral Responses

Dependent variable:	**Standardized test scores**					
	Family Responses (Grades 3 and 6)			**Student Responses (Grade 6)**		
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	-0.0011	-0.0007	-0.0010	-0.0023	-0.0042	-0.0082
	(0.0032)	(0.0032)	(0.0032)	(0.0038)	(0.0045)	(0.0052)
Anderson Indx. Parental involvement		-0.1049***				
		(0.0029)				
Hours of homework (parents)			-0.0039***			
			(0.0006)			
Hours of homework (students)				0.0085***		
				(0.0007)		
Studying with tutor					-0.3880***	
					(0.0076)	
Anderson Indx. Effort before test						-0.0132***
						(0.0051)
Baseline Mean	0.053	0.053	0.053	-0.032	-0.031	-0.052
Observations	210,444	210,444	210,444	161,316	112,265	110,515

Notes: This table shows whether behavioral responses of parents affect estimates of class size effects on test scores Grades are pooled together when possible, but students in Grade 3 did not answer survey questions in columns 4–7. Each regression controls for year, grade level, district and school type (public or semi-public) fixed-effects, as well as a second-order polynomial for enrollment. Clustered standard errors at the school level are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

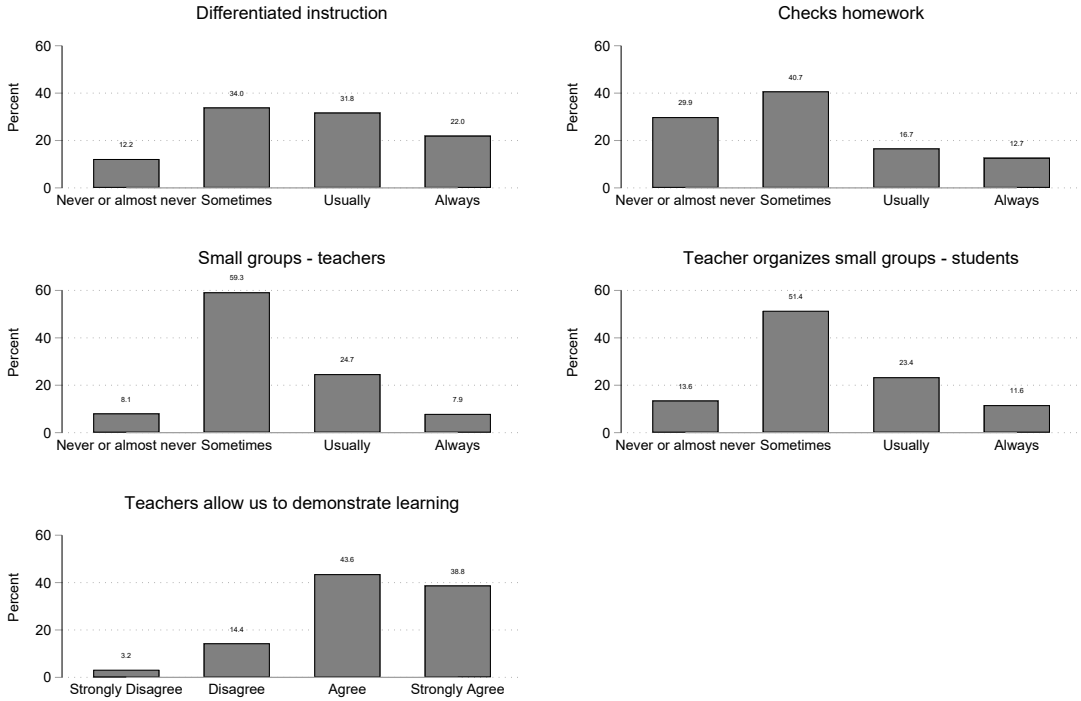
E Distribution of survey variables

Figure E4: Distribution of survey answers for questions on disruption



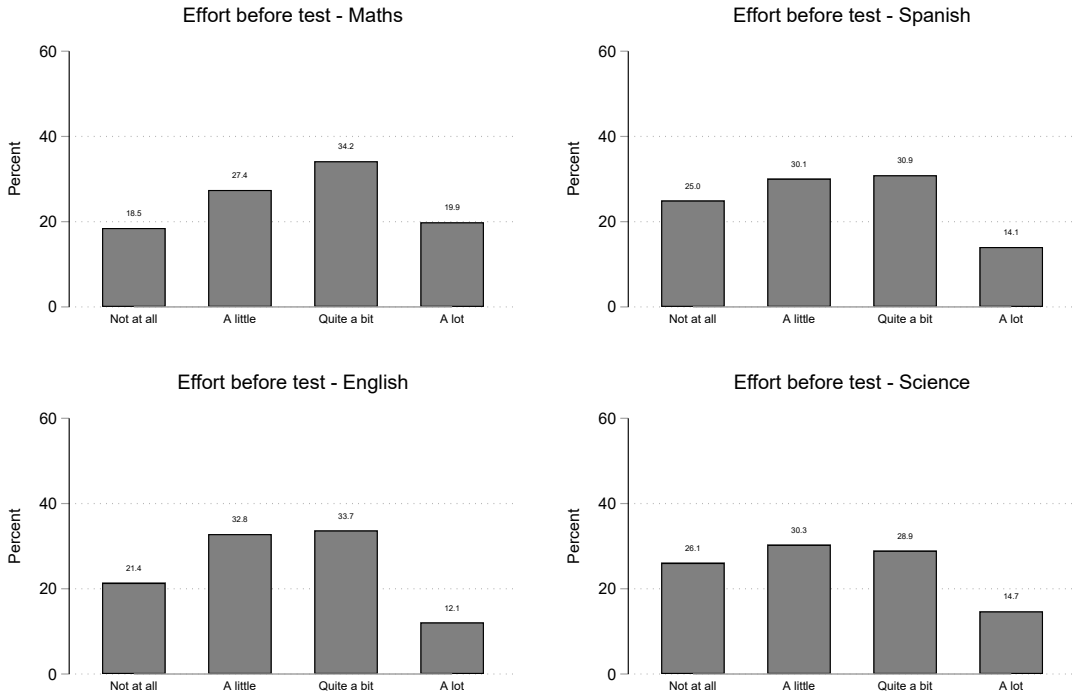
Note:Histogram of answers to survey .

Figure E5: Distribution of survey answers for questions on individualization



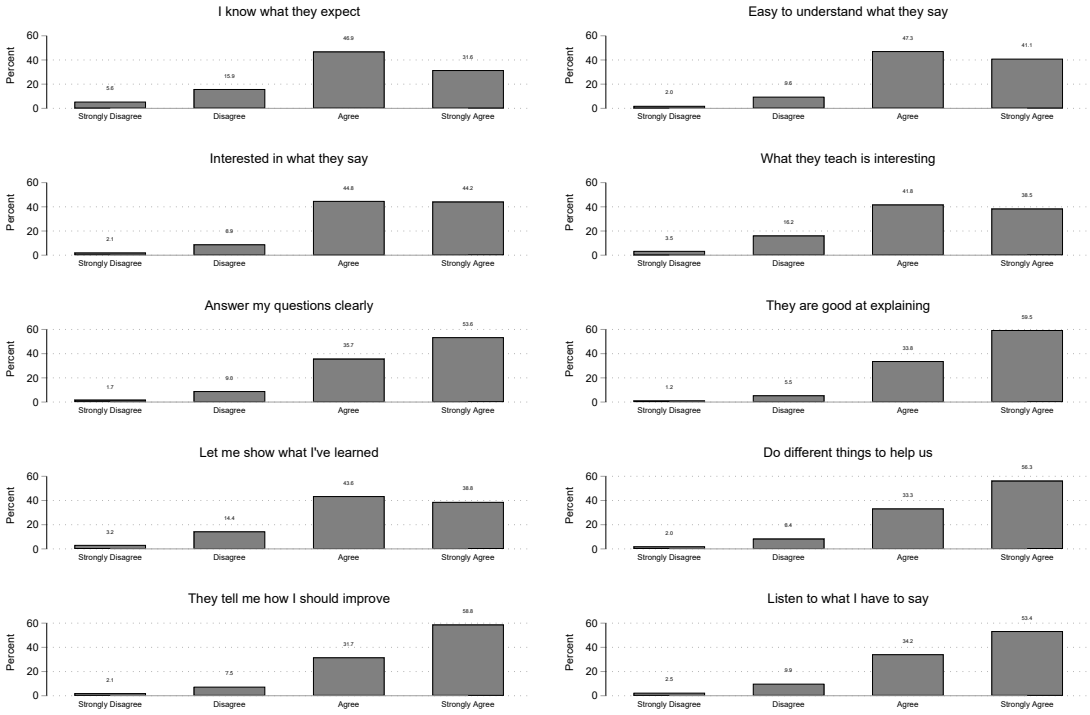
Note: .

Figure E6: Distribution of survey answers for questions on student effort



Note: .

Figure E7: Distribution of survey answers for questions on student satisfaction with teacher



Note: .