I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18025

How Do Classmates Matter for the Class-Size Effects?

Ryuichi Tanaka
Tong Wang

JULY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18025

How Do Classmates Matter for the Class-Size Effects?

Ryuichi Tanaka
University of Tokyo, RIETI and IZA

Tong Wang
Ritsumeikan University

JULY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

How Do Classmates Matter for the Class-Size Effects?*

This paper investigates the effect of class-size reduction on students' academic outcomes, with a particular emphasis on its heterogeneity based on classmates' characteristics. We estimate the causal effects of class-size reduction on students' mathematics and language test scores by controlling for student and teacher fixed effects. To address potential endogeneity, we employ the predicted class size with a cap as an instrumental variable for the actual class size. Utilizing rich panel data on Japanese primary school students, our findings indicate a positive and robust average effect of class-size reduction on mathematics test scores. Furthermore, we find that classes with high-ability classmates benefit even more from class-size reduction in terms of language test scores. The effect of class-size reduction on mathematics test scores is found to depend positively on the ability of the lowest-achieving student in a class. Additionally, classes with a higher proportion of female students tend to benefit more from class-size reduction. Our results lend support to the theoretical framework proposed by Lazear (2001).

JEL Classification: J13, J18, N35

Keywords: education, test scores, class-size reduction, ability, heterogeneity

Corresponding author:
Ryuichi Tanaka
Institute of Social Science
The University of Tokyo
7-chōme-3-1 Hongō
Bunkyo City
Tokyo 113-8654
Japan

E-mail: ryuichi.tanaka@iss.u-tokyo.ac.jp

* This study is conducted as a part of the project "Microeconometric Analysis of Education Policy" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The draft of this paper was presented at a DP seminar of the Research Institute of Economy, Trade and Industry (RIETI). The authors are grateful for valuable comments from Daiji Kawaguchi, Masayuki Morikawa, Shujiro Urata, and the participants at seminars and workshops. The authors acknowledge financial support for this project from the Japan Society for the Promotion of Science (Grant Numbers 20H05629 and 25H00538 for Ryuichi Tanaka.

1 Introduction

Class-size reduction stands as a prominent policy instrument aimed at enhancing the quality of educational environments. The underlying premise is that smaller classes benefit from richer educational resources per student, thereby fostering improved educational outcomes. However, the effectiveness of class-size reduction is difficult to be observed empirically to a large extent (e.g., Hanushek (1986), Hanushek (2003), Hanushek (2006)). While numerous studies have reported positive effects of smaller classes on academic achievement (e.g., Angrist and Lavy (1999); Krueger (1999), Krueger (2003)), the magnitudes of these *average* effects are often limited and sometimes statistically insignificant (e.g., Angrist et al. (2019)).

As evident in the existing literature, the reported size and significance of the effect of class-size reduction on students' academic achievement vary considerably. We suspect that these discrepancies may arise because class-size reduction has heterogeneous effects across different types of classes, such as those with varying student compositions or taught by different types of teachers. Investigating this potential heterogeneity in the effect of class-size reduction serves as the primary motivation for our study.

The characteristics of classmates are likely to influence the effectiveness of class-size reduction for at least two reasons. First, even in the absence of direct interaction among classmates, if the effect of class size varies according to individual student characteristics such as baseline ability and gender, the average effect will naturally be influenced by the composition of the student body. More importantly, within the context of schooling, where students interact with each other, the distribution of students' characteristics (e.g., gender and ability of classmates) can affect the impact of class-size reduction through various channels, including peer effects. In Lazear (2001)'s model, which conceptualizes schooling as a joint production process among students, disruptive behavior by even a single student can impede the entire educational process within the class, thereby negatively affecting the educational outcomes of all classmates. If the schooling process exhibits such strong complementarity among students, the composition of classmates will indeed generate heterogeneity in the effects of class size.

In this study, we examine the heterogeneous effect of class-size reduction on students' academic outcomes. As sources of this heterogeneity, we focus on the distribution of baseline academic outcomes, measured by test scores in the preceding grade (hereafter referred to as ability), and the gender of students. Drawing upon the spirit of Lazear (2001)'s model, the individual probability of students disturbing the education process is one of the fundamental determinants of educational outcomes. We employ student ability and gender as proxies for this probability. Generally, a negative correlation exists between students' ability and misbehavior in classrooms (e.g., Myers et al. (1987)). Regarding gender, boys may exhibit a higher propensity for disruptive behavior than girls, as evidenced by their greater likelihood of engaging in bullying and cyberbullying (Li (2006)) and their higher rates of referral, diagnosis, and treatment for Attention-Deficit Hyperactivity Disorder (ADHD) symptoms (e.g., Gaub and Carlson (1997); Gershon and Gershon (2002)). We estimate the heterogeneous effect of class-size reduction based on the average, maximum, and minimum ability of classmates, as well as the proportion of female students in a class.

Our data are taken from administrative records of primary school students in a large municipality within the Tokyo metropolis of Japan, collected between 2010 and 2016. This dataset comprises panel information on academic performance, as measured by a standardized test, linked with information on teachers and socioeconomic background, such as eligibility for school financial assistance (analogous to free lunch programs), for students in the second to sixth grades.

To identify the causal effects of class-size reduction on students' outcomes, it is essential to exploit either random variations in class size within a school (e.g., Krueger (1999)) or quasi-random variations arising from an institutional setting (e.g., Angrist and Lavy (1999)). In the absence of randomized controlled class formation in our data, we adopt the latter approach, leveraging quasi-experimental variations in class size. Japanese primary schools adhere to class-size regulations set by the Ministry of Education, the central authority for education policy in Japan, which cap class size at 40 students. Similar to the identification strategy employed by Angrist and Lavy (1999), we utilize the predicted class size based on grade size (the so-called Maimonides' rule) as an instrument for the actual class size. Furthermore, we control for teacher-student pair fixed effects to mitigate potential threats to the identification of the class-size reduction effect stemming from endogenous matching between students and teachers. In Japanese primary schools, it is common practice for class composition, as well as classroom teachers, to be shuffled and reassigned as students progress to the next grade. This generates variations in class size even when a student is taught by the same teacher across consecutive grades. By controlling for teacher-student fixed effects, we identify the causal effect of class-size reduction from changes in class size for students taught by the same teachers. This strategy, combined with the quasi-random variation in class size induced by the class-size cap, strengthens our identification strategy.

The results of our estimation analyses confirm that, on average, smaller classes are beneficial for students' academic performance, with a more pronounced effect observed for mathematics: a marginal reduction in class size is associated with an increase of 0.009 standard deviations in math test scores. We also find that students in classes with higher average baseline academic performance experience greater benefits from class-size reduction in Japanese language test scores. Moreover, the effect of class-size reduction is more sensitive to the ability of the lowest-performing student than to that of the highest-performing student. Classes where the lowest-performing student has a higher baseline score demonstrate significantly greater gains from class-size reduction in math test scores compared to classes with lower-performing bottom students. These findings align with the results of Lavy et al. (2012a) and Lavy et al. (2012b), who found that the proportion of low-performing students and/or the ability of the worst-performing student in a class significantly impact the performance of other students. Our findings are consistent with their results, and we extend their implications to the context of class-size reduction. Additionally, our analysis reveals that classes with a higher percentage of female students benefit more from class-size reduction.

The remainder of this paper is structured as follows. Section 2 provides a review of the relevant literature and outlines our contributions to the field. Section 3 details the institutional background. Section 4 our empirical strategy. Section 5 describes the data used in our analysis. Section 6 presents our empirical findings. Section 7 discusses the robustness and interpretation of our findings. Finally, Section 8 concludes the paper.

2 Literature Review and Our Contributions

Historically, a substantial body of literature examines the relationship between class size and student performance with observational data. A series of meta-analyses conducted by Hanushek (1986), Hanushek (2003), Hanushek (2006) found no consistent relationship between class size and student outcomes, showing the difficulty of identification of class size effects in the absence of experimental data or a quasi-experimental setting. With the experimental data from the Project STAR in Tennessee in the U.S., Krueger (1999) found a positive effect of small class size. While the mapping of percentile test scores to tangible outcomes remains unclear, the reported effect sizes, relative to the standard deviation of the average percentile score, were 0.20 in kindergarten, 0.28 in first grade, 0.22 in second grade, and 0.19 in third grade.

Angrist and Lavy (1999) is one of the first papers to estimate the class size effect with observational data within a quasi-experimental framework. They identified a negative relationship between class size and student academic performance, as measured by test scores in Israeli public primary schools. The institutional context of Israeli public primary schools, subject to a maximum class size regulation (the Maimonides' Rule), generates a discontinuous change in class size around multiples of the class size cap in grade size. Their work demonstrated the effectiveness of a regression discontinuity design (RDD) as an identification strategy for the causal effect of class size. Subsequently, numerous studies have applied RDD to observational data, often finding a positive effect of class size reduction (e.g., Urquiola (2006), Browning and Heinesen (2007); Akabayashi and Nakamura (2014); Hojo and Senoh (2019); and Gilraine (2020), among many). Similarly, Gary-Bobo and Mahjoub (2013) use a rich sample of students from French junior high schools with a panel structure to obtain small but significant and negative effects of class size on probabilities of being promoted to the next grade in grades 6 and 7.

The positive class size reduction effect is not necessarily guaranteed with this identification strategy. Hoxby (2000), utilizing a long panel dataset where class size changes are driven by idiosyncratic variation of cohort size and also applying the Maimonides' Rule, found an insignificant effect of class size on students' academic performance. Leuven et al. (2008) found insignificant class size effect using a Norwegian administrative database to estimate the effect of class size on student achievement at the end of lower-secondary school with identification strategy based on maximum class-size rules and population variation. Dobbelsteen et al. (2002) found that after correcting for endogeneity with different instruments driven by the budget for teacher salary depending on the total number of enrollment in a school, pupils in large classes do no worse – and sometimes even better – than identical pupils in small classes. Angrist et al. (2019) found that the effect was insignificant, which cast doubt on the effectiveness of class size reduction. Similarly, Ito et al. (2020) found no effect of class size reduction in Japanese compulsory schools.

While these studies contribute significantly to the debate on the effect of class size reduction, research using observational data exploring how this effect varies across different types of classes with varying student compositions is relatively limited. Applying a similar but more robust identification strategy to potential endogenous matching between students and teachers by controlling student-teacher fixed effects to observational data, our paper contributes to the literature by examining the heterogeneity of the class size effect based on the characteristics of classmates such as the distribution of student ability and gender.

There are several papers, studying heterogeneous class size effects in various dimensions. Using experimental data, Ding and Lehrer (2011) showed a heterogeneous effect of class size reduction based on teachers' characteristics. While utilizing experimental data, their simple categorization of classes into large and small limited the analysis of marginal changes in class size. With observational data, Bonesrønning (2003) found that the class size effect differs among student subgroups and that smaller classes yield greater benefits in Norwegian lower secondary schools with a high proportion of students from intact families. Bosworth (2014) reported that students struggling academically appear to benefit more from class size reductions than high-achieving students. Nandrup (2016) found that class size effects vary across grades in Danish public compulsory schools. Tanaka (2020) finds the heterogeneous effects of class size reduction by socioeconomic status of students' household using microdata from a large municipality in Japan. Kedagni et al. (2021) applied structural estimation methodologies to Greek administrative data, finding a hump-shaped effect of class size on academic achievement. Their analysis quantified the costs and benefits of teacher hiring and firing but did not examine the heterogeneity of the class size reduction effect by student characteristics. Our current paper emphasizes the role of distribution of baseline abilities and gender of classmates for the class size effects.

As a closely related paper to ours, Bandiera et al. (2010), using rich university data from the UK, demonstrated a heterogeneous marginal effect of class size reduction depending on the composition of students' ability within the class. Similarly, Diette and Raghav (2015), without using instrumental variables, found negative correlations between class size and college students' academic achievements, particularly pronounced for students with lower baseline achievements. Their identification strategy relied on controlling for student and/or teacher fixed effects separately. However, controlling for student and teacher fixed effects independently may be susceptible to the endogeneity of matching between students and teachers. Moreover, the absence of an effective class size cap in the university setting rendered the application of a regression discontinuity design infeasible. In contrast, our paper addresses the identification challenge by applying a regression discontinuity design while simultaneously controlling for fixed effects at the student-teacher matched pair level.

Several studies have highlighted the potential manipulation of grade size to create small classes, which invalidates the identification strategy based on the class size cap. Angrist et al. (2017) found that a significant portion of the class size effects in Italy could be attributed to such manipulation. Angrist et al. (2019) detected incentives for schools in Israeli public primary schools to manipulate grade size to achieve smaller classes, suggesting that predicted class size calculated using actual grade size might be endogenous. Urquiola and Verhoogen (2009) identified the possibility of grade size manipulation to attain small classes in Chilean schools with students from affluent households. To address this concern in our study, we employ the McCrary (2008) test to demonstrate that the potential for grade size manipulation is minor in our specific setting.

Finally, we offer interpretations of our findings based on a theoretical model built on Lazear (2001). The observed variations in the results of class size effects can be attributed to the heterogeneity of class composition. To explain the mechanism through which class size influences students' academic performance, Lazear (2001) interprets the education production process as a combat against students' disturbance behavior (the education production function is described in Appendix A). In larger classes, the probability of the educational process being disrupted by a student increases. Our findings provide supportive evidence for

Lazear (2001)'s theoretical framework.

3 Institutional Background

Japanese compulsory education is based on the Constitution, specifically, the Fundamental Law of Education, promulgated in 1947. Compulsory education consists of six years of primary education in primary school and three years of lower secondary education in middle school. Children who are six years old before April 2nd start first grade in primary school on April 1st and receive schooling for nine years in total compulsory. The majority of primary and middle schools are publicly financed and run by the education board of the local municipality. Each municipality establishes school districts and assigns students to designated public schools based on their residential addresses.[1]

Public primary schools (grades one to six) in Japan are subject to upper limits on class size, as stipulated by the Act on Standards for Class Formation and Fixed Number of School Personnel of Public Compulsory Education Schools. This law permits local government education boards to establish their own upper class size limits, provided these limits are below the national standard set by the Ministry of Education, Culture, Sports, Science and Technology of Japan. In the context of our study, primary schools had an upper limit of 35 students for the first and second grades, and 40 students for the remaining grades (third, fourth, fifth, and sixth grades) before 2021. Importantly, students experience simultaneous changes in the upper limits of class sizes and class reassignments when students transition from second to third grade.

All teachers in Japanese primary schools, regardless of whether the school is national, public, or private, are required to hold a teaching license. Teachers are assigned to schools by the prefectural government's education board, which oversees teacher personnel matters. Given that teachers with three to seven years of tenure at a school are typically transferred to another school, public school teachers do not have the opportunity for long-term self-selection into specific schools. Once assigned to elementary schools, classroom teachers are responsible for teaching all subjects to students in their classes. All teachers are certified to teach any grade within elementary schools and can be assigned as classroom teachers to any grade level. The assignment of teachers to specific classes within schools is at the discretion of the school principal, potentially leading to endogenous matching between teachers and students/classes. To address this potential endogeneity, our empirical analysis controls for teacher-student fixed effects.

[1] Students are also allowed to attend private or public schools run by the national government. In the case of school attendance at private and national schools, students need to take entrance examinations.

4 Empirical Strategy

We estimate the effect of class size reduction using the following regression model:

$$Y_{ijcgst} = \beta_0 + \beta_1 C_{jcgst} + \mathbf{X_{ijcgst}} \gamma_\mathbf{1} + \mathbf{P_{-ijcgst}} \gamma_\mathbf{2} + \mathbf{T_{jcgst}} \gamma_\mathbf{3} + f(E_{gst}) + d_{ij} + d_g + d_s + d_t + \epsilon_{ijcgst} \quad (1)$$

where Y_{ijcgst} is the outcome variable (i.e., student academic performance) for student i, in class c taught by teacher j, grade g, school s, and year t. C_{jcgst} is the size of class c taught by teacher j, in grade g at school s in year t. $\mathbf{X_{ijcgst}}$ is the vector of observable characteristics of students (e.g., socioeconomic status of households), $\mathbf{P_{-ijcgst}}$ is the baseline characteristics of classmates *excluding* i (the average, minimum, and maximum scores of baseline test, and the share of female students in a class), $\mathbf{T_{jcgst}}$ is the teacher characteristics (e.g., teaching experience), and $f(E_{gst})$ is a polynomial of enrollment in grade g at school s in year t, E_{gst}. We include up to the third polynomial of grade size. d_g, d_s, d_t are the grade, school and year fixed effects, respectively. d_{ij} represents the fixed effects for student-teacher pairs. ϵ_{ijcgst} is the idiosyncratic error term.

The inclusion of student-teacher fixed effects is important for the robust identification of class size effects, as it accounts for potential sorting between teachers and students based on unobservable factors. While controlling for student fixed effects and teacher fixed effects separately addresses unobserved characteristics of students and teachers independently, our institutional background suggests a more complex dynamic. As previously discussed, teachers and students are reassigned annually, and the matching of teachers to classes is determined by the school principal. Furthermore, the rules governing teacher assignment to classes are often highly school-specific, potentially leading to endogenous matching based on unobservable characteristics of both teachers and students. To address this, our identification strategy leverages the variation in class size within a specific teacher-student pairing. When we observe a student taught by the same teacher for at least two consecutive years but experiencing different class sizes across these grades, we interpret the resulting variation in academic outcomes as being caused by the change in class size. This approach provides a more robust estimate against potential endogenous matching between students and teachers.

Although our preferred identification strategy, employing student-teacher fixed effects, mitigates the risk of endogeneity, it may substantially reduce the variation in class size available for identifying the class size effects. In the subsequent section, we will demonstrate that sufficient variation in class size remains within our sample of students taught by the same teacher over multiple years to credibly identify the effect of class size reduction. In addition, as a robustness check to assess the external validity of our primary findings, we will present estimation results from a specification that controls for student fixed effects and teacher fixed effects separately. While this alternative identification strategy, without student-teacher pair fixed effects, may be more susceptible to endogenous matching, it retains greater variation in class size within individual students. By presenting both sets of results, we aim to provide a comprehensive and robust analysis of the class size effects.

Another threat to the identification of the class size effect, specifically the coefficient β_1, is the potential endogeneity of class size. For instance, districts with higher average student academic performance may benefit from richer educational resources outside of schools, such as the prevalence of cram schools and private tutoring. This attractive educational environment could lead more families to reside in these districts, resulting in increased student enrollment and, consequently, larger class sizes. This correlation between unobserved factors (e.g., external educational resources, family sorting) and class size would bias our estimates of the causal effect. To address this potential endogeneity problem, we adopt the instrumental variable (IV) approach, following the methodology of Angrist and Lavy (1999). We utilize the upper limit on class size to calculate a predicted class size, which serves as an instrument for the potentially endogenous actual class size variable. The rationale behind this instrument is that the predicted class size, derived from the grade enrollment and the mandated maximum class size, is correlated with the actual class size but is plausibly exogenous to the unobserved determinants of student academic performance. Given the number of students enrolled in grade g at school s in year t, assuming that classes are divided almost equally, we have

$$\hat{C}_{jcgst} = \frac{E_{gst}}{int[\frac{E_{gst}-1}{\bar{C_{gt}}}]+1} \tag{2}$$

where $\bar{C_{gt}}$ is the maximum possible number of students of a class. In our data, $\bar{C_{gt}} = 35$ for the first and second graders since year 2012, and $\bar{C_{gt}} = 40$ for the rest grade and before 2012.

To investigate the heterogeneity of the class size effect based on the characteristics of classmates, we incorporate interaction terms between class size and the baseline characteristics of classmates, as specified in the following equation:

$$Y_{ijcgst} = \beta_0 + \beta_1 C_{jcgst} + C_{jcgst}\mathbf{P_{-ijcgst}}\boldsymbol{\beta_2} + \mathbf{X_{ijcgst}}\boldsymbol{\gamma_1} + \mathbf{P_{-ijcgst}}\boldsymbol{\gamma_2} + \mathbf{T_{jcgst}}\boldsymbol{\gamma_3} + f(E_{gst}) + d_{ij} + d_g + d_s + d_t + \epsilon_{ijcgst} \tag{3}$$

In the subsequent analyses, we explore the heterogeneous effects of class size reduction by considering the distribution of classmates' ability and gender. Specifically, we include interaction terms between class size and the average baseline test scores, as well as the maximum and minimum baseline test scores of classmates. The average, maximum, and minimum baseline test scores are calculated using the test scores of classmates in year t, obtained from the previous year/grade $t-1$. In addition to the baseline academic performance of peers, we further investigate the heterogeneous effects of class size reduction based on the proportion of female classmates.

We estimate the models with interaction terms using instrumental variables, where the predicted class size interacted with the relevant classmate characteristics serves as instruments for these interaction terms. Our identification strategy combines both instrumental variable (IV) and fixed-effects approaches. Conditional on the included fixed effects, we exploit the variation in class size induced by changes in class composition from grade to grade, as well as the regulatory upper limit on class size, following the spirit of Angrist and Lavy (1999). The identification of heterogeneous class size reduction effects using equation (3) is also achieved through the within-student-teacher variation in class size across different grades. Specifically, conditioning on the fixed effects, we estimate the model by instrumenting the actual class size with the predicted class size derived from grade size. Furthermore, the interaction terms involving actual class size are instrumented by the interactions between the predicted class size and the respective class characteristics (e.g., mean baseline test scores, the proportion of female students in a class, etc.).

Figure 1 illustrates the relationship between grade size, predicted class size, and actual class size in our sample for each grade level. As the figure reveals, there are downward jumps in actual class size when the grade size reaches multiples of 35 for grade 2 and 40 for the remaining grades. This pattern indicates that class size caps are binding in some grades and schools, as class divisions are typically implemented when the grade size reaches these thresholds. These observations suggest a positive correlation between the predicted class size (our instrument) and the potentially endogenous actual class size in the first-stage regression. We will report the first-stage F-statistics in the subsequent instrumental variable regression analyses.

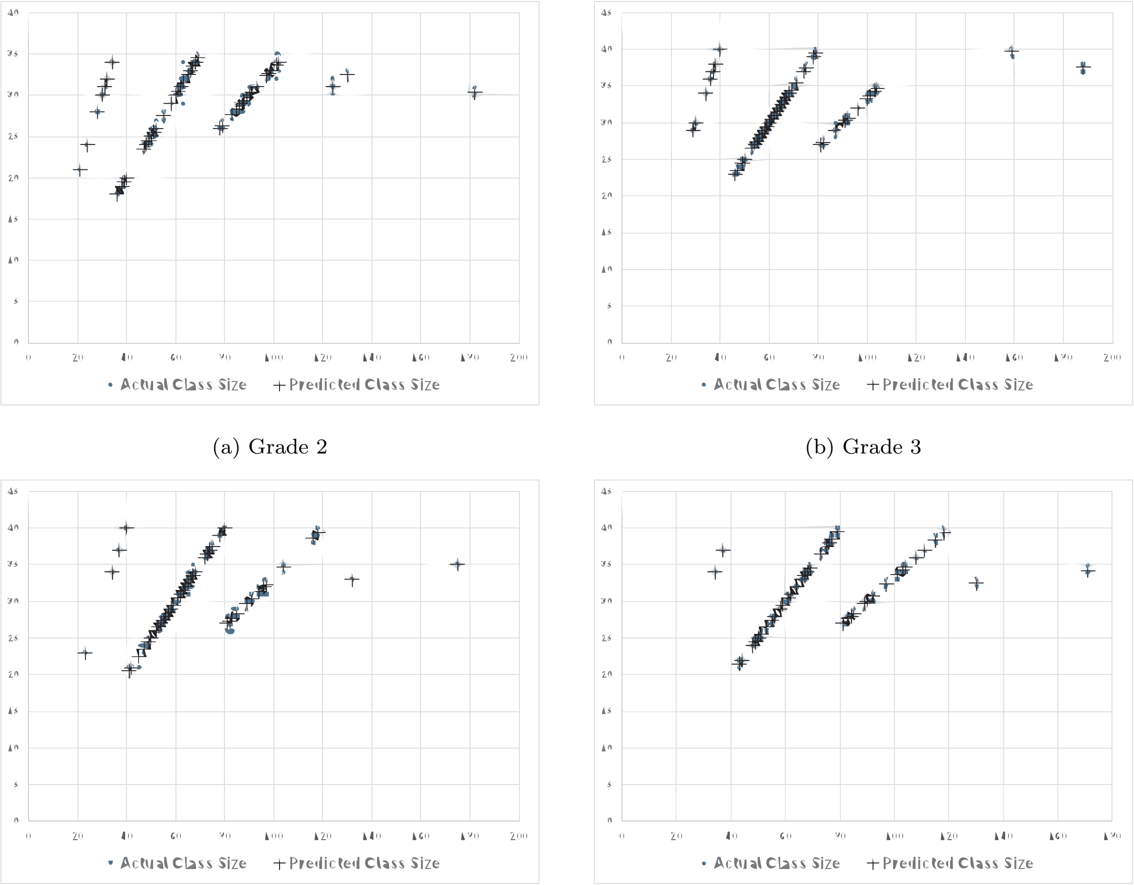
5 Data

5.1 Data Source and Descriptive Statistics

This study employs administrative data collected by the Education Board of a specific city from 2010 to 2016.[2] As of 2015, this city, a large municipality within the Tokyo Metropolis, comprised over 300,000 households and a population exceeding 600,000 residents. Our dataset encompasses information from all 74 public primary schools within the city, including students enrolled in the second through sixth grades.

[2] Due to the confidentiality agreement, we cannot identify the city by name in our paper.

Figure 1: *Actual and Predicted Class Size*



(a) Grade 2

(b) Grade 3

(c) Grade 4

(d) Grade 5

Note: The relationship between grade size (horizontal axis) and actual and predicted class sizes (vertical axis) is plotted. A class size cap of 40 students is in place, with the exception of grade 2, which has a cap of 35 students.

The municipality conducts an annual assessment of student learning in its public primary and lower secondary schools each April when the Japanese academic year starts. These tests are designed to assess students' achievements up to the previous year-grade. All students in these schools, with the exception of first-grade elementary school students, participate in these assessments. The assessments comprise examinations in mathematics and Japanese language, as well as a student questionnaire that gathers information on students' behavior and opinions regarding their school life. The primary objective of this assessment is to provide diagnostic feedback to both students and teachers to facilitate the improvement of student learning. It is considered a low-stakes examination for students, teachers, and schools, as there are no sanctions or penalties imposed on teachers or schools based on the assessment results. Furthermore, the results have no bearing on students' academic grades within their schools. In our analyses, we utilize test scores (specifically, the correction rate, ranging from 0 to 100) that have been normalized within each year and grade level across the municipality with mean 0 and standard deviation 1. In the regression analyses, we regress test scores in year $t+1$ on class size, the baseline scores of classmates, and other covariates from year t.

Our dataset comprises three components: student academic test data, student socioeconomic status data, and teacher survey data. The student academic test data includes standardized assessments that evaluate students' academic achievement. For students in the second through sixth grades, we have their scores in both Japanese language and mathematics. The student socioeconomic status data indicates whether students receive school financial assistance from the government. Students from lower-income households and/or those experiencing family hardships (e.g., parental divorce) are eligible for government aid. In our data, a student's socioeconomic status is represented by a binary variable indicating the receipt of any form of financial assistance. We restrict our sample to students without school transfer during our observation periods in addition to the availability of all information for the analysis and with class size larger than or equal to 10. This selection of samples reduced the total number of observations from 149,727 to 145,264.

Table 1 presents the summary statistics of students' test scores, class size, and grade size. The left panel reports the descriptive statistics of all students in the sample. Students' math and Japanese language scores have been normalized to have a mean of 0 and a standard deviation of 1 within each grade level and year before sample selection. The average grade size (total number of students in a grade at a school) is approximately 79 students, while the average class size is around 31 students. The standard deviation of grade size is approximately 28 students, indicating considerable variation in grade size across schools.

The right panel of Table 1 presents the descriptive statistics for the subsample of students who were taught by the same teacher for at least two years. Given that our primary identification strategy relies on the within-student-teacher pair variation in class size, these students are crucial for estimating the class size effects. Although the size of this subsample is reduced to approximately one-third of the full sample size, the descriptive statistics remain broadly similar to those observed for the entire sample.

Table 1: *Descriptive Statistics*

	Full				Same Teacher			
VARIABLES	mean	sd	min	max	mean	sd	min	max
Japanese score	0.010	0.986	-5.787	2.051	0.027	0.975	-5.057	2.031
Math score	0.010	0.987	-5.941	1.782	0.035	0.970	-5.477	1.587
Class size	31.481	4.351	10	41	31.731	4.329	10	40
Grade size	78.891	28.174	10	217	78.636	26.816	10	217
Female student	0.492	0.500	0	1	0.495	0.500	0	1
School financial assist.	0.347	0.476	0	1	0.357	0.479	0	1
Teaching experience	11.806	10.435	1	44	11.069	9.681	1	40
Tenure (current school)	3.473	2.052	0	17	3.592	1.939	0	16
Teacher age	37.294	10.276	22	65	36.455	9.691	22	62
Base Japanese score	-0.004	0.267	-1.335	0.877	0.013	0.269	-1.147	0.877
Base math score	-0.004	0.275	-1.237	0.861	0.013	0.278	-1.111	0.760
Sample size	145,264				44,122			

Note: The columns labeled ”Full” present results for the full sample, while the columns labeled ”Same Teacher” present results for the subsample of students taught by the same teacher for at least two consecutive years. Test scores are standardized within each year-grade-subject based on the correction rate, resulting in a distribution with a mean of 0 and a standard deviation of 1 before excluding observations with missing values for the variables listed in the table. Observations with a class size smaller than 10 and/or involving school transfers have been excluded from the analysis.

5.2 Source of Variations

Our identification strategy follows Angrist and Lavy (1999), with minor technical adaptations. While Angrist and Lavy (1999) utilize the Maimonides' Rule as an instrumental variable, the variation of which stems from year-to-year changes in grade size, we also employ the Maimonides' Rule as an IV. However, our unit of observation is the student-teacher pair. Consequently, if a student was instructed by different teachers in different years, these student-teacher pairs are treated as distinct observations. Therefore, our identification relies exclusively on students who were taught by the same teacher across multiple years but also experienced variations in class size. The variation in our instrumental variable arises from changes in the grade size of students taught by the same teacher over different years. These year-to-year fluctuations in grade size occur due to student mobility, including students transferring into and out of schools.

First, we present the number of students taught by the same teachers across different grade levels. Although class compositions are typically reshuffled as students progress to the next grade, the probability of being taught by the same teacher for two consecutive grades is non-negligible. Table 2 displays the number of students taught by the same teacher for multiple years (not necessarily consecutive). As shown in the table, the vast majority of students in our sample (71.88%) changed teachers every year. This implies that our identification of the class size reduction effect is derived from the remaining 28.12% of students. Among the students who had the same teacher for multiple years, the majority (27.12% of the total sample) experienced the same teacher for two years.

Table 2: *Total Number of Students Taught by Same Teacher*

Observed year	Number of Students	Percent
1 year	101,142	71.88
2 years	39,402	27.12
3 years	3,972	2.73
4 years	728	0.50
5 years	20	0.01

Note: The share of students taught by the same teachers is reported. The total number of observations in the analysis is 145,264.

Next, we present summary statistics on the variation in class size experienced by the student taught by the same teacher across different years. This is crucial because our identification of the effect of class size reduction is predicated on the existence of such variation for this specific subsample of students. Table 3 examines the difference in class size between the current year (t) and the previous year ($t-1$), two years prior ($t-2$), and three years prior ($t-3$), respectively, for students taught by the same teacher for two, three, and four consecutive years. The first row indicates that the average change in class size between two consecutive grades for students taught by the same teacher in those grades is 0.632 students. This relatively small average change is attributable to the fact that many students experience no change in class size between two consecutive grades.

Although the majority of students experienced minor changes in class size, a subset of students experienced substantial changes due to the upper class size limit regulation. As shown in Column (6) of Table 3, approximately 4-11% of students taught by the same teachers experienced class size changes resulting from new class formation and/or the closure of existing classes due to enrollment fluctuations and the upper class size limit. Consequently, some students experienced a reduction in class size of up to 13 students within a single year, while others experienced an increase of up to 21 students within one year. As we extend the period of consecutive years considered, the mean of the class size variation increases. The standard deviation, minimum, and maximum values of class size changes within a student-teacher pair serve as key sources of identification in our analysis.

Table 3: *Changes of Class Size of Students Taught by Same Teacher*

	(1)	(2)	(3)	(4)	(5)	(6)
Changes of class size	obs.	mean	std. dev.	min	max	% big changes
1-year difference	21,662	0.632	2.242	-13	21	4.07
2-year difference	2,312	1.214	3.155	-4	20	4.97
3-year difference	662	1.873	3.481	-2	20	10.88
4-year difference	169	2.349	3.048	-2	8	7.10

Note: Changes in class size for students taught by the same teachers across consecutive grades are reported. Column (6) presents the percentage of students who experienced a change in the total number of classes within one grade from the current grade to the subsequent grade.

The variation in grade size arises from students transferring between schools and, in some instances, school consolidations. Students are permitted to transfer schools at any point during the academic year, contingent on their specific circumstances. These inter-school transfers, including those resulting from school consolidations, coupled with the annual reshuffling of classes, lead students to experience different teachers, classmates, and class sizes across different years. Table 4 presents summary statistics of the within-school cohort change in grade size experienced by students from one year to the subsequent year. The first row treats each student as an observation, while the second row treats each student-teacher pair as an observation. In both cases, each school-cohort experienced changes in cohort size with a standard deviation of approximately 7 students, on average. These within-school cohort variations in grade size provide a crucial source of exogenous variation for the identification of class size effects.

5.3 Balancing Tests

In Japanese public primary schools, class compositions are often reshuffled between grade levels, leading to changes in classmates. Furthermore, teacher assignments to classes can also vary from one grade to the next. Given the absence of explicit rules governing the

Table 4: *Standard Deviation of Grade Size within School-Cohort*

	mean	std. dev.	min	max
For all student				
Standard deviation of grade size	6.890	6.885	0	39.748
Student-teacher pairs				
Standard deviation of grade size	7.216	6.483	0	39.748

Note: Summary statistics of the standard deviation of grade size are reported for each school-cohort. The full sample comprises 787 school-cohorts, while the student-teacher pair subsample includes 461 school-cohorts.

assignment of students and teachers to specific classes, it is worth exploring the potential systematic correlation between teachers', classes' and students' characteristics. To examine the possibility of sorting based on teachers' characteristics and students' ability, we conduct balancing tests, the results of which are presented in Tables 5 and 6.

Regarding the assignment of teachers to classes based on teachers' characteristics, we regressed teaching experience, teacher's tenure at the current school, and teacher's age on class characteristics such as class size and the baseline characteristics of students within the class. The results of these regressions are reported in Table 5. We find no strong evidence of a systematic correlation between teaching experience, teacher's tenure at the current school, or teacher's age and class size, nor with the baseline ability of students in our sample. These findings suggest that there is no systematic matching between teachers and classes based on these observed teacher and class characteristics.

As a further balancing test, we compare students whose data contribute to the identification of class size effects (i.e., those who were taught by the same teacher but experienced a change in class size due to class reshuffling) against those whose class size remained stable. This comparison is based on observable characteristics of both students and teachers. We conduct this balancing test by regressing a binary indicator for students who had the same teacher but experienced a change in class size on the following covariates: baseline Japanese language and mathematics test scores, receipt of school financial assistance, class size in the previous grade, teacher's teaching experience, tenure at the current school, and age, in addition to school, year, and grade fixed effects. We estimate this model using the subsample of students who were taught by the same teacher for at least two grades, incorporating student-teacher pair fixed effects. It is important to note that, within this specification, time-invariant characteristics of students and teachers, such as gender, cannot be included as they are absorbed by the student-teacher pair fixed effects.

The results of this balancing test are reported in Column (1) of Table 6. Most importantly,

Table 5: *Balancing Tests: Class Level*

	(1)	(2)	(3)
VARIABLES	Teach exp.	Sch. tenure	Age
Baseline Japanese score	0.054	0.019 0	-0.001
	(0.158)	(0.183)	(0.025)
Baseline Math score	-0.107	0.035	0.003
	(0.109)	(0.167)	(0.025)
Class size	0.000	0.010	0.000
	(0.009)	(0.009)	(0.001)
%Financial assist	-0.173	0.167	0.034
	(0.237)	(0.371)	(0.040)
%Female students	-0.245	-0.014	-0.050
	(0.320)	(0.475)	(0.060)
Max Japanese score	-0.028	0.031	-0.020
	(0.132)	(0.186)	(0.020)
Max math score	-0.102	0.203	-0.014
	(0.253)	(0.284)	(0.038)
Min Japanese score	-0.045	-0.037	-0.003
	(0.035)	(0.037)	(0.005)
Min math score	0.033	-0.016	0.002
	(0.030)	(0.038)	(0.006)

Note: Estimates using class-level data, controlling for school, year, and grade fixed effects, as well as a third-order polynomial of grade size, are reported. Standard errors clustered at the school level are presented in parentheses. The total number of observations is 5,362. *Significant at 10%; **Significant at 5%; ***Significant at 1%

the baseline Japanese language and mathematics test scores, as well as the receipt of school financial assistance, did not exhibit a systematic correlation with the change in class size for students taught by the same teachers. Furthermore, the teacher's total teaching experience, tenure at the current school, and age are uncorrelated with the class size change. Class size in the previous grade shows a positive correlation with the change in class size, which is plausible given that larger grade sizes (and consequently, larger class sizes) increase the likelihood of class size changes due to grade-level reshuffling. Based on these results, we conclude that the class size change is not driven by students' characteristics, particularly their academic ability, conditional on being taught by the same teacher.

Table 6: *Balancing Tests: Individual Level*

	(1)	(2)
VARIABLES	Class size change	Same teacher
Baseline Japanese score	0.009	0.015***
	(0.032)	(0.005)
Baseline Math score	-0.029	0.003
	(0.039)	(0.0005)
Class size	0.044***	-0.011
	(0.020)	(0.010)
Financial assist	-0.013	-0.008
	(0.057)	(0.008)
Teach exp.	-0.184	0.028
	(0.162)	(0.041)
Sch. tenure.	0.025	0.012
	(0.031)	(0.018)
Age	-0.053	-0.053
	(0.067)	(0.086)
Observations	21,662	56,806

Note: Column (1) presents estimates using the subsample of students taught by the same teacher for at least two consecutive grades, controlling for student-teacher pair fixed effects and a third-order polynomial of grade size. Column (2) presents estimates using the subsample of students observed for two consecutive years, controlling for student fixed effects, teacher fixed effects, and a third-order polynomial of grade size. Standard errors clustered at the school level are reported in parentheses. *Significant at 10%; **Significant at 5%; ***Significant at 1%

Building on the previous balancing test, which confirmed the absence of a systematic correlation between baseline test scores and class size changes within student-teacher pairs, we address the possibility that teachers might differentially retain certain types of students based on their prior experience. To investigate this, we regress a binary indicator of being taught by the same teacher for two consecutive years on students' characteristics (i.e., baseline Japanese language and mathematics test scores, and the receipt of school financial assistance in the previous grade), incorporating both student and teacher fixed effects.

The results of this analysis are presented in Column (2) of Table 6. These findings indicate that students' and teachers' characteristics are generally not strongly correlated with the probability of being taught by the same teacher, with the exception of students' baseline Japanese language test score. This score exhibits a positive correlation with the probability of having the same teacher for two consecutive grades. This correlation potentially introduces selection bias into our estimation results that employ student-teacher fixed effects. We will address the robustness of our main findings in light of this potential bias in Section 7.1.

5.4 Manipulation Check

A potential threat to the identification of the effects of class size reduction is the manipulation of grade size to achieve smaller classes. As highlighted by Angrist et al. (2019), if schools have an incentive to manipulate grade size to obtain smaller classes, the predicted class size calculated using the actual grade size could be endogenous, thereby invalidating our instrumental variable. Following the recommendation of Angrist et al. (2019), we test for the possibility of grade size manipulation by conducting the McCrary (2008) density test. We use grade size as the running variable to test for discontinuities in the density of grade size at multiples of the class size cap for each grade level. For the second grade, the class size cap was 35 after 2012, while for all other grades, the cap was 40.

Table 7: *Manipulation Tests*

Cutoff	1st	2nd	3rd
Grade			
2	-0.259	-2.187	-0.051
	(0.795)	(0.029)	(0.960)
3	0.437	-0.980	-1.112
	(0.662)	(0.327)	(0.266)
4	1.587	0.572	0.211
	(0.113)	(0.567)	(0.833)
5	-0.120	-1.129	0.368
	(0.904)	(0.259)	(0.713)
6	0.684	-1.033	0.323
	(0.494)	(0.299)	(0.747)

Note: Results from McCrary (2008)'s density discontinuity test, using grade size as the running variable, are reported to assess the discontinuity in the density of grade size at the multiples of the class size cap for each grade. For the second grade, the class size cap was 35 after 2012. For all other grades, the class size cap was 40. The t-statistic of the test, with the null hypothesis of no bunching at each cutoff point, is reported. The p-value is presented in parentheses.

Table 7 reports the t-statistic of the results from the McCrary (2008) test, with the corresponding p-value in parentheses. We found no statistically significant discontinuities in the grade size density around the cutoff points, with the exception of some bunching of grade size below the second cutoff point (i.e., grade size of 70) for the second graders. While the manipulation of grade size to create smaller classes poses a significant concern for the identification of class size effects, as discussed in detail by Angrist et al. (2019), our findings indicate the opposite: bunching occurring *below* the cutoff point, which would lead to larger class sizes. Therefore, we conclude that the potential manipulation of grade size to achieve smaller class sizes is not a major concern in our dataset.

6 Results

We apply our benchmark model (1) to evaluate the average effect of class size reduction and the regression model (3) to evaluate different types of heterogeneity in class size reduction effects. This section presents the results of our analysis.

6.1 Average Effect

Table 8 presents the results of our benchmark model. The first two columns display the findings for Japanese language scores, while the last two columns show those for math test scores. The columns labeled "FE" report the estimation results from the model with student-teacher fixed effects using Ordinary Least Squares (OLS), and the columns labeled "FEIV" report the results from the model with student-teacher fixed effects using the instrumental variable (IV) method. Overall, the coefficient for class size is negative, suggesting that smaller class sizes are associated with higher academic achievement in both Japanese language and math. Without the use of an instrument, the coefficients on class size are negative and statistically significant, as shown in Columns (1) and (3). These OLS results indicate that a one-student decrease in class size is associated with an average increase of 0.0046 standard deviations in Japanese scores and 0.0073 standard deviations in math scores. However, when class size is instrumented by the predicted class size, the coefficient on class size is statistically significant only for math test scores (-0.0087 in Column (4)), and the coefficient on class size is negative but not statistically significant for Japanese language (-0.0028 in Column (2)). The results from our preferred specification, which includes student-teacher fixed effects and the instrumental variable, suggest that, on average, class size reduction is beneficial for math achievement but less effective for Japanese language. One potential explanation for this difference is that mathematics, as a scientific discipline, may benefit more from focused instruction and specific techniques. Conversely, Japanese language learning might

Table 8: *Average Class Size Effect*

VARIABLES	Japanese		Math	
	(1)	(2)	(3)	(4)
MODEL	FE	FEIV	FE	FEIV
Class size	-0.00464**	-0.00278	-0.00730***	-0.00870***
	(0.00211)	(0.00328)	(0.00195)	(0.00309)
School financial assistance	0.00317	0.00321	0.000670	0.000641
	(0.0188)	(0.0188)	(0.0168)	(0.0168)
Total teaching experience	0.00957**	0.00955**	0.00293	0.00295
	(0.00469)	(0.00469)	(0.00365)	(0.00365)
Tenure at the current school	0.00148	0.00149	0.00492	0.00491
	(0.00581)	(0.00581)	(0.00543)	(0.00543)
Teacher age	0.0649***	0.0645***	-0.0101	-0.00982
	(0.0238)	(0.0238)	(0.0209)	(0.0209)
Baseline ability of language	-0.223***	-0.223***	0.0332	0.0334
	(0.0256)	(0.0256)	(0.0230)	(0.0230)
Baseline ability of math	-0.00604	-0.00606	-0.196***	-0.196***
	(0.0243)	(0.0243)	(0.0225)	(0.0225)
Percent of female students	-0.0155	-0.0140	-0.0290	-0.0301
	(0.186)	(0.186)	(0.170)	(0.170)
Share of financial assistance receivers	-0.0303	-0.0290	-0.0866	-0.0876
	(0.0699)	(0.0699)	(0.0639)	(0.0639)
Max Japanese score	-0.00420	-0.00594	-0.000535	0.000779
	(0.0217)	(0.0219)	(0.0198)	(0.0199)
Max math score	0.0226	0.0205	0.0952***	0.0968***
	(0.0340)	(0.0341)	(0.0299)	(0.0301)
Min Japanese score	0.0145**	0.0150**	0.0136***	0.0133**
	(0.00584)	(0.00588)	(0.00517)	(0.00519)
Min math score	-0.00663	-0.00656	0.00522	0.00517
	(0.00605)	(0.00605)	(0.00573)	(0.00573)
Predicted class size in the 1st stage		0.5785***		0.5785***
		(0.0185)		(0.0185)
F-stat. of the 1st stage		313.22***		313..22***

Note: All specifications include third-order polynomials of grade size, school fixed effects, year fixed effects, grade fixed effects, and student-teacher pair fixed effects, in addition to the variables listed in the table. A student's own status is excluded from the calculation of the mean, minimum, and maximum scores, as well as the share of female students and school financial assistance recipients. The FE models are estimated using Ordinary Least Squares (OLS). The FEIV models are estimated using Two-Stage Least Squares (2SLS), with the predicted class size as the instrument for the actual class size. Standard errors are clustered at the student-teacher pair level. The number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%

depend more on communicative interaction and may not always have definitive right or wrong answers. Consequently, more concentrated teaching may be less crucial for Japanese compared to math.

In the bottom panel of Table 8, we report the F-statistic of the first stage, as well as the estimated coefficient and standard error of the predicted class size as the excluded instrument for the actual class size in the first-stage regression. The coefficient on the predicted class size in the first-stage regression is positive and statistically significant at the 1% level. Furthermore, the first stage is significant at the 1% level, based on the F-test result. These results indicate the relevance of our instrument (i.e., its correlation with the endogenous variable). Combined with the findings on the absence of manipulation in grade size, as discussed in Figure 1 and Section 5.4 (which suggests the excludability of our instrument), these findings collectively support its validity.

It is worthwhile to compare the results of Ordinary Least Squares (OLS), including fixed effects for student-teacher pairs with those of the instrumental variable (IV) method. For the Japanese language score, the coefficients on class size in both models are negative, and their magnitudes are comparable (-0.0046 with FE and -0.0028 with FEIV). However, only the OLS coefficient is statistically significant, while the IV coefficient is imprecisely estimated. For the math score, both coefficients are negative and statistically significant. Notably, the estimated coefficient with the IV method is larger in absolute value than the one obtained with OLS. This finding aligns with the discussion in Angrist and Lavy (1999), suggesting that OLS estimates might be biased upward due to a potential positive correlation between class size and unobserved heterogeneity, such as school "quality." However, we find that the results from OLS and the IV method are generally similar and close to each other. This may indicate that our results are primarily identified by variations in class size within student-teacher pairs, rather than variations driven by the class size cap rule.[3]

6.2 Heterogeneous Effects

Table 9 presents the heterogeneous effects of class size reduction on Japanese language scores across classes with varying baseline ability and female student share. Column (1) replicates the coefficient on class size from Column (2) in Table 8. Column (2) examines the effect of

[3] Although the coefficient on class size for Japanese is imprecisely estimated with the IV method, it is interesting to discuss the potential reason for the different estimated effects across subjects. Given that the set of compliers in the IV estimates is consistent across subjects, we interpret this result (specifically, the larger coefficients for mathematics than for Japanese in general) as indicating that our IV method, which leverages large class size variations due to the class size cap as a source of identification, is more effective in identifying the causal effect of class size for mathematics than for Japanese. This may be because mathematics content is relatively easier to adjust to varying class sizes than Japanese, leading to a more responsive causal effect of class size in mathematics.

class size reduction on Japanese scores in classes with different mean baseline Japanese scores of classmates. The coefficient on the interaction term between class size and mean baseline Japanese score is negative and statistically significant, indicating that the positive effect of class size reduction is more pronounced in classes with higher mean baseline Japanese scores. Similarly, Column (3) shows the class size effect interacting with the maximum and minimum baseline Japanese scores of classmates. The coefficient on the interaction term between class size and maximum baseline Japanese score is positive and statistically significant at the 10% level, suggesting that the positive effect of class size reduction is weaker in classes with higher maximum baseline Japanese scores.

Column (4) reports the results when all three interaction terms with class size are included. The coefficient on the interaction term with the mean score remains negative and statistically significant, and that on the interaction term with the maximum score remains positive and statistically significant. Column (5) reports the results with the interaction term between class size and the proportion of female students in a class. The coefficient on this interaction term is negative and statistically significant, indicating that the positive effect of class size reduction is stronger in classes with a higher proportion of female students. Finally, Column (6) presents the model including all interaction terms. We can confirm the robustness of our findings: the positive effect of class size reduction is stronger for classes with high mean baseline scores and a high proportion of female students, and weaker for classes with high maximum baseline scores.

Table 10 similarly presents the heterogeneous effects of class size reduction on math scores across classes with varying baseline ability and female student share. Column (1) replicates the coefficient on class size from Column (4) in Table 8. Column (2) examines the effect of class size reduction on math scores in classes with different mean baseline math scores of classmates. The coefficient on the interaction term between class size and mean baseline math score is negative and statistically significant, indicating that the positive effect of class size reduction is more pronounced in classes with higher mean baseline math scores. Column (3) shows the class size effect interacting with the maximum and minimum baseline math scores of classmates. The coefficients on both interaction terms are negative and statistically significant, suggesting that the positive effect of class size reduction is stronger in classes with higher maximum and minimum baseline math scores.

Column (4) reports the results when all three interaction terms are included. While the statistical significance of the coefficient on the interaction term with the mean score is no longer evident, the coefficients on the maximum and minimum baseline math scores remain negative and statistically significant. Column (5) reports the results with the interaction term between class size and the proportion of female students in a class. The coefficient

Table 9: *Heterogeneous Class Size Effect: Japanese Score*

Japanese score	(1)	(2)	(3)	(4)	(5)	(6)
Class size	-0.00278	-0.00259	-0.0117*	-0.0101	0.0399***	0.0305**
	(0.00328)	(0.00328)	(0.00642)	(0.00648)	(0.0142)	(0.0151)
x Mean score		-0.0103**		-0.0163***		-0.0146***
		(0.00428)		(0.00510)		(0.00512)
x Max score			0.00672*	0.00964**		0.00964**
			(0.00393)	(0.00402)		(0.00402)
x Min score			-0.000390	0.00169		0.00168
			(0.00114)	(0.00132)		(0.00132)
x Female share					-0.0956***	-0.0911***
					(0.0303)	(0.0304)
F-stat. of the 1st stage						
Class size	313***	300***	300***	287***	316***	291***
x Mean score		388650***		373136***		354613***
x Max score			18311***	21383***		20979***
x Min score			54072***	59110***		57062***
x Female share					1030***	936***

Note: All specifications for math scores include the class mean, maximum, and minimum baseline Japanese scores, the class share of female students, the class share of school financial assistance recipients, teacher's total teaching experience, teacher's tenure at the current school, teacher's age, student's own status of school financial assistance, school fixed effects, year fixed effects, grade fixed effects, and student-teacher fixed effects, in addition to the variables listed in the table. A student's own status is excluded from the calculation of the class means. Class size and its interaction terms are instrumented by the predicted class size and its interactions. Standard errors are clustered at the student-teacher pair level. The number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%

on this interaction term is negative and statistically significant at the 10% level, indicating that the positive effect of class size reduction is stronger in classes with a higher proportion of female students, consistent with the findings for Japanese scores. Finally, Column (6) presents the model including all interaction terms. The positive effect of class size reduction is stronger for classes with high minimum baseline scores, high maximum scores (significant at the 10% level), and a high proportion of female students (significant at the 10% level).

Similarly to the case for the average class size effects, we report the F-statistics and its significance for the first stage in the bottom panel of Tables 9 and 10. For the specifications including the interaction terms with class size, we report the F-statistics of the first stage regressions using the interaction terms with actual class size as dependent variable of the first stage in addition to the first stage with class size itself as the dependent variable. We can

Table 10: *Heterogeneous Class Size Effect: Math Score*

Math score	(1)	(2)	(3)	(4)	(5)	(6)
Class size	-0.00870***	-0.00845***	-0.00530	-0.00538	0.0150	0.0148
	(0.00309)	(0.00309)	(0.00855)	(0.00871)	(0.0127)	(0.0144)
x Mean score		-0.00959***		0.000302		0.000477
		(0.00340)		(0.00456)		(0.00456)
x Max score			-0.0121**	-0.0122**		-0.0112*
			(0.00602)	(0.00607)		(0.00611)
x Min score			-0.00398***	-0.00403***		-0.00408***
			(0.00106)	(0.00139)		(0.00139)
x Female share					-0.0531*	-0.0480*
					(0.0272)	(0.0273)
F-stat. of the 1st stage						
Class size	313***	302***	281***	258***	316***	288***
x Mean score		459516***		432250***		440401***
x Max score			6562***	6345***		6260***
x Min score			47786***	46490***		45518***
x Female share					1030***	971***

Note: All specifications include the class mean, maximum, and minimum baseline math scores, the class mean, maximum, and minimum baseline Japanese scores, the class share of female students, the class share of school financial assistance recipients, teacher's total teaching experience, teacher's tenure at the current school, teacher's age, student's own status of school financial assistance, school fixed effects, year fixed effects, grade fixed effects, and student-teacher fixed effects, in addition to the variables listed in the table. A student's own status is excluded from the calculation of the class means. Class size and its interaction terms are instrumented by the predicted class size and its interactions. Standard errors are clustered at the student-teacher pair level. The number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%

confirm that the first stage is statistically significant at the 1% level across all specifications. Although the coefficients on the excluded instruments (namely, predicted class size and its interaction terms with other class characteristics) are not reported in the tables, they are also strongly significant. These results support the validity of our instrument.

7 Discussions

7.1 Robustness

7.1.1 Clustering Level of Standard Errors

In our analyses thus far, we have reported standard errors clustered at the student-teacher pair level. As a robustness check, we also report standard errors clustered at the school-cohort level. The first and second columns in Table 11 present the estimation results of the model from Column (6) of Tables 9 and 10, but with standard errors clustered at the school-cohort level. Although some coefficients lose statistical significance under this alternative clustering, the coefficients on the interaction terms with mean baseline scores and female share remain statistically significant for Japanese scores, and the coefficients on the interaction terms with minimum baseline scores remain statistically significant for math scores.

7.1.2 Alternative Specification of Fixed-Effects

In our benchmark specification, we include student-teacher pair fixed effects to control for potential endogeneity arising from student-teacher matching. Although our specification is robust against this type of endogenous matching, it reduces the variation in class sizes used for the identification of the class size effect. We utilize the variation in class size within student-teacher pairs for identification, as discussed in Subsection 5.2. However, concerns may arise regarding the robustness and external validity of our benchmark results. To address these concerns, we estimate the model (3) with student fixed effects and teacher fixed effects included separately as an alternative specification. This specification leverages the variation in class size within students and within teachers to identify the class size effects. Consequently, we can utilize the variation in class size across grades for students taught by different teachers, in addition to the variation for students taught by the same teachers.

Columns (3) and (4) of Table 11 report the estimated class size effects with student fixed effects and teacher fixed effects included separately. The results are largely consistent with those obtained using student-teacher pair fixed effects. Notably, our main findings—the importance of class mean baseline ability for Japanese class size effects and the minimum baseline achievement of classmates for math class size effects—remain robust in the specification with separate student and teacher fixed effects.

In Subsection 5.3, we found a positive correlation between baseline Japanese test scores and the probability of being taught by the same teacher in two consecutive grades. This correlation may potentially introduce sample selection bias in the benchmark specification with

Table 11: *Robustness Checks*

	Clustering		Separate FE		Raw score	
	(1)	(2)	(3)	(4)	(5)	(6)
	Japanese	Math	Japanese	Math	Japanese	Math
Class size	0.0305	0.0148	0.0222***	0.0076	0.355	0.573**
	(0.0213)	(0.0206)	(0.0085)	(0.0083)	(0.270)	(0.273)
x Mean score	-0.0146**	0.0005	-0.0071**	0.0008	-0.362***	0.0290
	(0.0069)	(0.0079)	(0.0032)	(0.0029)	(0.0921)	(0.0863)
x Max score	0.0096	-0.0112	0.0072**	-0.0086**	0.149**	-0.182
	(0.0066)	(0.0094)	(0.0028)	(0.0038)	(0.0734)	(0.117)
x Min score	0.0017	-0.0041**	0.0018**	-0.0035***	0.0520**	-0.0775***
	(0.0018)	(0.0020)	(0.0009)	(0.0008)	(0.0235)	(0.0255)
x Female share	-0.0911**	-0.0480	-0.0640***	-0.0302**	-0.843	-1.611***
	(0.0418)	(0.0393)	(0.0148)	(0.0135)	(0.548)	(0.514)

Note: All specifications include the class mean, maximum, and minimum baseline math scores, the class mean, maximum, and minimum baseline Japanese scores, the class share of female students, the class share of school financial assistance recipients, teacher's total teaching experience, teacher's tenure at the current school, teacher's age, student's own status of school financial assistance, school fixed effects, year fixed effects, grade fixed effects, and student-teacher fixed effects, in addition to the variables listed in the table. A student's own status is excluded from the calculation of the class means. Class size and its interaction terms are instrumented by the predicted class size and its interactions. Standard errors are clustered at the school-cohort level for Columns (1) and (2), at the student level for Columns (3) and (4), and at the student-teacher pair level for Columns (5) and (6). The total number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%

student-teacher fixed effects. The results reported in Columns (3) and (4) of Table 11 also provide suggestive evidence regarding sample selection bias based on baseline Japanese test scores. In these specifications, we utilize the variation in class sizes within students, including both those retained and those who changed teachers. The robustness results reported indicate that the potential bias due to the positive correlation between baseline Japanese test scores and the probability of being taught by the same teachers is not substantial in our context.

7.1.3 Raw Test Scores

In our main specifications, we standardized test scores within each year-grade-subject. Given that we utilize student panel data, this standardization allows us to compare test results across grades with varying levels of test difficulty. To assess the robustness of our results to the standardization of test scores, we estimate the model using non-standardized test scores as dependent variables. The results of this analysis are presented in Columns (5) and (6) of Table 11. For Japanese language scores, the finding that the class size reduction effect is positively associated with the baseline mean score remains robust when using raw scores as the dependent variable. Similarly, for math test scores, the finding that the class size reduction effect is positively associated with the minimum baseline score also proves robust. Therefore, we conclude that our main findings regarding the heterogeneous effects of class size are robust to the standardization of test scores.

7.2 Other Forms of Heterogeneity

In our analyses thus far, we have identified heterogeneous class size effects based on classmates' characteristics. We now discuss other potential forms of heterogeneous effects: heterogeneity across quantiles and the non-linearity of the class size effect.

7.2.1 Quantile Regression

It is worthwhile to explore the potential heterogeneity of the class size effects across different quantiles of the outcome test scores. Given that the estimation results with and without the instrument are similar, as observed in Table 8, we discuss the potential heterogeneity across quantiles based on the results obtained from the quantile regression model without the instrument. These results are reported in Table A1. Although the class size effects were not statistically significant in this analysis, the magnitude of the coefficients is comparable to those in Table 8 and remains relatively stable across different quantiles.

7.2.2 Non-linearity

In the main specifications examined in this paper, we assumed a linear relationship between class size and outcome test scores. However, this relationship may be non-linear, as demonstrated by Kedagni et al. (2021) in the Greek context. To explore this possibility, we extended our benchmark analysis by including the squared class size as an additional regressor, and we instrumented both class size and its squared term with predicted class size and its squared counterpart. The results of this extension are reported in the Appendix as Table A2. Our analysis did not provide evidence of a significant non-linear relationship in our study.

7.3 Interpretation

To explain how class size impacts educational activities, Lazear (2001) provides a theoretical framework with explanatory power and applicability. He posits that each student has a probability, denoted by p, of being well-behaved and not disturbing other students. In the absence of disruptions, the total value of educational production in a class is V. However, if even one student misbehaves, the teacher must address this, temporarily halting the educational production process. Thus, each student's disruptive behavior creates a negative externality that affects the educational output of the entire classroom.

Lazear (2001)'s model can be extended in various dimensions, and our results provide support to Lazear (2001)'s model in an extended dimension. The heterogeneity in students' gender or socioeconomic status are factors that can determine their value within a unit of time and/or the possibility of being well-behaved. If we kick out a student with a higher possibility of disturbing other students, the possibility of disturbing others will be reduced, which can even be viewed as a benefit for the education activity. However, if we kick out a student with a lower possibility of disturbing others, that will be a lost for the education because such a student could contribute more to class education by helping others or asking thought-provoking questions.

Our finding that class size reduction is more effective when the average baseline academic performance is higher aligns with this framework. We interpret this as students with higher academic performance being less likely to disrupt the class and having a greater capacity for educational production within a given time. Consequently, when the class is disrupted, the loss of educational production is more substantial.

Our results regarding the heterogeneity based on female student percentage also fit Lazear (2001)'s model. Classes with a higher proportion of female students benefit more from class size reduction in both Japanese and mathematics, likely because these classes experience fewer disruptions. Data from the Japanese Ministry of Education in 2021 indicates that among all recognized primary school bullying cases, 246,211 were perpetrated by males and 174,686 by females, suggesting that males are more prone to engaging in problematic behaviors that can disrupt the educational process.

Another finding that can be explained by, and provides supporting evidence for, Lazear (2001)'s model is the heterogeneous effect of class size reduction in classes with different baseline scores of the lowest-performing student. We consider a student's baseline score to reflect their ability, which may encompass both cognitive skills and non-cognitive behaviors. It is reasonable to assume that the lowest-performing student in a class is more likely to disrupt learning than other students. Conversely, a student with a higher baseline score is less likely to disrupt the educational process, for example, by asking questions with obvious answers. Therefore, when reducing class size in a class where the lowest-performing student has a particularly low baseline score, the potential for disruption is likely reduced more significantly than in a class where the lowest-performing student's baseline score is higher. A detailed discussion with a formal mathematical model is provided in Appendix A.

Our results regarding the average effect of class size reduction are comparable to, though somewhat smaller in magnitude than, those reported in other literature such as Bandiera et al. (2010) and Urquiola (2006). Bandiera et al. (2010) found that a one standard deviation reduction in class size increases students' test scores by 0.074 standard deviations, while Urquiola (2006) reported an increase of up to 0.3 standard deviations. Angrist and Lavy (1999) found effects ranging from 0.13 to 0.27 standard deviations for pupils. While these studies measure class size reduction in terms of standard deviations, our measure is the number of students in the class. Given that the standard deviation of class size in our dataset is 4.3 students, the effect of a one standard deviation decrease in class size can be approximated as 0.04 standard deviations in mathematics scores and about 0.02 standard deviations in Japanese scores.[4] Our estimates are slightly smaller than those in the literature, partly due to the inclusion of teacher-student fixed effects.[5]

8 Conclusion

This paper primarily estimated the heterogeneous effects of class size reduction on academic outcomes, focusing on class average baseline test scores, the percentage of female students in a class, and the baseline scores of the highest- and lowest-performing students in a class. Our findings indicated that the main effect of class size reduction is positive for students' academic performance and particularly strong for mathematics.

We also found that smaller classes are more beneficial for students in Japanese language classes with higher average peer baseline scores. Another notable finding is that for math scores, classes where the lowest-performing student has a higher baseline math score benefit more from class size reduction. These findings suggest that the heterogeneity of class size reduction effects is *heterogeneous across subjects.* In particular, focusing more attention

[4] Hojo and Senoh (2019) report that a class size reduction of one student results in an improvement of 0.018 in math scores and 0.014 in Japanese scores within the Japanese context.

[5] A cost-benefit analysis based on our estimates might be of interest. However, a comprehensive calculation of the cost-benefit ratio would require estimates of the long-run effects on not only academic outcomes but also other skills, such as non-cognitive skills, in addition to a number of assumptions specific to the educational context. This is beyond the scope of the current paper and will be considered as future research.

on the lowest-performing students in their classes is effective in improving the academic performance of students in math through class size reduction. We proposed an interpretation of our empirical findings by extending the theoretical framework proposed by Lazear (2001).

One important avenue of future research is to unpack the mechanism why the effects of class size reduction are heterogeneous. Although our paper focuses on the disruptive peer/bad apple effect, we could also explore more mechanisms such as individualized attention, boutique, and rainbow (e.g., Hoxby and Weingarth (2005)). Unpacking the mechanism of heterogeneous effects, we could obtain rich implications for optimal education policy in schools.

References

Akabayashi, H. and R. Nakamura (2014): "Can Small Class Policy Close the Gap? An Empirical Analysis of Class Size Effects in Japan," *The Japanese Economic Review*, 65, 253–281.

Angrist, J., V. Lavy, J. Leder-Luis, and A. Shany (2019): "Maimonides' Rule Redux," *American Economic Review: Insights*, 1, 309–24.

Angrist, J. D., E. Battistin, and D. Vuri (2017): "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, 9, 216–49.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114, 533–575.

Bandiera, O., V. Larcinese, and I. Rasul (2010): "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students," *Economic Journal*, 120, 1365–1398.

Bonesrønning, H. (2003): "Class Size Effects on Student Achievement in Norway: Patterns and Explanations," *Southern Economic Journal*, 69, 952–965.

Bosworth, R. (2014): "Class size, class composition, and the distribution of student achievement," *Education Economics*, 22, 141–165.

Browning, M. and E. Heinesen (2007): "Class Size, Teacher Hours and Educational Attainment," *Scandinavian Journal of Economics*, 109, 415–438.

Diette, T. M. and M. Raghav (2015): "Class Size Matters: Heterogeneous Effects of Larger Classes on College Student Learning," *Eastern Economic Journal*, 41, 273–283.

Ding, W. and S. Lehrer (2011): "Experimental estimates of the impacts of class size on test scores: robustness and heterogeneity," *Education Economics*, 19, 229–252.

Dobbelsteen, S., J. Levin, and H. Oosterbeek (2002): "The causal effect of class size on scholastic achievement: distinguishing the pure class size effect from the effect of changes in class composition," *Oxford Bulletin of Economics and Statistics*, 64, 17–38.

Gary-Bobo, R. J. and M.-B. Mahjoub (2013): "Estimation of Class-Size Effects using "Maimonides' Rule" and Other Instruments: the Case of French Junior High Schools," *Annals of Economics and Statistics*, 193–225.

Gaub, M. and C. L. Carlson (1997): "Gender Differences in ADHD: A Meta-Analysis and Critical Review," *Journal of the American Academy of Child and Adolescent Psychiatry*, 36, 1136–1139.

Gershon, J. and J. Gershon (2002): "A Meta-Analytic Review of Gender Differences in ADHD," *Journal of Attention Disorders*, 5, 143–154, pMID: 11911007.

Gilraine, M. (2020): "A Method for Disentangling Multiple Treatments from a Regression Discontinuity Design," *Journal of Labor Economics*, 38, 1267–1311.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177.

——— (2003): "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113, 64–98.

——— (2006): "School Resources," in *Handbook of the Economics of Education*, ed. by E. Hanushek and F. Welch, Elsevier, vol. 2 of *Handbook of the Economics of Education*, chap. 14, 865–908.

Hojo, M. and W. Senoh (2019): "Do the disadvantaged benefit more from small classes? Evidence from a large-scale survey in Japan," *Japan and the World Economy*, 52.

Hoxby, C. M. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics*, 115, 1239–1285.

Hoxby, C. M. and G. Weingarth (2005): "Taking race out of the equation: School reassignment and the structure of peer effects," *NBER Conference Paper*.

Ito, H., M. Nakamuro, and S. Yamaguchi (2020): "Effects of class-size reduction on cognitive and non-cognitive skills," *Japan and the World Economy*, 53.

Kedagni, D., K. Krishna, R. Megalokonomou, and Y. Zhao (2021): "Does class size matter? How, and at what cost?" *European Economic Review*, 133, 103664.

Krueger, A. (1999): "Experimental Estimates of Education Production Functions," *The Quarterly Journal of Economics*, 114, 497–532.

Krueger, A. B. (2003): "Economic Considerations and Class Size," *Economic Journal*, 113, 34–63.

Lavy, V., M. D. Paserman, and A. Schlosser (2012a): "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom," *The Economic Journal*, 122, 208–237.

Lavy, V., O. Silva, and F. Weinhardt (2012b): "The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools," *Journal of Labor Economics*, 30, 367–414.

Lazear, E. P. (2001): "Educational Production," *The Quarterly Journal of Economics*, 116, 777–803.

Leuven, E., H. Oosterbeek, and M. Rønning (2008): "Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway," *Scandinavian Journal of Economics*, 110, 663–693.

Li, Q. (2006): "Cyberbullying in Schools: A Research of Gender Differences," *School Psychology International*, 27, 157–170.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714, the regression discontinuity design: Theory and applications.

Myers, D. E., A. M. Milne, K. Baker, and A. Ginsburg (1987): "Student Discipline and High School Performance," *Sociology of Education*, 60, 18–33.

Nandrup, A. B. (2016): "Do class size effects differ across grades?" *Education Economics*, 24, 83–95.

Tanaka, R. (2020): "Toward the Evidence-Based Education Policy Making by Local Governments-An Analysis of Heterogenous Effects of Class-size Reduction on Students Educational Achievements Using Administrative Panel Data," *Journal of Social Security Research (in Japanese)*, 5, 325–340.

Urquiola, M. (2006): "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia," *The Review of Economics and Statistics*, 88, 171–177.

Urquiola, M. and E. Verhoogen (2009): "Class-Size Caps, Sorting, and the Regression-Discontinuity Design," *American Economic Review*, 99, 179–215.

A Illustration of Education Process

We illustrate our results regarding the heterogeneous effects of class size reduction by an extended version of the model by Lazear (2001). In our formal model, in a class, there are n students, where each student belongs to a type $t \in \{1, 2, ..., T\}$. There are n_t students of each type t, such that $\sum_{t=1}^{t=T} n_t = n$, and we assume $n_t > 0$ for all t. The probability that each student with type t will well-behave in class is p_t, and without loss of generality, we assume $1 > p_1 > p_2.... > p_T > 0$. Following Lazear (2001)'s model, let V denote the value of a unit of learning, which is determined by the market value of human capital and the likelihood that a student is focusing on learning during the given instant. We assume that $V > 0$, i.e., human capital has a strictly positive value. Therefore, the expected value of education in a time unit can be specified as follows:

$$\pi = V \times \Pi_{j=1}^{j=T} p_j^{n_j}$$

In our model, π can be viewed as the students' academic performance, measured by test scores. The possibility of each type being well-behaved, p_j, is determined by its baseline academic performance, i.e., students with higher baseline scores have higher possibility of being well-behaved. It is relatively intuitive that students with higher baseline scores are less likely to disturb other students, for example, less likely to ask questions which all other students know the answer.

We first want to show that reducing the class size will have a positive effect on the education production, regardless of the type of students whose number is reduced. We take the derivative of π with respect to n_j for any given j, we get

$$\frac{\partial \pi}{\partial n_j} = V ln(p_j) \Pi_{j=1}^{j=T} p_j^{n_j}$$

Because $V > 0$, $\Pi_{j=1}^{j=T} p_j^{n_j} > 0$ and $ln(p_j) < 0$, $\frac{\partial \pi}{\partial n_j} < 0$, for any type j. This can explain our result that the main effect of class size reduction is positive since the first derivative of education production with respect to class size is always negative.

Take the derivative of π with respect to p_k, we have

$$\frac{\partial \pi}{\partial p_k} = (V \Pi_{j \neq k} p_j) \times \frac{\partial p_k^{n_k}}{\partial p_k} = (V \Pi_{j \neq k} p_j) n_k p_k^{n_k - 1}$$

Since all elements in the formula, V, p_j and n_j are strictly positive, we have $\frac{\partial \pi}{\partial p_k} > 0$. Therefore, the education production is a strictly increasing function of the possibility of any type of students being well-behaved, which is also quite intuitive.

We are now going to provide some possible theoretical explanation to the heterogeneous effect of class size reduction. In particular, we explain why class size reduction has a significantly positive effect for classes whose bottom student is better. To do this, we take the derivative to $\frac{\partial \pi}{\partial n_k}$ with respect to p_k. That is, we look at the effect of class size reduction by reducing the number of a type of students with the possibility of well-behave of p_k.

$$\frac{\partial^2 \pi}{\partial n_k \partial p_k} = V\frac{1}{p_k}\Pi_{j=1}^{j=T} p_j^{n_j} + \frac{1}{p_k}V ln(p_k) n_k \Pi_{j=1}^{j=T} p_j^{n_j} = V\Pi_{j=1}^{j=T} p_j^{n_j}[1 + n_k ln(p_k)] \times \frac{1}{p_k}$$

We can see that the sign of $\frac{\partial^2 \pi}{\partial n_k \partial p_k}$ depends on the sign of the term $[1+n_k ln(p_k)]$. If p_k is large and n_k is small, it would be more likely that the term is positive, which means the effect of class size reduction is less positive only if we reduce the number of the type of students whose possibility of well-behave is high enough, and meanwhile, there are enough number of this type of students.

If we put some specific numbers into the term $[1 + n_k ln(p_k)]$, we can get a sense of how large p_k and how small n_k should be, in order to make it positive. If $n_k = 10$ and $p_k = 0.9$, we have $ln(0.9)$=-0.105 and $n_k ln(p_k)$=-1.05, so $[1 + n_k ln(p_k)]$=-0.05, still slightly smaller than zero. And if we reduce n_k by 1, i.e. $n_k = 9$, we get $[1 + n_k ln(p_k)]$=0.05, which is slightly greater than zero. When p_k=0.8, if $n_k = 5$, $[1+n_k ln(p_k)] = -0.1 < 0$, and if $n_k = 4$, $[1 + n_k ln(p_k)] = 0.12 > 0$. We can see that when p_k is larger, for the term to be positive, the number of students in this type k can be allowed to be larger than when p_k is smaller.

This theoretical result can explain our finding that the heterogeneity in the effect of class size reduction is not sensitive to the baseline score of the highest-performing student in the class, but is sensitive to the baseline score of the lowest-performing student. Class size reduction has a significantly larger positive effect when the class's lowest-performing student has a higher baseline score, but the effect is insignificant when the highest-performing student's baseline score varies.

For the student with the worst baseline score in a class, p_k is relatively small, so $\frac{\partial^2 \pi}{\partial n_k \partial p_k} < 0$ is more likely to hold. Therefore, when reducing the number of this type of students n_k, and increase p_k, we can see that the class size reduction effect on education production is even more positive. However, for the top student, p_k is high, and when the number of this type is smaller enough, $\frac{\partial^2 \pi}{\partial n_k \partial p_k} > 0$ is more likely to hold. When we reduce this type of students, the effect of class size reduction is less obvious, and that's why the coefficient associated with the interaction term of class size and the score of bottom student in the class is negative and significant, but the coefficient associated with the class size and the score of the top student is not significant.

B Appendix Tables

Table A1: *Quantile Regression*

	(1)	(2)	(3)	(4)	(5)
Quantile	.1	.3	.5	.7	.9
Japanese score					
Class size	-0.00458	-0.00458	-0.00464	-0.00470	-0.00470
	(0.00967)	(0.00959)	(0.00360)	(0.00302)	(0.00309)
Math score					
Class size	-0.00728	-0.00728	-0.00730	-0.00731	-0.00731
	(0.0513)	(0.0507)	(0.0167)	(0.0132)	(0.0138)

Note: Class size effects in this table are estimated by Ordinary Least Squares (OLS) with student-teacher pair fixed effects. All specifications include the class mean, maximum, and minimum baseline math scores, the class mean, maximum, and minimum baseline Japanese scores, the class share of female students, the class share of school financial assistance recipients, teacher's total teaching experience, teacher's tenure at the current school, teacher's age, student's own status of school financial assistance, grade dummies, and a third-order polynomial of grade size. A student's own status is excluded from the calculation of the class means. Standard errors are clustered at the student-teacher pair level. The number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%

Table A2: *Nonlinear class size effect*

	(1)	(2)	(3)	(4)
Japanese score				
Class size	-0.0359*	-0.0312	-0.0280	-0.0358
	(0.0208)	(0.0209)	(0.0312)	(0.0315)
Class size squared	0.000509	0.000430	0.000396	0.000530
	(0.000334)	(0.000336)	(0.000517)	(0.000523)
Math score				
Class size	-0.0565***	-0.0529***	-0.0432	-0.0418
	(0.0195)	(0.0197)	(0.0289)	(0.0292)
Class size squared	0.000800**	0.000740**	0.000548	0.000532
	(0.000316)	(0.000319)	(0.000478)	(0.000483)
Controls	NO	YES	NO	YES
IV	NO	NO	YES	YES

Note: All specifications include grade dummies, a third-order polynomial of grade size, and student-teacher fixed effects. The following variables are included as controls: the class mean, maximum, and minimum baseline math scores; the class mean, maximum, and minimum baseline Japanese scores; the class share of female students; the class share of school financial assistance recipients; teacher's total teaching experience; teacher's tenure at the current school; teacher's age; and student's own status of school financial assistance. A student's own status is excluded from the calculation of the class means. For the instrumental variable (IV) estimates, class size and its squared term are instrumented by the predicted class size and its squared term. Standard errors are clustered at the student-teacher pair level. The number of observations is 145,264. *Significant at 10%; **Significant at 5%; ***Significant at 1%