RIETI Discussion Paper Series 26-E-060

The Effect of Class-size Reduction on Quality of Classroom Environment

SUGITA, Soichiro
Keio University

NAKAMURO, Makiko
RIETI

Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 26-E-060

August 2026

The Effect of Class-size Reduction on Quality of Classroom Environment*

Soichiro Sugita (Keio University)

Makiko Nakamuro (Keio University)

Abstract

This paper estimates the effects of class-size reduction on the quality of the classroom environment, exploiting the discontinuous variation in class size generated by Japan's statutory 40-student cap. A 10-student reduction improves composite indices of teacher-student relationships and classroom climate by 0.03–0.04 standard deviations respectively; school-reported discipline and violence also improve. Effects are stable across the pre- and post-COVID periods, concentrated in elementary grades, and largest where the prior classroom environment was weak, consistent with Lazear's disruption model. Effects on test scores are essentially zero. The benefits of smaller classes operate on the classroom environment rather than on academic achievement.

Keywords: teacher-student relationship, classroom climate, class size reduction,

JEL classification: I21, I28, H52

*This study is conducted as a part of the Project "Comprehensive Study to Promote Evidence-Based Policy Making (EBPM) in Japan" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). An earlier version of this paper was presented at a Discussion Paper Seminar for the paper at RIETI. I am grateful to the seminar participants for their valuable comments. I also benefited from insightful comments from Hirotake Ito (CyberAgent). This study was made possible by access to anonymized microdata from the Saitama Prefectural Academic Achievement and Learning Survey — including student test scores, and student and school questionnaires — provided by the Saitama Prefectural Board of Education. I would like to express my sincere gratitude for their generous support.

1 Introduction

In 2021, Japan enacted a legislative reform to progressively lower the statutory class-size cap in public elementary schools from 40 to 35 students, one of the largest input-based education investments in recent decades. The reform sparked an intense policy debate over whether class-size reduction is worth its considerable cost. That debate, however, was conducted almost entirely on the terrain of academic achievement, where the evidence in Japan as elsewhere points to null or at best modest effects (e.g., Ito et al., 2020; Hanushek, 2003; Angrist et al., 2019). Evidence on what smaller classes deliver beyond test scores was, and remains, remarkably scarce.

This paper asks whether the benefits of class-size reduction operate through a different channel: the quality of the classroom environment itself. Specifically, we examine whether smaller classes improve students' relationships with their teachers and the overall classroom climate as well as school-reported behavioral indicators including chronic absenteeism, learning discipline, and student-on-student violence. This focus aligns with recent international policy frameworks that treat the quality of classroom environment as core outcomes of effective schooling (OECD, 2017a, 2021a).

There are good reasons to believe these outcomes matter in their own right. School climate has been identified as a foundational condition for healthy student development (Darling-Hammond and DePaoli, 2020), and the quality of teacher–student relationships predicts adolescent development over a seven-year horizon (Ansari et al., 2020). Exposure to disruptive peers and the broader learning environment has been linked to long-run earnings and educational attainment (Chetty et al., 2011a; Carrell et al., 2018). Conversely, negative school experiences carry lasting costs: childhood bullying victimization is associated with worse adult health (Momose and Ishida, 2024), chronic absence predicts lower adult earnings and employment (Gershenson et al., 2017; Cattan et al., 2023), and school violence is linked to adverse developmental trajectories (Diagne, 2009).

Estimating the causal effect of class size is complicated by endogeneity. Struggling students may be deliberately assigned to smaller classes, or experienced teachers to more difficult ones; naive regressions treating class size as exogenous would then be biased. To address this, we follow Angrist and Lavy (1999) and exploit the discontinuous variation in class size generated by Japan's 40-student cap under the Act on Standards for Class Formation and Staffing Levels at Public Compulsory Education Schools: a cohort of 40 students forms a single class, while a cohort of 41 is split into two. Using predicted class size based on grade enrollment as an instrument, we estimate two-stage least squares (2SLS) models on individual-level panel data from the Saitama Skill Assessment, a census of roughly 300,000 public school students per cohort spanning 2017 to 2024. The panel structure allows us to include school and student fixed effects, absorbing time-invariant unobserved heterogeneity at both levels.

We find that class-size reduction significantly improves both dimensions of the quality of classroom environment. A 10-student reduction raises composite indices of teacher–student relationships and classroom climate by 0.03–0.04 standard deviations (SD) each. School-reported behavioral outcomes also improve by 0.26 SD for learning discipline and 0.15 SD for violence, although these school-level estimates are likely upper bounds, as aggregation tends to inflate estimated resource effects (Hanushek et al., 1996). The effects are concentrated in elementary grades, larger in schools serving more disadvantaged students, and stable across the pre- and post-COVID periods. In terms of a more interpretable metric, a 10-student reduction raises the probability that a student reports a calm learning environment by roughly 3 percentage points.

This study makes two contributions. First, it provides new evidence on outcomes that prior work has largely overlooked. As this study, Ito et al. (2020) find no significant effects of class size on students' cognitive and non-cognitive skills. We shift the outcome of interest from student skills to the quality of the classroom environment and complement students' subjective assessments with school-reported behavioral indicators, finding consistent beneficial effects across both data sources. Second, we speak to the question of whether class-size effects survive in newer data. Revisiting the original Israeli setting with more recent data, Angrist et al. (2019) find that the class-size effects documented in Angrist and Lavy (1999) have essentially vanished, raising the possibility that class-size effects were an artifact of particular times and places. Our data extend through 2024, spanning the COVID-19 pandemic and the rollout of Japan's 2021 class-size reform. Although we confirm no achievement effects, the effects on the classroom environment, by contrast, are remarkably stable throughout: they are of similar magnitude before and after the pandemic and across the reform period.

The remainder of this paper is organized as follows. Section 2 reviews the prior literature on class-size effects. Section 3 describes the data and presents summary statistics. Section 4 outlines the empirical strategy. Section 5 reports the estimation results and presents additional analyses, including robustness checks and subgroup comparisons. Section 6 concludes with a summary of the findings, a discussion of the study's limitations, and directions for future research.

2 Literature Review

This paper sits at the intersection of two literatures: a large body of work on class-size effects, which has focused overwhelmingly on academic achievement, and a growing literature establishing teacher–student relationships and classroom climate as consequential outcomes in their own right.

2.1 Class-Size Effects on Academic Achievement

Since the Coleman Report (Coleman, 1968), researchers have debated whether school inputs affect student outcomes, and class size has been the most studied input of all. The evidence is voluminous but unsettled. Experimental evidence from Tennessee's Project STAR found positive achievement effects concentrated in the earliest grades (Krueger, 1999; Hanushek, 1999). Quasi-experimental studies exploiting statutory class-size caps, beginning with Angrist and Lavy (1999) in Israel, found significant effects in some settings (Urquiola, 2006; Fredriksson et al., 2013) but not others (Hoxby, 2000; Leuven et al., 2008). Most strikingly, Angrist et al. (2019), returning to the original Israeli setting with more recent data, find that the effects documented two decades earlier have essentially vanished, and Angrist et al. (2017) reach a similarly skeptical conclusion for Italy. Reviews of this literature conclude that input-based policies deliver little robust achievement gain (Hanushek, 2003; Hanushek and Woessmann, 2017).

The Japanese evidence mirrors this ambiguity. Akabayashi and Nakamura (2014) find essentially no achievement effects; Senoh and Hojo (2016) find effects concentrated in low-SES schools; Hojo (2013) reports significant but small effects using TIMSS, while Niki (2013) finds none; and Ito et al. (2020), using the same Saitama Skill Assessment employed in this study, find no effects on either cognitive or non-cognitive skills. Household responses may partly explain these modest effects: Asakawa et al. (2026) show that families compensate for larger classes with more home study time and parental involvement.

More Recent work suggests that average effects may conceal heterogeneity tied to classroom composition. Tanaka and Wang (2025), using administrative panel data from a large Tokyo-area municipality with student–teacher pair fixed effects, find positive average effects of class-size reduction on mathematics scores. Drawing on Lazear (2001)'s model of education as production subject to disruption externalities, they show that these effects vary systematically with the ability distribution of classmates, including the baseline score of the lowest-achieving student. Their interpretation is that class size matters through the probability that instruction is disrupted. Notably, however, disruption itself is unobserved in their data and proxied by peer ability and gender composition. This points to a gap the present study is positioned to fill: our data contain direct measures of the classroom environment and school reports of discipline and violence, which may be the very margin that the Lazear (2001)'s framework identifies as the operative channel.

2.2 Class-Size Effects on Non-Academic Achievement

A smaller but growing literature examines class-size effects beyond academic achievement. The STAR follow-ups show that early classroom quality affects earnings, college attendance, and home ownership at age 27, even where test-score effects fade (Chetty et al., 2011a), and Fredriksson et al. (2013) find effects of class size on non-cognitive traits and long-run wages in Sweden. In Japan, Nakamuro (2017) finds that smaller elementary classes reduce absenteeism; Hojo (2021) links higher student-teacher ratios to teacher workload stress; Oikawa et al. (2022) show that smaller classes slowed influenza transmission; and Sugita (2025), using cross-sectional TIMSS data, finds positive effects on students' satisfaction with school life. These studies beyond the test-score margin suggest the positive returns to smaller classes. However, none directly examines the quality of classroom with individual-level panel data and a credible identification strategy.

2.3 The Quality of Classroom Environments as Educational Outcomes

The outcomes we study are the quality of classroom environments, such as teacher–student relationships and classroom climate, which have historically been treated in the economics of education as inputs into achievement. A substantial body of evidence from developmental psychology and, increasingly, economics indicates that they merit treatment as outcomes in their own right.

They have documented long-run consequences that operate independently of contemporaneous test scores. The quality of classroom environment predicts adolescent development, including academic motivation, behavioral adjustment, and mental health(Ansari et al., 2020; Darling-Hammond and DePaoli, 2020). The economics literature reaches parallel conclusions from a different direction: exposure to disruptive peers in elementary school reduces earnings decades later (Carrell et al., 2018), and the quality of early classroom environments affects adult outcomes through channels that persist after test-score effects fade (Chetty et al., 2011a). Conversely, negative school experiences carry lasting costs. Bullying victimization in childhood predicts worse health in adulthood (Momose and Ishida, 2024); chronic absence predicts lower adult earnings and employment (Gershenson et al., 2017; Cattan et al., 2023); and school violence is associated with adverse educational and developmental trajectories (Diagne, 2009).

We believe that education policy should not be evaluated on test scores alone. Behavioral problems such as absenteeism and violence are outcomes that parents care deeply about, arguably more immediately than marginal differences in achievement. Moreover, students' sense of belonging, perceptions of classroom disciplinary climate, and relationships with teachers are treated by the OECD as core indicators of school system performance (OECD, 2017b, 2021b).

What determines teacher–student relationships and classroom climate, however, remains largely unexplored in economics, particularly whether they respond to policy-manipulable inputs. Class size is a natural candidate: mechanically, a smaller class raises teacher attention per student and, in Lazear (2001)'s framework, reduces the probability that instruction is disrupted at any moment. Yet we are aware of no study that estimates the causal effect of class size on directly measured classroom climate and teacher–student relationships using individual-level panel data. Sugita (2025) comes closest but relies on cross-sectional data and a coarser satisfaction measure; Tanaka and Wang (2025) invoke the disruption channel but cannot observe it.

Three gaps emerge. First, evidence suggests that if smaller classes deliver benefits, they do so through channels other than test scores; yet those channels have rarely been measured directly. Second, the literature establishing the long-run importance of classroom environments has developed largely separately from the class-size literature, leaving open whether the environment responds to this canonical policy lever. Third, the leading theoretical account of class-size effects (Lazear, 2001) assigns the central role to classroom disruption, but empirical work has proxied disruption with student composition rather than observing it. This study addresses all three gaps: we estimate the causal effect of class size on directly measured classroom climate, teacher–student relationships, and behavioral outcomes, using large-scale individual-level panel data through 2024 and a Maimonides-rule identification strategy.

3 Data

3.1 Data Description

In this study, we use individual-level panel data from the Saitama Skill Assessment (hereafter, the Saitama Survey), conducted annually between 2017 and 2024. The Survey is a census covering all public elementary and junior high school students from grade 4 through grade 9 in every municipality of the prefecture, excluding the prefectural capital, with each cohort comprising approximately 300,000 students. A notable feature of the Survey is that test scores are scaled using item response theory (IRT), allowing comparisons of academic performance across years. In addition, the Survey collects student questionnaire responses on socioeconomic background (e.g., participation in private tutoring, number of books at home) and school experiences, as well as school questionnaires completed by principal or vice principal.

3.1.1 Timing and Sample Construction

The sample construction follows Ito et al. (2020), who also use the same data set[1]. The Saitama Survey is administered at the beginning of each school year, in April. Because Japanese schools reshuffle classes between school years, responses collected in April of year t primarily reflect students' experiences in year $t-1$. We therefore link the class size a student experienced in year $t-1$ to outcomes reported in year t, and index class size by t throughout for notational simplicity.

3.1.2 Outcome Variables

Student-level Outcomes

Our primary outcomes are two composite indices constructed from the student questionnaire, each serving as a proxy for the quality of the classroom environment: the Teacher–Student Relationship Index and the Classroom Climate Index. The Teacher–Student Relationship Index aggregates three items: "Did your teachers recognize your strengths and good qualities?" (*Recognized*), "Did your teachers listen to your worries and concerns?" (*Listened*), and "Did your teacher explain things you did not understand in class or on tests until you fully understood them?" (*Explained*). The Classroom Climate Index aggregates the following three items: "Did you enjoy your life in the classroom?" (*Satisfaction*), "Was your classroom a calm and focused place for learning?" (*Calm Environment*), and "Do you think your class worked together as a team in various activities (including events like sports day and field trips)?" (*Cooperation*)

We use the term classroom climate in the sense employed in the school climate literature

[1] Two groups are excluded by this design. First, grade 4 students are dropped, as no lagged class-size information exists for them. Second, following the 2021 legislative reform that progressively lowered the statutory cap to 35 students in elementary grades, the cohort observed in grade 5 in 2024 is excluded.

(e.g., Darling-Hammond and DePaoli (2020)) and in the OECD's assessment frameworks (OECD (2017a)), which encompass disciplinary climate, peer relations, and students' affective experience of the classroom. Our three items capture these dimensions in abbreviated form; they are not a validated multidimensional climate scale, and we interpret the index as a summary measure of students' experienced classroom environment rather than as a measurement of any single psychological construct.

Each item is answered on a 4-point Likert scale. Items are reverse-coded where necessary so that 4 represents the most affirmative response. Each index is the sum of its three items and then standardizes to mean 0 and standard deviation 1, so that estimated effects are interpretable in index standard deviation units. Internal consistency is high, with Cronbach's alpha of 0.748 for the Teacher–Student Relationship Index and 0.617 for the Classroom Climate Index. It is worth mentioning that Kaiser and Oswald (2022) provide evidence supporting the use of subjective ordinal measures of this kind, showing that they predict future behavior better than objective socioeconomic variables.

School-Reported Behavioral Outcomes

Our second set of outcomes is drawn from the school questionnaire, in which each school reports on behavioral conditions in each grade. We use three items asking the extent to which the following were a problem at the school: chronic student absenteeism, defined as 30 or more absent days per year (*Absenteeism*); maintaining learning discipline in the classroom (*Discipline*); and student-on-student violence or other serious misconduct (*Violence*). Each item is answered on a 4-point Likert scale by the school principal or vice principal. Items are reversed and standardized to mean 0 and standard deviation 1 so that higher values indicate better conditions: less chronic absence, better-maintained discipline, and less violence or misconduct.

Unlike the student questionnaire outcomes, these measures are reported at the school–grade–year level rather than for individual students. They nonetheless serve two purposes. First, they provide behavioral counterparts to the students' subjective assessments: absenteeism, discipline, and violence are precisely the margins that the literature reviewed in Section 2.3 identifies as consequential for long-run outcomes. Second, because they come from reports by school administrators rather than student self-reports, they allow us to verify that our findings do not depend on a single respondent type[2].

3.2 Summary Statistics

Table 1 presents descriptive statistics for the main variables used in the analysis. These statistics are based on the pooled estimation sample of students in grades 5 through 9. The average class size is 32.53, while the average predicted class size under the Maimonides rule is 33.47, with a maximum of 40, consistent with the 40-student cap.

There are student-level covariates, serving as proxies for parental socioeconomic status. Approximately 39% of students have no access to private tutoring, and 11% have no books at home. Distributions of the outcome variables are strongly skewed toward affirmative responses, with most students answering 3 or 4 on the 4-point scale. We return to this issue in Section 5.4.4 when discussing the scope for detecting treatment effects.

4 Empirical Strategy

4.1 Identification Problem and The Maimonides Rule Instrument

A naive regression of student outcomes on class size is unlikely to recover a causal effect, because class size is endogenously determined. Schools may assign struggling or disruptive students to smaller classes, allocate experienced teachers to larger ones, or parents may sort into schools partly on the basis of class size. Any of these mechanisms induces correlation between class size and unobserved determinants of student outcomes, biasing OLS estimates in ways that are difficult to sign a priori.

To address this endogeneity, we follow Angrist and Lavy (1999) and exploit the discontinuous variation in class size generated by Japan's statutory class-size cap. Under the Act on Standards for Class Formation and Staffing Levels at Public Compulsory Education

[2] An important feature of the school questionnaire is its timing. Schools report on conditions in the previous academic year, so that data collected at the beginning of grade 5 describe the cohort's classroom environment in grade 4. The school questionnaire therefore captures class-size effects one grade earlier than the student questionnaire, which is why grade 4 class-size assignments enter the school-level analysis despite being excluded from the student-level sample.

Table 1: Summary Statistics

Variables	Observations	Units	Mean	Median	SD	Min	Max	% Chg.
Teacher–Student Relationship								
Composite index	1,692,153	535,012	0.00	0.27	0.82	−3.46	0.73	
Items included in the composite index								
Recognized	1,707,538	535,605	3.51	4.00	0.68	1	4	38.5
Listened	1,702,867	535,716	3.44	4.00	0.75	1	4	41.6
Explained	1,713,881	536,147	3.47	4.00	0.72	1	4	41.8
Classroom Climate								
Composite Index	1,706,846	535,929	0.00	0.04	0.76	−3.14	0.80	
Items included in the composite index								
Satisfaction	1,719,539	536,375	3.57	4.00	0.74	1	4	36.5
Calm Environment	1,714,696	536,193	2.97	3.00	0.88	1	4	56.8
Cooperation	1,715,187	536,143	3.55	4.00	0.69	1	4	41.0
School Reported Outcomes								
Absenteeism (reversed)	28,485	1,081	2.66	3.00	1.12	1	4	
Discipline (reversed)	28,479	1,081	3.38	4.00	0.82	1	4	
Violence (reversed)	28,488	1,081	3.17	3.00	0.88	1	4	
Variable of Interest								
Class Size	1,727,691	536,848	32.53	33.00	4.67	1	43	
Instrumental Variable								
Predicted Class Size (instrument)	1,727,691	536,848	33.47	34.25	4.74	1	40	
Covariates								
Grade Enrollment	1,727,691	536,848	117.20	107.00	58.66	1	362	
School Enrollment	1,727,691	536,848	348.67	319.00	172.66	1	988	
Gender (female=1)	1,727,691	536,848	0.49	0.00	0.50	0	1	
Private Tutoring (No tutoring=1)	1,703,319	536,031	0.39	0.00	0.49	0	1	
Books at Home (No books at home=1)	1,717,855	536,587	0.11	0.00	0.31	0	1	

Note:
Observations column shows number of student-year observations; Units shows number of unique students. For school-level outcomes (Absenteeism, Discipline, Violence), observations are at the school×grade×year level and Units shows number of unique schools. The composite indices shown in Table are computed as the row-mean of their constituent raw (non-standardized) items on the original 1–4 scale, and are presented for descriptive purposes only. For estimation, some of each item is standardized.

Schools, the maximum class size is 40 students[3]. A grade cohort of 40 students therefore forms a single class, while a cohort of 41 must be split into two classes of 20 and 21. More generally, predicted class size, PS, is a sawtooth function of grade enrollment:

$$PS_{ts} = \frac{GS_{ts}}{\text{int}\left(\frac{GS_{ts}-1}{40}\right) + 1} \tag{1}$$

where t, s, and g index year, school, and grade; GS is the grade enrollment; and `int()` returns the greatest integer less than or equal to the given real number (i.e., the floor function). Because the discontinuous drops in PS at multiples of 40 are driven by small, arguably idiosyncratic fluctuations in enrollment, class size in the neighborhood of these thresholds can be treated as exogenous, conditional on a smooth function of enrollment. Since PS is positively correlated with actual class size, Angrist and Lavy (1999) treated PS as an instrumental variable and conducted two-stage least squares (2SLS) estimation to identify the causal effect of class size reduction.

Figure 1 plots predicted and actual class size against grade enrollment. Actual class size tracks the sawtooth pattern of the Maimonides rule closely, while also rising with enrollment overall—which is why all specifications control for grade enrollment directly.

4.2 Empirical Specifications

Our student-level analysis estimates the following system by two-stage least squares (2SLS):

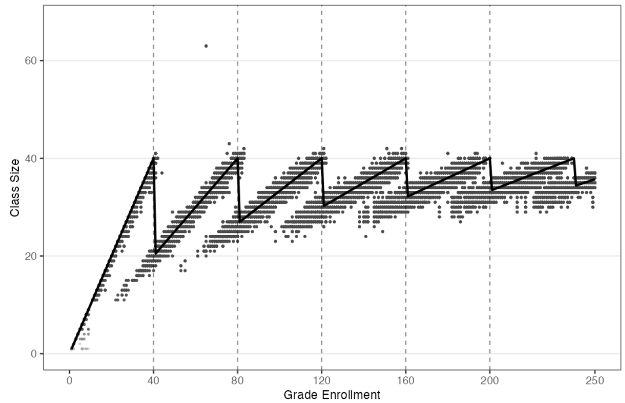
$$CS_{itsg} = \theta_0 + \theta_1 PS_{tsg} + \theta_2 GS_{tsg} + \theta_3 GS^2_{tsg} + \boldsymbol{\pi}^T \boldsymbol{X}_{itsg} + \varepsilon_{itsg} \tag{2}$$

$$Y_{itsg} = \delta_0 + \delta_1 \widehat{CS}_{itsg} + \delta_2 GS_{tsg} + \delta_3 GS^2_{tsg} + \boldsymbol{\gamma}^T \boldsymbol{X}_{itsg} + \omega_{itsg} \tag{3}$$

where i indexes students, Y is variables on the quality of classroom environment measured by a composite index of teacher–student relationships or classroom climate, CS is actual class size, and $\boldsymbol{X}$ is a vector of controls including gender, variables on parental socio-economic status measured by the access to private tutoring and the availability of books at home, and grade and year fixed effects. We estimate three specifications of increasing stringency: (1)

[3] Since 2011, the cap for first grade has been 35 students; a 2021 reform progressively extended the 35-student cap to higher elementary grades. The affected cohort is excluded from the sample.

Figure 1: Relationship between Grade Enrollment and Class Size



Note:
The vertical axis shows the number of students at a given class, and the horizontal axis shows grade enrollment. The solid line represents the predicted class size based on Maimonides' rule. The plotted points indicate the actual observed class sizes.

controlling only for student-level covariates (baseline); (2) adding school fixed effects, which absorb time-invariant school characteristics; and (3) adding student fixed effects, which absorb all time-invariant individual heterogeneity, so that identification comes solely from within-student changes in class size across grades.

However, as documented in Appendix A.2 (Figure 5), within-student variation in class size is limited because students typically remain in the same school. The median within-student standard deviation of class size is only 1.50 students, so the student-FE estimates are identified from relatively modest variation. We therefore treat the school-FE specification as our benchmark for interpretation.

4.3 Validity of the Instrument

Instrument validity requires relevance and exogeneity. Relevance is directly testable: as shown in Section 5.1, predicted class size is a strong predictor of actual class size in all specifications, with first-stage effective F statistics exceeding both the conventional threshold of Stock and Yogo (2005) and the more stringent heteroskedasticity-robust threshold of Olea and Pflueger (2013).

Exogeneity requires that, conditional on the smooth enrollment controls, predicted class size affects outcomes only through actual class size. The leading threat is manipulation of enrollment around the thresholds. For example, schools or parents strategically placing cohort sizes just above a multiple of 40 to secure smaller classes. Three pieces of evidence weigh against this concern. First, McCrary-style density tests Cattaneo et al. (2018) detect no bunching in grade enrollment at any of the four thresholds (Appendix A.5). Second, predetermined student characteristics, such as gender composition, parental socio-economic status, are smooth through the thresholds (Appendix A.5). Third, schools near the thresholds do not differ systematically from other schools in observable composition (Appendix A.6), suggesting the complier population is not a selected group, although the 2SLS estimates remain local to medium-sized schools near the 40-, 80-, 120-, and 160-student cutoffs.

5 Results

5.1 First Stage

Table 2 presents the first-stage estimate of equation (2) under three specifications. In all specifications, the predicted class size is a strong predictor of actual class size throughout: the coefficient is 0.776 with covariates only, 0.635 with school fixed effects, and 0.479 with school and student fixed effects, all precisely estimated. Effective F statistics exceed both the Stock and Yogo (2005) threshold and the more stringent heteroskedasticity-robust threshold of Olea and Pflueger (2013) in all specifications. We conclude that the instrument is strong and proceed to 2SLS estimation.

5.2 Main Results

5.2.1 Effects on Composite Indices for Teacher–Student Relationships and Classroom Climate

Table 3 presents the 2SLS estimates of the effect of class size on the Teacher–Student Relationship Index and the Classroom Climate Index. The class-size coefficient is negative and significant at the 1% level in all six columns, and remarkably stable across specifications: adding school fixed effects, and then student fixed effects, leaves the point estimates es-

Table 2: First Stage

	(1)	(2)	(3)
Predicted Class Size	0.779***	0.630***	0.479***
	(0.008)	(0.012)	(0.015)
Num.Obs.	1 696 247	1 696 247	1 696 247
R^2	0.626	0.689	0.839
F-stat (Predicted Class Size)	8714.8***	2867.6***	1003.8***

Note: All specifications include grade and year fixed effects. Column (1) includes no school or student fixed effects. Column (2) adds school fixed effects. Column (3) further adds student fixed effects.
All specifications include the following student-level covariates: gender (female=1), no private tutoring (=1), and no books at home (=1). In column (3), time-invariant covariates are absorbed by student fixed effects.
The instrument (Predicted Class Size) is derived from the Maimonides Rule: $\lfloor (GS-1)/40 \rfloor + 1$, where GS denotes grade enrollment.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

sentially unchanged at -0.003 to -0.004. This stability suggests that the instrument is not merely picking up school- or student-level sorting. In our benchmark school-FE specification, a 10-student reduction in class size improves both indices by 0.03–0.04 SD.

With respect to teacher–student relationships, smaller classes may increase teacher attention per student through reduced teacher workload; Hojo (2021) documents that larger classes are associated with longer working hours and greater workload stress among Japanese teachers, suggesting that teachers in smaller classes may simply have more capacity to attend to each student, to notice their difficulties, and to respond to their concerns.

With respect to classroom climate, the framework of Lazear (2001) models education as a joint production process vulnerable to disruption: a single misbehaving student interrupts learning for the entire class, and the probability that instruction is disrupted at any moment rises mechanically with the number of students. Smaller classes therefore reduce the incidence of disruption directly, and the greater per-student attention noted above may reduce it further by allowing teachers to detect and defuse problem behavior early. If the disruption channel is operative, it should be clearer in school-reported outcomes examined in Section 5.2.2.

The covariates behave as expected: girls report better relationships and climate, while students without access to private tutoring and, especially, students with no books at home report substantially worse experiences. The no-books gap of roughly 0.06–0.09 SD exceeds the effect of a 10-student class-size reduction, a useful benchmark for the magnitude of the treatment effect. Appendix A.1 reports estimates for the six individual items. Effects are significant for all items, with the largest for *Calm Environment* (-0.006 to -0.008), the item that also exhibits the greatest year-to-year variation (Section 5.4.4).

Table 3: Class-size Effects on Composite Indices

Variable	Baseline		+ School FE		+ School FE + Student FE	
	(1) Teacher–Student Relationship	(2) Classroom Climate	(3) Teacher–Student Relationship	(4) Classroom Climate	(5) Teacher–Student Relationship	(6) Classroom Climate
Class Size	-0.005***	-0.004***	-0.003***	-0.004***	-0.004***	-0.005***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Grade Enrollment	0.000	0.000*	0.000*	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment^2	-0.000	-0.000**	-0.000*	-0.000**	-0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.093***	0.059***	0.093***	0.059***	0.046***	0.060***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.014)	(0.013)
Private Tutoring	-0.013***	-0.041***	-0.013***	-0.041***	-0.004***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Books at Home	-0.095***	-0.067***	-0.094***	-0.066***	-0.030***	-0.028***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	1,663,925	1,678,450	1,663,925	1,678,450	1,663,925	1,678,450

Note: Teacher–Student Index is the row-mean of *Recognized*, *Listened*, and *Explained*, and Classroom Index is the row-mean of *Satisfaction*, *Calm Environment*, and *Cooperation*. All constituent outcomes are standardized. Grade and year fixed effects included in all specifications.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5.2.2 Effects on School-Reported Behavioral Outcomes

Table 4 presents the 2SLS estimates for the school-reported outcomes. A 10-student reduction improves absenteeism by 0.08 SD (significant at 1%), discipline by 0.26 SD, and violence by 0.15 SD (all significant at 1%). Because these outcomes are measured at the school–grade–year level, the estimates are subject to the aggregation bias documented by Hanushek et al. (1996) and should be read as upper bounds; indeed, when we aggregate the student questionnaire outcomes to the school level (Appendix A.3), estimated effects grow relative to their individual-level counterparts, consistent with this concern.

The value of these results lies less in their magnitudes than in their source: they come from school reports rather than student self-reports. Students assessing their own experience and school administrators reporting behavioral conditions are independent respondent groups, and both register improvements. This consistency strengthens the case that class-size reduction changes the classroom environment in ways observable from multiple perspectives.

Table 4: Class-size Effects on School-Reported Behavioral Outcomes

Variable	(1) Absenteeism	(2) Discipline	(3) Violence
Class Size	-0.008***	-0.026***	-0.015***
	(0.003)	(0.003)	(0.003)
Grade Enrollment	-0.000	-0.003	-0.005***
	(0.002)	(0.002)	(0.002)
Grade Enrollment^2	-0.000	0.000	0.000**
	(0.000)	(0.000)	(0.000)
% of Female	0.225**	0.971***	0.538***
	(0.097)	(0.106)	(0.097)
% of No Books at Home	-0.283**	-0.356**	-0.234*
	(0.135)	(0.143)	(0.128)
% of No Private Tutoring	-0.016	-0.088	-0.102
	(0.074)	(0.080)	(0.076)
School FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	25,493	25,491	25,496

Note: School-grade-year level 2SLS estimation. Outcomes are school questionnaire measures of absenteeism, classroom discipline, and violence/trouble, reversed and standardized (higher = better).
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5.3 Contrast with Academic Achievement

To place these findings in context, Table 5 estimates the same specifications with standardized reading and mathematics scores as outcomes. Effects on achievement are small (approximately -0.002 for both reading and mathematics in the school-FE specification) and vanish once student fixed effects are included. This result replicates Ito et al. (2020) who used the same Survey, even though our data extend the observation window through 2024, and mirrors the broader Japanese literature. The contrast is the central message of this paper: class-size reduction leaves test scores essentially unchanged while producing detectable improvements in the classroom environment.

Table 5: Class-size Effects on Academic Achievement

	(1)	(2)	(3)	(4)
	School FE		School FE + Student FE	
Variable	Reading (Std.)	Math (Std.)	Reading (Std.)	Math (Std.)
Class Size	-0.002***	-0.002***	-0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Grade Enrollment	0.000	-0.000	-0.000	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment^2	-0.000	0.000	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.258***	-0.041***	0.085***	0.060***
	(0.003)	(0.002)	(0.017)	(0.016)
Private Tutoring	-0.011***	-0.175***	-0.017***	-0.076***
	(0.003)	(0.005)	(0.001)	(0.002)
Books at Home	-0.514***	-0.447***	-0.019***	-0.011***
	(0.004)	(0.004)	(0.002)	(0.002)
Observations	1,691,901	1,684,181	1,691,901	1,684,181

Note: Outcomes are standardized within year-grade cells. Grade 5 observations in 2024 are excluded because the 35-student class-size rule reform invalidates the Maimonides Rule instrument for that cohort.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5.3.1 Mechanisms: Does Prior Classroom Environment Moderate the Class-Size Effect?

The leading theoretical account of class-size effects, due to Lazear (2001), assigns the central role to classroom disruption: in larger classes, the probability that instruction is interrupted at any moment is higher, and the returns to reducing class size should therefore be greatest where disruption is most prevalent. Tanaka and Wang (2025) find support for this framework in Japanese data, showing that class-size effects on achievement vary with the ability composition of classmates, which they interpret as a proxy for the probability of disruption. Our data allow a more direct test: rather than proxying the classroom environment with peer characteristics, we observe it.

We estimate specifications that interact class size with the lagged classroom environment. The moderators are the school-grade-year means of the Teacher–Student Relationship Index and the Classroom Climate Index, measured in the survey administered one year before the outcome. Because the survey conducted in April of year t−1 reflects students' experience in year t−2, the moderators capture the state of the classroom environment before exposure to the class size we study, and are thus predetermined with respect to treatment. Both moderators are standardized to mean 0 and standard deviation 1. Class size and its interactions are instrumented by predicted class size and the corresponding interactions, following Tanaka and Wang (2025), and all specifications include school, grade, and year fixed effects. Because the moderators require a lagged survey wave, the estimation sample is smaller than in the main analysis.

Table 6 reports the results for the two composite indices and for standardized reading and mathematics scores. With standardized moderators, the coefficient on class size itself measures the effect in a school-grade cell with average prior conditions; it remains negative and significant for both climate outcomes, consistent with the main results.

Two findings emerge. First, there is evidence suggesting that the effect of class-size reduction on teacher–student relationships is larger where prior relationships were poor. The interaction between class size and the lagged Teacher–Student Relationship Index is positive and significant for the Teacher–Student Index (0.0013, $p < 0.01$); the corresponding coefficient for the Classroom Index points in the same direction but does not reach conventional significance (0.0006). Since the uninteracted class-size coefficient is negative, a positive interaction implies that the beneficial effect of smaller classes on teacher–student relationships shrinks as prior relationships improve, and is correspondingly larger where prior relationships were weak. In a cell whose prior Teacher–Student Index is one standard deviation below the mean, the implied effect of a 10-student reduction on the Teacher–Student Index is 0.032 SD ($=10 \times (0.0019 + 0.0013)$), compared with 0.006 SD ($=10 \times (0.0019 - 0.0013)$) in a cell one standard deviation above the mean. This pattern is suggestive of class-size effects being larger where the prior classroom environment was under strain, consistent with the heterogeneity by poverty rate documented in Section 5.4.2.

Second, there is suggestive evidence that the achievement effect, which is null on average, is not null everywhere. The interaction between class size and the lagged Classroom Climate Index is positive for both reading (0.0006, $p < 0.10$) and mathematics (0.0007, $p < 0.10$), implying that class-size reduction improves achievement more in cells where the prior classroom climate was poor. This is precisely the pattern the Lazear framework predicts:

where instruction is more frequently disrupted, reducing the number of students yields larger gains in learning. We emphasize that these interactions are significant only at the 10 percent level and involve multiple hypothesis tests, and we therefore treat them as suggestive rather than conclusive. Read alongside the first finding, however, they point to a coherent account. Class-size reduction operates on the classroom environment, and its benefits are largest where the prior environment was weakest. In those same settings, there are hints of achievement gains that the average estimate conceals.

5.4 Heterogeneity

5.4.1 Pre- vs. Post-COVID Comparison

Table 7 splits the sample into pre-COVID ($\leq$2020) and post-COVID ($\geq$2021) subperiods and reports 2SLS estimates for both composite indices across the two subperiods. The estimates are broadly consistent across both periods. Under our benchmark school FE specification, a 10-student reduction improves the teacher-student relationship index by 0.04 SD and the classroom climate index by 0.04 SD in the pre-COVID period, and by 0.04 and 0.05 SD respectively in the post-COVID period. These results suggest that the main findings are not driven by any particular period and are stable over the extended sample period from 2017 to 2024.

Table 6: Heterogeneous Effects of Class Size by Prior Classroom Environment (School-Grade-Year Mean)

Variable	Classroom Environment		Academic Achievements	
	(1) Teacher–Student Relationship	(2) Classroom Climate	(3) Reading (SD)	(4) Math (SD)
Class Size (C)	-0.0019***	-0.0029***	-0.0013*	-0.0024***
	(0.0005)	(0.0006)	(0.0007)	(0.0008)
$C\times$ Teacher–Student$_{t-1}$	0.0013***	0.0006	-0.0005	-0.0008*
	(0.0003)	(0.0004)	(0.0004)	(0.0005)
$C\times$ Classroom$_{t-1}$	0.0001	0.0005	0.0006*	0.0007*
	(0.0003)	(0.0004)	(0.0004)	(0.0004)
Teacher–Student Index$_{t-1}$	0.0149	-0.0117	0.0269**	0.0342**
	(0.0109)	(0.0127)	(0.0130)	(0.0145)
Classroom Index$_{t-1}$	0.0028	0.0478***	-0.0125	-0.0065
	(0.0096)	(0.0132)	(0.0116)	(0.0129)
Grade Enrollment	0.0001	-0.0003	0.0003	-0.0008*
	(0.0002)	(0.0003)	(0.0004)	(0.0004)
Grade Enrollment^2	-0.0000	-0.0000	-0.0000	0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Gender (female=1)	0.0855***	0.0456***	0.2512***	-0.0518***
	(0.0018)	(0.0018)	(0.0030)	(0.0028)
Private Tutoring	-0.0108***	-0.0383***	-0.0277***	-0.2031***
	(0.0014)	(0.0014)	(0.0035)	(0.0048)
Books at Home	-0.0867***	-0.0601***	-0.5159***	-0.4471***
	(0.0024)	(0.0025)	(0.0046)	(0.0046)
Observations	858,466	865,861	873,732	870,452
F-stat. of the 1st stage				
Class Size (C)	677***	675***	676***	674***
$C\times$ Teacher–Student$_{t-1}$	1076***	1080***	1073***	1071***
$C\times$ Classroom$_{t-1}$	1299***	1285***	1291***	1290***

Note: Moderators (Teacher–Student Relationship Index$_{t-1}$, Classroom Climate Index$_{t-1}$) are standardized (mean 0, SD 1). Class Size and its interactions are instrumented by Predicted Class Size and its interactions. School, grade, and year fixed effects included. The sample is smaller than in the main analysis (Table 3) because constructing the lagged moderators requires data from the preceding survey wave, causing the earliest cohort in the panel to be excluded.

Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 7: Class-Size Effects on Composite Indices (Pre- vs. Post-COVID 19)

Variable	Pre-COVID (2017–2019)				Post-COVID (2020–2024)			
	School FE		School+Student FE		School FE		School+Student FE	
	(1) Teacher–Student Relationship	(2) Classroom Climate	(3) Teacher–Student Relationship	(4) Classroom Climate	(5) Teacher–Student Relationship	(6) Classroom Climate	(7) Teacher–Student Relationship	(8) Classroom Climate
Class Size	-0.004***	-0.004***	-0.004***	-0.005***	-0.004***	-0.005***	-0.005***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Grade Enrollment	-0.000	-0.001	0.000	-0.001	0.000	0.000	0.001**	0.001
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
Grade Enrollment^2	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.109***	0.075***	-0.020	-0.009	0.076***	0.040***	0.075***	0.086***
	(0.002)	(0.002)	(0.024)	(0.024)	(0.002)	(0.002)	(0.020)	(0.020)
Private Tutoring	-0.009***	-0.039***	-0.004*	-0.006***	-0.017***	-0.042***	-0.002	-0.007***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Books at Home	-0.104***	-0.075***	-0.030***	-0.026***	-0.080***	-0.053***	-0.024***	-0.021***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Observations	877,976	882,088	877,976	882,088	785,949	796,362	785,949	796,362

Note: Teacher–Student Index is the row-mean of *Recognized*, *Listened*, and *Explained*, and Classroom Index is the row-mean of *Satisfaction*, *Calm Environment*, and *Cooperation* (see Table 1). All constituent outcomes are standardized. Grade and year fixed effects are included.

Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 8 reports the corresponding results for school-reported behavioral outcomes. The effect on classroom discipline is statistically significant in both periods. The post-COVID estimates for discipline and violence tend to be larger in magnitude than the pre-COVID estimates, possibly reflecting heightened behavioral challenges after pandemic-related disruptions to schooling.

Table 8: Class-size Effects on School-Reported Behavioral Outcomes (Pre- vs. Post-COVID 19)

	Pre-COVID (2017–2019)			Post-COVID (2020–2024)		
Variable	(1) Absenteeism	(2) Discipline	(3) Violence	(4) Absenteeism	(5) Discipline	(6) Violence
Class Size	-0.006	-0.023***	-0.010**	-0.008*	-0.030***	-0.020***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Grade Enrollment	0.000	-0.004	-0.006**	-0.001	-0.003	-0.005*
	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Grade Enrollment^2	-0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% of Female	0.331**	0.880***	0.592***	0.219	1.065***	0.531***
	(0.129)	(0.136)	(0.134)	(0.141)	(0.152)	(0.141)
% of No Books at Home	-0.225	-0.406**	-0.305*	-0.215	-0.062	-0.063
	(0.171)	(0.182)	(0.174)	(0.201)	(0.194)	(0.189)
% of No Private Tutoring	0.123	0.203*	-0.122	-0.099	-0.270**	-0.160
	(0.103)	(0.104)	(0.100)	(0.104)	(0.113)	(0.109)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,516	12,516	12,516	12,977	12,975	12,980

Note: School-grade-year level 2SLS estimation. Outcomes are school questionnaire measures of absenteeism, classroom discipline, and violence/trouble, reversed and standardized (higher = better). School, grade, and year fixed effects included.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5.4.2 Heterogeneity by Poverty

Table 9 splits schools at the median poverty rate (12.2%), measured as the share of students receiving public financial assistance and shows the results from benchmark school FE specification. In high-poverty schools, a 10-student reduction improves both the teacher–student index and the classroom index by 0.04 SD, with effects significant at the 1% level. In schools with low educational aid rates, the estimated effects are smaller at 0.02 SD for both indices, though still statistically significant. This pattern suggests that class-size reduction may yield larger benefits in socioeconomically disadvantaged schools, though the difference in magnitudes between the two groups is not statistically tested here.

Table 9: Heterogeneity by School-level Poverty Rate

Variable	High Poverty Rate (≥12.2%)		Low Poverty Rate (<12.2%)	
	(1) Teacher–Student Relationship	(2) Classroom Climate	(3) Teacher–Student Relationship	(4) Classroom Climate
Class Size	-0.004***	-0.004***	-0.002***	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)
Grade Enrollment	0.001**	0.001*	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment^2	-0.000***	-0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.086***	0.052***	0.094***	0.057***
	(0.002)	(0.002)	(0.002)	(0.002)
Private Tutoring	-0.014***	-0.041***	-0.012***	-0.040***
	(0.002)	(0.002)	(0.002)	(0.002)
Books at Home	-0.093***	-0.059***	-0.091***	-0.070***
	(0.003)	(0.003)	(0.003)	(0.003)
Observations	1,042,805	1,051,539	1,032,486	1,041,825

Note: Sample split by median poverty rate (= 12.2%), measured as the share of students receiving public financial assistance. Teacher–Student Relationship is the row-mean of *Recognition*, *Listened*, and *Explained*, and Classroom Climate is the row-mean of *Satisfaction*, *Calm Environment*, and *Cooperation* (see Table 1). All constituent outcomes are standardized. Grade and year fixed effects included. Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5.4.3 Heterogeneity by Grade

Table 10 presents estimates by grade for the composite indices. Effects are significant in grades 5–7 (-0.004 to -0.005) but small and insignificant in grades 8–9. Table 11 presents results for school-reported behavioral outcomes. The effects on discipline and violence are significant in Grades 4, 5, and 6 (all elementary school grades), while Grade 7 estimates are small and insignificant. Grade 8, however, shows significant effects for both discipline (0.41 SD) and violence (0.39 SD) per 10-student reduction. Taken together, the school-level results are strongest in elementary grades, weakest at the transition to junior high school, and partially re-emerge in grade 8, a pattern we do not over-interpret given the smaller sample size[4]. Overall, the evidence indicates that class-size reduction matters most in elementary school, consistent with the concentration of effects in early grades found in the experimental literature (Krueger, 1999; Hanushek, 1999).

5.4.4 Ceiling Effects and Outcome Variation

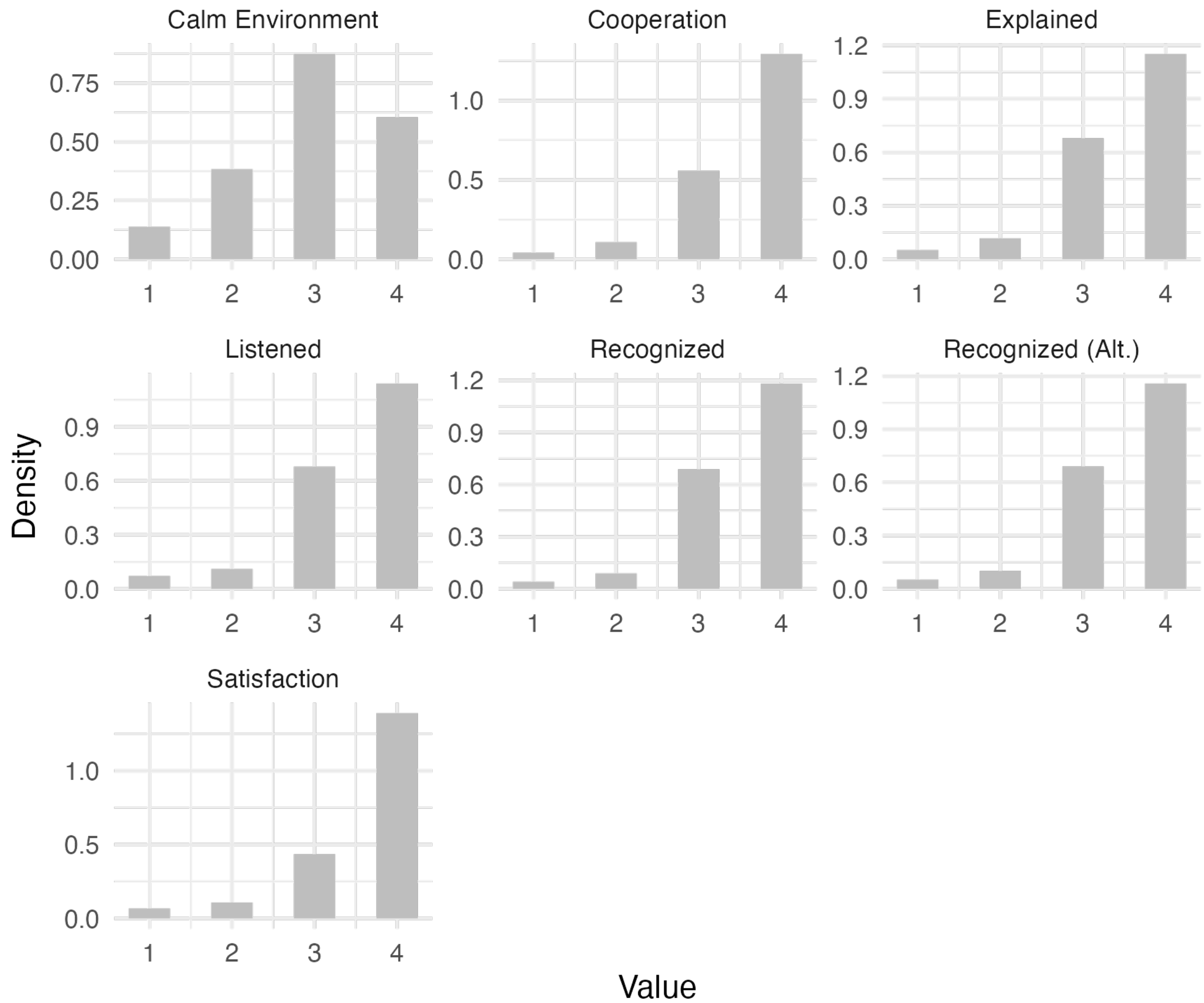
Two features of the outcome data bear on the interpretation of effect sizes. First, as shown in Figure 2, the distributions are heavily skewed toward affirmative responses, suggesting that many students responded with either a 3 or 4 on the 4-point Likert scale. Second, as shown in Figure 4a, year-to-year variation is modest. For five of the six items, roughly 40% of students change their response across years, with *Calm Environment* exception at over 60%, and also the item with the largest estimated effect. Both features imply that our estimates likely understate the latent improvement in classroom environments: ceiling effects compress the observable response to treatment. The 0.04 SD estimates should be read with this attenuation in mind.

5.4.5 Cost-effectiveness

Are these effects worth their cost? Reducing class size by 10 students requires roughly 135,000 JPY per student per year at a baseline class size of 30, based on an average teacher

[4]Note that grade 4 is included in Table 11 because the school questionnaire records outcomes for the previous academic year's cohort; grade 4 class-size assignments are therefore matched to school-level outcomes reported in the following year's survey.

Figure 2: Distribution of Outcomes



Notes: Distribution of student responses on the four-point Likert scale (1 = Never, 2 = Rarely, 3 = Often, 4 = Always) for each of the six student questionnaire items. Items are grouped by composite index. **Teacher–Student Relationship Index**: *Recognized* (Did your teachers recognize your strengths and good qualities?), *Listened* (Did your teachers listen to your worries and concerns?), *Explained* (Did your teacher explain things you did not understand until you fully understood?). **Classroom Climate Index**: *Satisfaction* (Did you enjoy your life in the classroom?), *Calm Environment* (Was your classroom a calm and focused place for learning?), *Cooperation* (Did your class work together as a team?). Responses are pooled across Grades 4–9. Data from the Saitama Survey, 2017–2024.

Figure 3: Outcomes by Grade

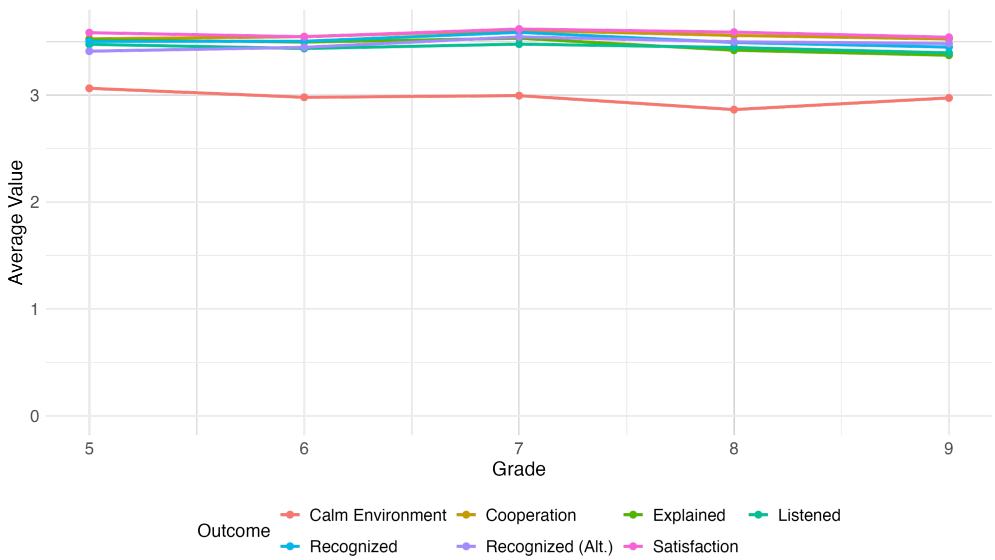
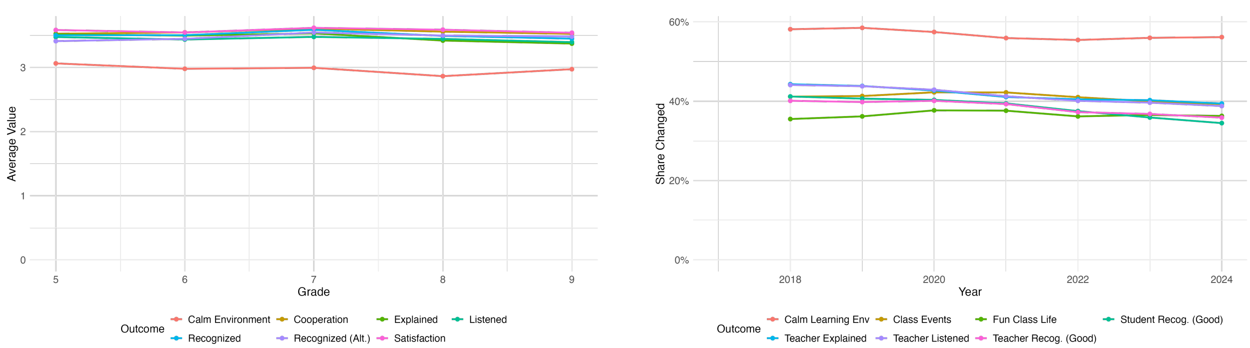


Figure 4: Variation of Outcomes



(a) Variation by Grade

(b) Share of Students with Outcome Change by Year

Notes: Each line represents one of the six student questionnaire outcome items. Items in the **Teacher–Student Relationship Index** are *Recognized*, *Listened*, and *Explained*; items in the **Classroom Climate Index** are *Satisfaction*, *Calm Environment*, and *Cooperation*. Panel (a) shows, for each grade (Grades 4–9), the share of students whose response to the item changed relative to the previous school year, illustrating the extent of within-student year-to-year variation. Panel (b) shows the same share by survey year; the 2015 cohort is excluded because no prior-year response is available to compute year-to-year changes for that cohort. Data from the Saitama Survey, 2017–2024.

Table 10: Heterogeneity by Grade (Student-level Composite Indices)

Grade	(1) Teacher–Student Relationship	(2) Classroom Climate
Grade 5	-0.004***	-0.004***
	(0.001)	(0.001)
Grade 6	-0.004***	-0.004***
	(0.001)	(0.001)
Grade 7	-0.004***	-0.005***
	(0.001)	(0.001)
Grade 8	0.000	-0.001
	(0.002)	(0.003)
Grade 9	-0.001	-0.004*
	(0.002)	(0.003)
N (grade 5)	293,526	296,734
N (grade 6)	350,836	354,695
N (grade 7)	338,664	341,123
N (grade 8)	339,577	342,534
N (grade 9)	341,374	343,414

Note: Each cell reports the class-size coefficient from a separate 2SLS regression estimated on the indicated grade subsample. Teacher–Student Relationship is the row-mean of *Recognized*, *Listened*, and *Explained*, and Classroom Climate is the row-mean of *Satisfaction*, *Calm Environment*, and *Cooperation* (see Table 1). All constituent outcomes are standardized. School and year fixed effects included. The Grade 5 sample excludes 2024 because the 35-student class-size rule reform invalidates the Maimonides Rule instrument for that cohort.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

employment cost of 8.1 million JPY (MEXT, 2019). A back-of-the-envelope calculation that monetizes the environmental improvements through established wage returns to non-cognitive skills yields benefit-cost ratios above one under central parameter values (e.g., 1.52 at a 3% discount rate), and consistently above the corresponding ratios computed through the achievement channel. Appendix A.9 details the assumptions and sensitivity analysis; the results are sensitive to the assumed persistence of effects and to the mapping

Table 11: Heterogeneity by Grade (School-Reported Behavioral Outcomes)

Grade	(1) Absenteeism	(2) Discipline	(3) Violence
Grade 4	-0.005	-0.038***	-0.027***
	(0.005)	(0.005)	(0.005)
Grade 5	-0.006	-0.028***	-0.015***
	(0.005)	(0.005)	(0.005)
Grade 6	-0.008*	-0.024***	-0.015***
	(0.005)	(0.005)	(0.005)
Grade 7	0.012	-0.001	-0.010
	(0.014)	(0.014)	(0.014)
Grade 8	-0.014	-0.041***	-0.039**
	(0.013)	(0.015)	(0.016)
N (grade 4)	3,925	3,928	3,926
N (grade 5)	6,104	6,105	6,110
N (grade 6)	6,377	6,375	6,373
N (grade 7)	2,375	2,372	2,376
N (grade 8)	3,888	3,887	3,887

Note: School-year level 2SLS estimation by grade. Grade labels indicate the grade in which class size was assigned and outcomes were measured. School questionnaire outcomes reflect conditions in the same academic year as class size assignment, as schools report on the previous year's cohort in the following year's survey. Outcomes reversed and standardized (higher = better). School and year fixed effects included. Covariates: enrollment, enrollment^2, proportion female, proportion with no books, and proportion with no private tutoring.
Standard errors clustered at the school level in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

from classroom environment measures to non-cognitive skills, and the two channels should not be summed, as they are unlikely to be independent. Cost-effectiveness comparisons with alternative policies also require institutional context. Per million JPY of per-student expenditure, class-size reduction yields roughly half the improvement attainable through a one-SD gain in teacher value-added (Ito and Tabata, 2022). In Japan, however, public school teachers are recruited and assigned by prefectural boards of education; schools and school districts have virtually no discretion over teacher selection. Improving teacher value-added is therefore not a policy lever available at the school or local level, whereas class size is set by national staffing standards and can be adjusted through national or prefectural decisions.

Within the Japanese institutional environment, class-size reduction is one of the few available instruments for improving the classroom environment at scale—a point of direct relevance to the ongoing rollout of the 35-student cap.

6 Conclusion

This paper has examined the effect of class-size reduction on classroom climate, teacher–student relationships, and school reported behavioral outcomes using individual-level panel data from Japan. Leveraging an instrumental variable approach based on the Maimonides rule, which stipulates a maximum class size of 40 students, we find that a 10-student reduction in class size improves composite indices for classroom climate and teacher–student relationships by 0.03–0.04 standard deviations (SD), and substantially reduces school-reported behavioral problems, with effects of 0.26 SD for classroom discipline and 0.15 SD for violence, though these aggregated estimates should be read as upper bounds (Hanushek et al., 1996). These results are robust across the pre- and post-COVID subperiods, concentrated in elementary grades, and larger in schools serving more disadvantaged students. In the same data and design, effects on test scores are essentially zero, replicating prior Japanese findings and consistent with recent international evidence that achievement effects have faded (Angrist et al., 2019). The benefits of smaller classes, where they exist, appear to run through the classroom environment rather than through measured achievement. As shown in A.8, a back-of-the-envelope calculation suggests that these environmental improvements alone are large enough to justify the cost of class-size reduction under central parameter values.

These findings carry a distinct institutional implication. Per unit of expenditure, improving teacher quality may deliver larger gains than reducing class size (Ito and Tabata, 2022). In Japan, schools and school districts have virtually no discretion over teacher selection. Improving teacher value-added through selective recruitment is therefore not a policy lever available at the school or local government level, whereas class size is set by national staffing standards and can be adjusted through national or prefectural decisions. Given this institutional constraint, class-size reduction is one of the few available policy instruments for improving quality of classroom environment at scale, a point of direct relevance to the ongoing rollout of the 35-student cap.

Two caveats on external validity apply. The 2SLS estimate identified by the Maimonides rule applies specifically to schools near the enrollment cutoffs (40, 80, 120, 160 students per grade). As shown in Appendix A.6, schools near the cutoffs do not differ systematically from other schools in terms of student characteristics, suggesting that the compliers are not a particularly selected group. Nevertheless, the 2SLS estimate is most directly applicable to medium-sized schools, and generalization to very small or very large schools should be done with caution. In addition, the estimates identify the effect of approximately one school year of exposure to a smaller class; they do not speak to the cumulative effect of sustained exposure across grades.

Several limitations remain. The observation window covers approximately one academic year per student, and whether the improvements in classroom climate and teacher–student relationships documented here translate into long-term educational attainment or labor market outcomes is an open question. Future research combining class-size data with longer-term follow-up would help assess the durability and downstream consequences of these effects. In addition, the school questionnaire outcomes analyzed in Section 5.2.2 are measured at the school–grade–year level. As Hanushek et al. (1996) caution, estimates based on aggregated data tend to be upward biased relative to those from individual-level data; accordingly, the reported effects of 0.26 SD for classroom discipline and 0.15 SD for violence should be interpreted as upper bounds rather than precise point estimates.

As Japan proceeds with the largest class-size reform in decades, evaluations that look only at test scores risk concluding that the policy has failed. Our results suggest a different standard of evaluation: the reform's effects on the daily environment in which children spend their school lives.

References

Akabayashi, H. and R. Nakamura (2014) "Can Small Class Policy Close the Gap? An Empirical Analysis of Class Size Effects in Japan", *Japanese Economic Review*, **65** (3), pp. 253–281.

Angrist, J. D. and V. Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement", *The Quarterly Journal of Economics*, **114** (2), pp. 533–575.

——— , 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** (4), pp. 216–249.

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

Ansari, A., T. L. Hofkens, and R. C. Pianta (2020) "Teacher–Student Relationships across the First Seven Years of Education and Adolescent Outcomes", *Journal of Applied Developmental Psychology*, **71**, p. 101200.

Asakawa, S., M. Abe, F. Ohtake, and S. Sano (2026) "Class Size and Household Educational Investment Responses: Evidence from Amagasaki City Data", RIETI Discussion Paper Series 26-J-022, Research Institute of Economy, Trade and Industry (RIETI), pp. 1–35, In Japanese.

Carrell, S. E., M. Hoekstra, and E. Kuka (2018) "The Long-Run Effects of Disruptive Peers", *American Economic Review*, **108** (11), pp. 3377–3415.

Cattan, S., D. A. Kamhöfer, M. Karlsson, and T. Nilsson (2023) "The Long-Term Effects of Student Absence: Evidence from Sweden", *The Economic Journal*, **133** (650), pp. 888–903.

Cattaneo, M. D., M. Jansson, and X. Ma (2018) "Manipulation Testing Based on Density Discontinuity", *The Stata Journal*, **18** (1), pp. 234–261.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011a) "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR", *The Quarterly Journal of Economics*, **126** (4), pp. 1593–1660.

——— , ——— , ——— , ——— , ——— , and ——— (2011b) “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR”, *The Quarterly Journal of Economics*, **126** (4), pp. 1593–1660.

Coleman, J. S. (1968) “Equality of Educational Opportunity”, *Integrated education*, **6** (5), pp. 19–28.

Darling-Hammond, L. and J. DePaoli (2020) “Why School Climate Matters and What Can Be Done to Improve It”, *State Education Standard*, **20** (2), p. 7.

Diagne, D. (2009) “School Violence: Evidence from the Economics Literature and Related Disciplines”, *Swiss Journal of Educational Research*, **31** (1), pp. 135–150.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013) “Long-Term Effects of Class Size”, *Journal of Human Resources*, **48** (1), pp. 1–29.

Gershenson, S., A. Jacknowitz, and A. Brannegan (2017) “Are Student Absences Worth the Worry in U.S. Primary Schools?”, *Education Finance and Policy*, **12** (2), pp. 137–165.

Hanushek, E. A. (1999) “Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects”, *Educational Evaluation and Policy Analysis*, **21** (2), pp. 143–163.

——— (2003) “The failure of input-based schooling policies”, *The economic journal*, **113** (485), pp. 64–98.

——— and L. Woessmann (2017) “Cognitive Abilities and Educational Outcomes”, in *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson*: Springer, pp. 149–171.

——— , S. G. Rivkin, and L. L. Taylor (1996) “Aggregation and the Estimated Effects of School Resources”, *The Review of Economics and Statistics*, **78** (4), pp. 611–627.

——— , G. Schwerdt, S. Wiederhold, and L. Woessmann (2015) “Returns to Skills around the World: Evidence from PIAAC”, *European Economic Review*, **73**, pp. 103–130.

Heckman, J. J., J. Stixrud, and S. Urzua (2006) "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior", *Journal of Labor Economics*, **24** (3), pp. 411–482.

Hojo, M. (2013) "Class-size Effects in Japanese Schools: A Spline Regression Approach", *Economics Letters*, **120** (3), pp. 583–587.

——— (2021) "Association between student-teacher ratio and teachers' working hours and workload stress: evidence from a nationwide survey in Japan", *BMC public health*, **21**, pp. 1–8.

Hoxby, C. M. (2000) "The Effects of Class Size on Student Achievement: New Evidence from Population Variation", *The Quarterly Journal of Economics*, **115** (4), pp. 1239–1285.

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**, p. 100977.

Ito, H. and S. Tabata (2022) "The Role of Teachers from the Perspective of Teacher Value-Added: Evidence from Japanese Elementary School Students", *Japanese Journal of Labour Studies* (740), pp. 48–69, In Japanese.

Japan Institute for Labour Policy and Training (JILPT)"Useful Labor Statistics: Processed Labor Statistics Indicators",Technical report, Japan Institute for Labour Policy and Training, In Japanese.

Kaiser, C. and A. J. Oswald (2022) "The Scientific Value of Numerical Measures of Human Feelings", *Proceedings of the National Academy of Sciences*, **119** (42), p. e2210412119.

Krueger, A. B. (1999) "Experimental estimates of education production functions", *The quarterly journal of economics*, **114** (2), pp. 497–532.

Lazear, E. P. (2001) "Educational production", *The quarterly journal of economics*, **116** (3), pp. 777–803.

Leuven, E., H. Oosterbeek, and M. Rønning (2008) "Quasi-experimental estimates of the effect of class size on achievement in Norway", *The Scandinavian Journal of Economics*, **110** (4), pp. 663–693.

Lindqvist, E. and R. Vestman (2011) "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment", *American Economic Journal: Applied Economics*, **3** (1), pp. 101–128.

Ministry of Education, Culture, Sports, Science and Technology (2019) "School Teacher Statistics Survey, Fiscal Year 2019".

Momose, Y. and H. Ishida (2024) "Bullying Experiences in Childhood and Health Outcomes in Adulthood", *PLOS ONE*, **19** (7), p. e0305005.

Mulligan, C. B. (1999) "Galton versus the Human Capital Approach to Inheritance", *Journal of Political Economy*, **107** (S6), pp. S184–S224.

Murnane, R. J., J. B. Willett, Y. Duhaldeborde, and J. H. Tyler (2000) "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?", *Journal of Policy Analysis and Management*, **19** (4), pp. 547–568.

Nakamuro, M. (2017) "Do Small Classes Reduce Bullying, Violence, and Truancy?", RIETI Discussion Paper Series 17-J-074, Research Institute of Economy, Trade and Industry (RIETI), pp. 1–37, in Japanese.

Niki, M. (2013) "The Effects of Class Size on Cognitive and Non-Cognitive Abilities", *Economic Analysis*, **186**, pp. 30–49, in Japanese.

OECD (2017a) *PISA 2015 Results (Volume III): Students' Well-Being*, Paris: OECD Publishing, PISA.

——— (2017b) *PISA 2015 Results (Volume III): Students' Well-Being*, Paris: OECD Publishing.

——— (2021a) *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*, Paris: OECD Publishing.

——— (2021b) *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*, Paris: OECD Publishing.

Oikawa, M., R. Tanaka, S. ichiro Bessho, and H. Noguchi (2022) "Do Class Size Reductions Protect Students from Infectious Diseases? Lessons for COVID-19 Policy from a Flu Epidemic in the Tokyo Metropolitan Area", *American Journal of Health Economics*, **8** (4), pp. 449–476.

Olea, J. L. M. and C. Pflueger (2013) "A Robust Test for Weak Instruments", *Journal of Business & Economic Statistics*, **31** (3), pp. 358–369.

Senoh, W. and M. Hojo (2016) "Does the class size affect the academic achievement of junior high school students? : an empirical study with "National Assessment of Academic Ability" of Japan", *NIER Research Bulletin*, **145**, pp. 119–128, In Japanese.

Stock, J. H. and M. Yogo (2005) "Testing for Weak Instruments in Linear IV Regression", in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: Cambridge University Press, pp. 80–108.

Sugita, S. (2025) "The Effect of Class-Size Reduction on Student Satisfaction with School Life", *KEIO SFC JOURNAL*, **24** (2), pp. 161–171, In Japanese.

Tanaka, R. and T. Wang (2025) "How do classmates matter for the class-size effects?", *Economics of Education Review*, **108**, p. 102689.

Urquiola, M. (2006) "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia", *The Review of Economics and Statistics*, **88** (1), pp. 171–177.

A Appendix

A.1 Results for Six Individual Items

Tables 12, 13, and 14 present 2SLS estimates for each of the six individual items (of teahcer–student relationships and classroom climate) under three specifications: (1) controlling only for student-level covariates (baseline); (2) adding school fixed effects; and (3) adding student fixed effects.

Table 12: Results on 6 Individual Items Included in Composite Indices (Baseline)

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.004***	-0.005***	-0.005***	-0.003***	-0.006***	-0.003***
	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Grade Enrollment	0.000**	0.000	-0.000	0.001***	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment^2	-0.000**	-0.000	0.000	-0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.113***	0.115***	0.052***	0.032***	0.068***	0.078***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Private Tutoring	-0.014***	-0.014***	-0.011***	-0.033***	-0.061***	-0.029***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Books at Home	-0.113***	-0.081***	-0.092***	-0.078***	-0.042***	-0.082***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Observations	1,678,517	1,673,981	1,684,788	1,690,464	1,685,753	1,685,982

Note: Each column reports a separate 2SLS regression for one constituent item. Grade and year fixed effects included.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2 Within-student Changes in class size Across Grades

To document the limited within-student variation in class size that the student fixed effects model relies upon for identification, Figure 5 shows the distribution of within-student standard deviations in class size. The distribution is right-skewed, with mass concentrated near zero, reflecting the fact that students in the Japanese school system typically remain in the same school across grades.

Table 13: Results on 6 Individual Items Included in Composite Indices (School FE)

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.003***	-0.003***	-0.004***	-0.003***	-0.007***	-0.002***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Grade Enrollment	0.000	0.001*	0.000	0.001***	0.001	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Grade Enrollment^2	-0.000	-0.000*	-0.000**	-0.000***	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.113***	0.115***	0.051***	0.031***	0.067***	0.078***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Private Tutoring	-0.012***	-0.014***	-0.012***	-0.032***	-0.061***	-0.029***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Books at Home	-0.111***	-0.080***	-0.093***	-0.077***	-0.040***	-0.082***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Observations	1,678,517	1,673,981	1,684,788	1,690,464	1,685,753	1,685,982

Note: Each column reports a separate 2SLS regression for one constituent item. School, grade, and year fixed effects included.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 14: Results on 6 Individual Items Included in Composite Indices (School FE + Student FE)

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.004***	-0.004***	-0.004***	-0.002***	-0.008***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Grade Enrollment	0.000	0.000	-0.000	0.001*	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Grade Enrollment^2	-0.000	-0.000	-0.000	-0.000***	-0.000***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.066***	0.046**	0.022	0.084***	0.045**	0.049***
	(0.015)	(0.018)	(0.016)	(0.017)	(0.019)	(0.016)
Private Tutoring	-0.004**	-0.002	-0.005***	-0.010***	-0.011***	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Books at Home	-0.030***	-0.029***	-0.031***	-0.026***	-0.026***	-0.031***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	1,678,517	1,673,981	1,684,788	1,690,464	1,685,753	1,685,982

Note: Each column reports a separate 2SLS regression for one constituent item. School, student, grade, and year fixed effects included.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.3 Estimates Using Aggregated School-Level Data

Table 16 presents 2SLS estimates using school-level aggregated data[5]. Table **??** shows the first-stage estimates for these models. All specifications satisfy the weak IV criterion of

[5] A similar robustness check was conducted in Ito et al. (2020).

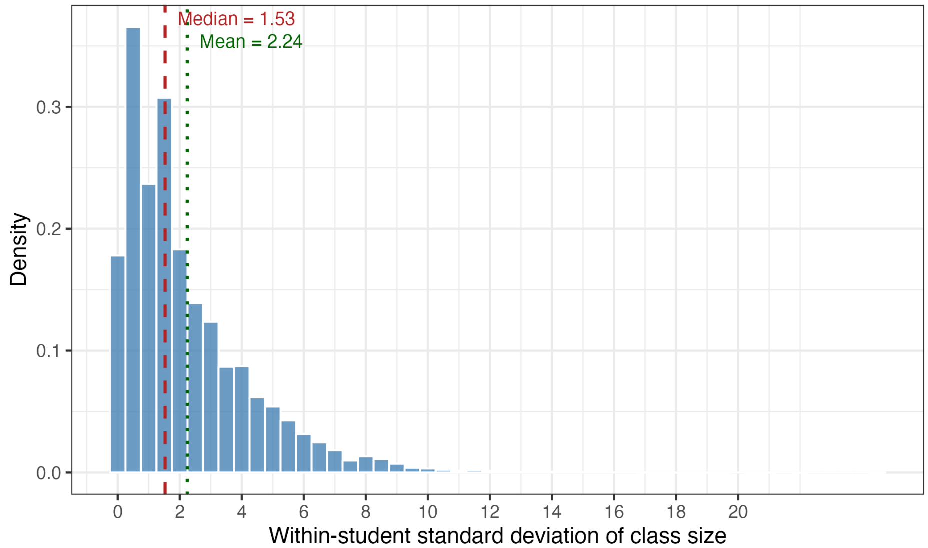


Figure 5: Distribution of Within-Student Standard Deviation of Class Size
Note: The figure shows the distribution of within-student standard deviations of class size across all observation years, computed for students with two or more observations. Dashed line (red): median (1.50); dotted line (green): mean (2.11).

$F > 10$.

All coefficients are negative and significant at the 1% level. The estimated effects range from 0.03 to 0.09 SD for a 10-student reduction in class size. The overall pattern of results is consistent across specifications.

The largest effect (0.09 SD) is observed in Column (2): Calm Learning Environment, and is greater than the estimates based on individual-level data. Hanushek et al. (1996) caution that estimates using aggregated data can be upward biased compared to those using student-level data. Our findings are consistent with that concern.

A.4 Results for Binary Outcomes

Table 17 presents the 2SLS estimation results using binary outcome variables for classroom climate and teacher–student relationships. In this specification, each outcome is coded as 1 if the student responded with a 3 or 4, and 0 otherwise. Since we estimate a linear probability model using these binary variables, the coefficients represent percentage point changes in the probability of reporting a high level of classroom climate or teacher–student relationship

Table 15: First Stage Estimates (School-Level Data)

	(1)	(2)	(3)
Predicted Class Size	0.874***	0.714***	0.708***
	(0.007)	(0.012)	(0.012)
Num.Obs.	23 116	23 116	23 116
R^2	0.809	0.846	0.848
F-stat (Predicted Class Size)	14981.5***	3575.8***	3508.4***

Note: The dependent variable is average class size at the school–grade–year level. Column (1) includes no fixed effects. Column (2) adds school fixed effects. Column (3) further adds year fixed effects.
All specifications include the following school-level covariates: share of female students, share with no private tutoring, and share with no books at home.
The instrument (Predicted Class Size) is derived from the Maimonides Rule and aggregated to the school–grade–year level ($\lfloor (GS-1)/40 \rfloor + 1$, where GS denotes grade enrollment).
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 16: Results on 6 Individual Items Included in Composite Indices (School-Level Aggregated Data)

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.004***	-0.003***	-0.005***	-0.004***	-0.009***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	23,074	23,076	23,084	23,093	23,079	23,087

Note: Class size coefficients from school-level 2SLS estimation, aggregating student questionnaire outcomes to school-grade-year cells.
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

quality in response to a one-student reduction in class size.

As shown in Table 17, the pattern of coefficient signs and statistical significance remains largely unchanged. A class-size reduction of 10 students is associated with an increase of approximately 1 to 3 percentage points in the likelihood of students reporting high classroom climate or teacher–student relationship quality. According to the results in Section 5.4.4, *Calm Environment* is the outcome with the greatest year-to-year variation. This is consistent with the result in this section, where the largest treatment effect is also observed for *Calm*

Environment.

Table 17: Results for Binary Outcomes

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.001***	-0.001***	-0.001***	-0.000***	-0.003***	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment	0.000	0.000	0.000	0.000*	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Grade Enrollment^2	-0.000	-0.000	-0.000**	-0.000**	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.035***	0.032***	0.019***	0.008***	0.030***	0.022***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Private Tutoring	0.000	-0.001	-0.000	-0.008***	-0.024***	-0.004***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Books at Home	-0.027***	-0.025***	-0.028***	-0.022***	-0.012***	-0.022***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1,678,517	1,673,981	1,684,788	1,690,464	1,685,753	1,685,982

Note: Each outcome is coded 1 if the student responded 3 or 4 on the 4-point scale, and 0 otherwise. Linear probability models estimated by 2SLS; coefficients are percentage-point changes (divided by 100) in the probability of an affirmative response per one-student change in class size. School, grade, and year fixed effects included.

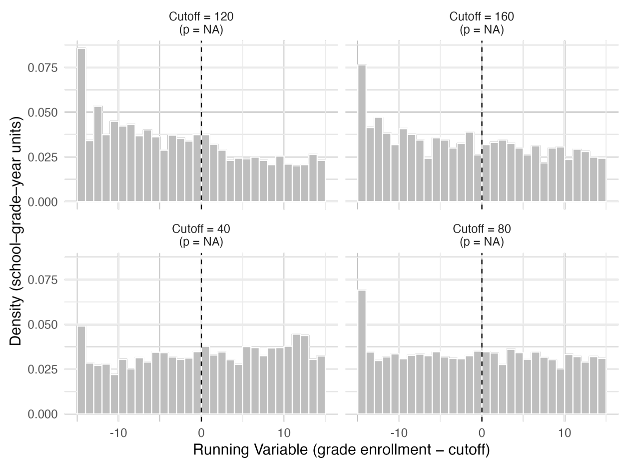
Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.5 RD Validity Checks

This section verifies the internal validity of the regression discontinuity design underlying the Maimonides rule instrument.

Density Continuity (McCrary Test)

Figure 6: Density of Grade Enrollment Near Cutoffs



Notes:
The dashed lines indicate cutoff points at multiples of 40. The histogram is based on samples with grade sizes within ±15 students of each cutoff.

Figure 6 presents histograms of grade enrollment near each cutoff. No apparent bunching or discontinuity in the density is visually detectable. To formally test for manipulation, we apply the density discontinuity test of Cattaneo et al. (2018) at the school–grade–year level. Results are reported in Table 18. The p-values at all four cutoffs range from 0.34 to 0.62, all well above conventional significance thresholds, providing no evidence of manipulation. Bonferroni and Holm corrections for multiple testing yield adjusted p-values of 1.000 at all four cutoffs.

Covariate Balance

To further assess RD validity, we test for discontinuities in pre-determined student characteristics at each cutoff. Table 19 reports the estimated discontinuity in the share female, share with no private tutoring, and share with no books at home, for bandwidths of ±5 and

Table 18: McCrary Density Test at Each Cutoff

		Adjusted p-value				
Cutoff	T-statistic	p-value	N (left)	N (right)	Bonferroni	Holm
Cutoff = 40	0.951	0.341	5346	2314	1.000	1.000
Cutoff = 80	0.923	0.356	5973	2947	1.000	1.000
Cutoff = 120	0.827	0.408	3445	2061	1.000	1.000
Cutoff = 160	-0.499	0.618	1631	911	1.000	1.000

Note:
The null hypothesis is that there is no discontinuity in the density of the running variable (grade enrollment) at the cutoff. The test statistic is the jackknife-based T-statistic from Cattaneo, Jansson, and Ma (2018). Bonferroni and Holm corrections are applied across the four cutoffs to account for multiple testing. The sample is restricted to observations within ± 15 students of each cutoff. Observations are at the school–grade–year level.

± 10 students. No statistically significant discontinuity is detected in any covariate at any cutoff, confirming that observable student characteristics are smooth through the cutoff.

First-Stage Validity in Subsamples

To verify that the Maimonides rule instrument remains valid in the subsamples used for heterogeneity analysis, we present first-stage regression discontinuity plots for each subsample. Figure 7 shows the first-stage RD plot for the subsample split by school-level educational aid rate, and Figure 8 shows the corresponding plot for the pre- and post-COVID subsamples. In both cases, a clear discontinuity in actual class size is visible at the enrollment cutoffs, confirming that the instrument retains sufficient first-stage variation across subgroups.

A.6 Complier Characteristics

The 2SLS estimate based on the Maimonides rule instrument applies to schools near the enrollment cutoffs—the compliers. Table 20 compares student characteristics between schools near any cutoff (± 10 students) and schools farther from the cutoffs (> 10 students).

There are no statistically significant differences in the share female, share with no private tutoring, or share with no books at home between near-cutoff and far-from-cutoff schools. This indicates that complier schools are not systematically different from non-complier schools in terms of observable student characteristics. The 2SLS estimate therefore

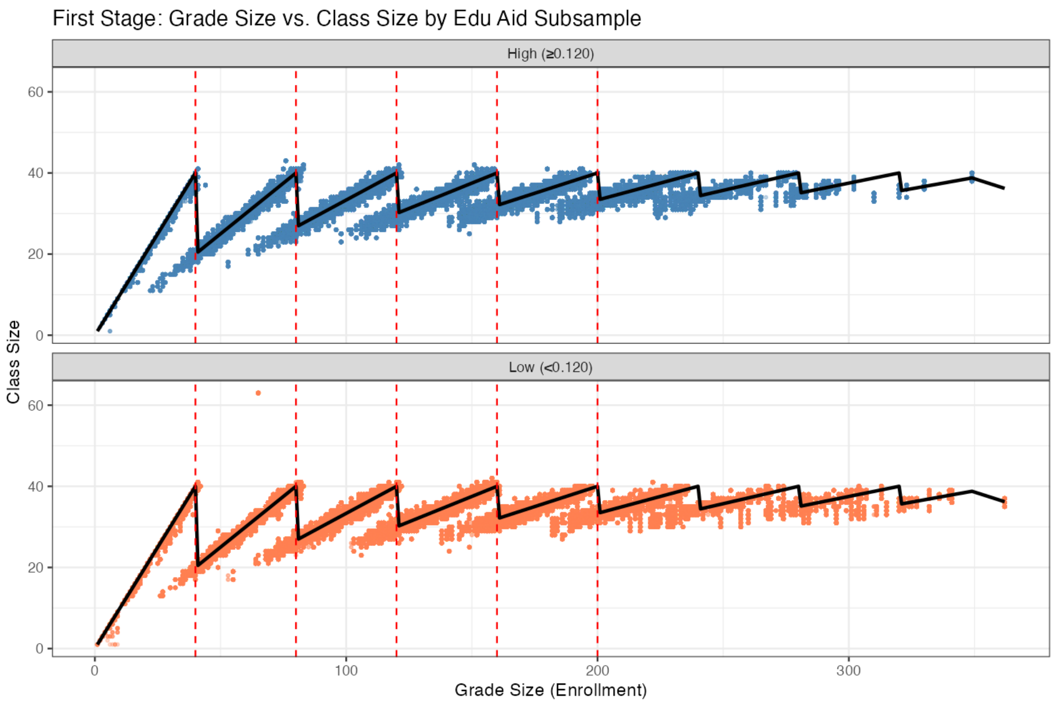


Figure 7: First-Stage RD Plot: Poverty Rate Subsamples

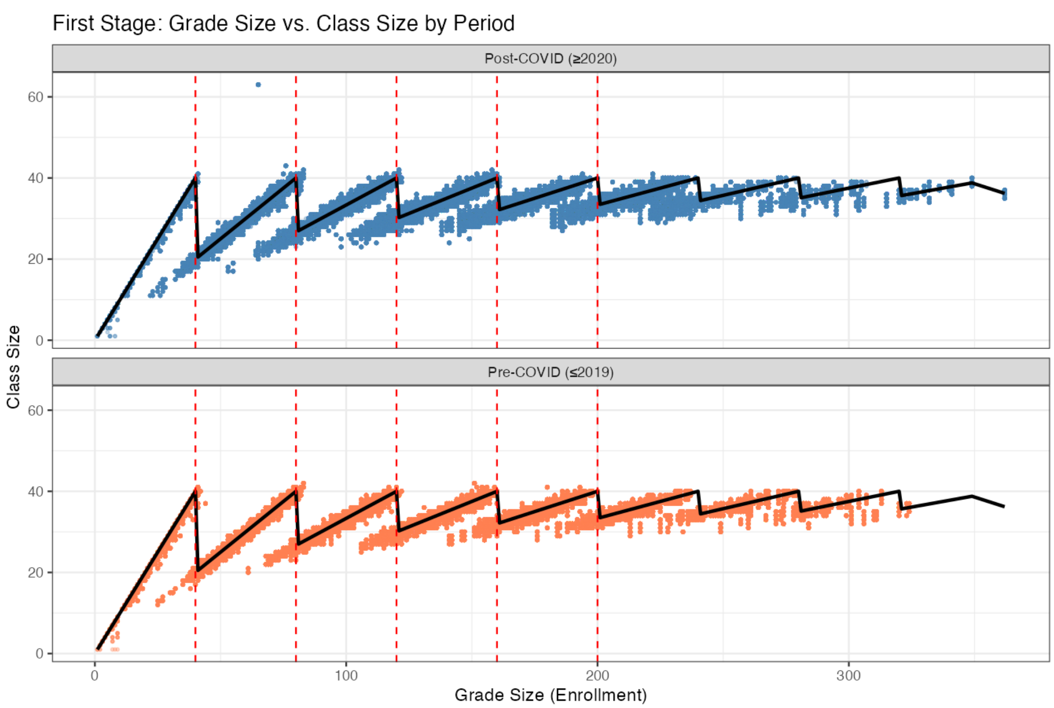


Figure 8: First-Stage RD Plot: Pre- and Post-COVID Subsamples

Table 19: Covariate Balance at the Cutoff

	Bandwidth	
Variable	±5	±10
% Female	-0.0065	-0.0025
	(0.0086)	(0.0065)
% No Private Tutoring	-0.0154	-0.0046
	(0.0093)	(0.0064)
% No Books at Home	-0.0057	-0.0006
	(0.0058)	(0.0044)
Observations	5,779	10,887

Note:
Each cell reports the coefficient on an indicator for being above the Maimonides cutoff from an OLS regression of the covariate on the above-cutoff indicator, a quadratic in the running variable, and cutoff fixed effects. The unit of observation is a school–grade–year cell. Covariates are the share of students within each cell with the given characteristic. Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 20: Characteristics of Schools Near and Far from Maimonides Cutoffs

Variable	Near Cutoff (±10)	Far from Cutoff (> 10)	t-statistic
% of Female	0.487	0.485	1.053
	(0.150)	(0.154)	
% of No Private Tutoring	0.417	0.417	-0.307
	(0.166)	(0.175)	
% No Books at Home	0.114	0.115	-0.980
	(0.109)	(0.117)	
Observations (school-grade-year)	10,910	12,249	

Note:
The unit of observation is a school–grade–year cell. 'Near cutoff' is defined as being within ±10 students of any of the four Maimonides cutoffs (40, 80, 120, 160). Mean reported with standard deviation in parentheses. t-statistics test equality of means between near- and far-cutoff groups. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

applies primarily to medium-sized schools (near the 40-, 80-, 120-, or 160-student cutoffs), and generalization to very small or very large schools should be made with caution.

A.7 Grade 5 (2024) Cohort under the 35-Student Class-Size Cap

Table 21 presents 2SLS estimates for Grade 5 students in 2024—the first cohort subject to the revised 35-student class-size cap introduced under Japan's 2021 reform of the Act on Standards for Compulsory Education. We apply the standard 40-student Maimonides rule instrument (predicted class size based on a 40-student cap) to this cohort to examine whether the instrument retains its first-stage relevance under the new statutory framework.

The first-stage F-statistic of 0.26 falls far below any conventional threshold for instrument relevance, indicating that the 40-student Maimonides rule generates essentially no variation in actual class size for students already subject to the 35-student cap. This is expected: once the statutory cap shifts to 35 students, the enrollment cutoffs that drive discontinuous variation in class size under the 40-student rule no longer bind, rendering the instrument irrelevant for this cohort. Accordingly, this cohort is excluded from the main analysis (see Section 3.1), and all main estimates are identified from cohorts to which the 40-student Maimonides rule continues to apply.

A.8 Cost-Effectiveness: Detailed Calculations

This section extends the cost-effectiveness discussion in Section 6 by estimating the lifetime earnings benefits of class-size reduction through two distinct channels: (A) academic achievement and (B) classroom climate and teacher–student relationships.

A.8.1 Method

The per-child lifetime benefit through channel $k \in \{A, B\}$, expressed as a discounted present value at the time of intervention, is defined as:

$$\text{Benefit}_k = \Delta\text{SD}_k \times \rho_k \times \text{PDV}(L) \tag{4}$$

where ΔSD_k is the estimated effect of a 10-student class-size reduction on outcome k (in SD units), ρ_k is the wage return per 1 SD improvement in outcome k, and $\text{PDV}(L)$ is the discounted present value of lifetime earnings. Treating nominal lifetime earnings L as a

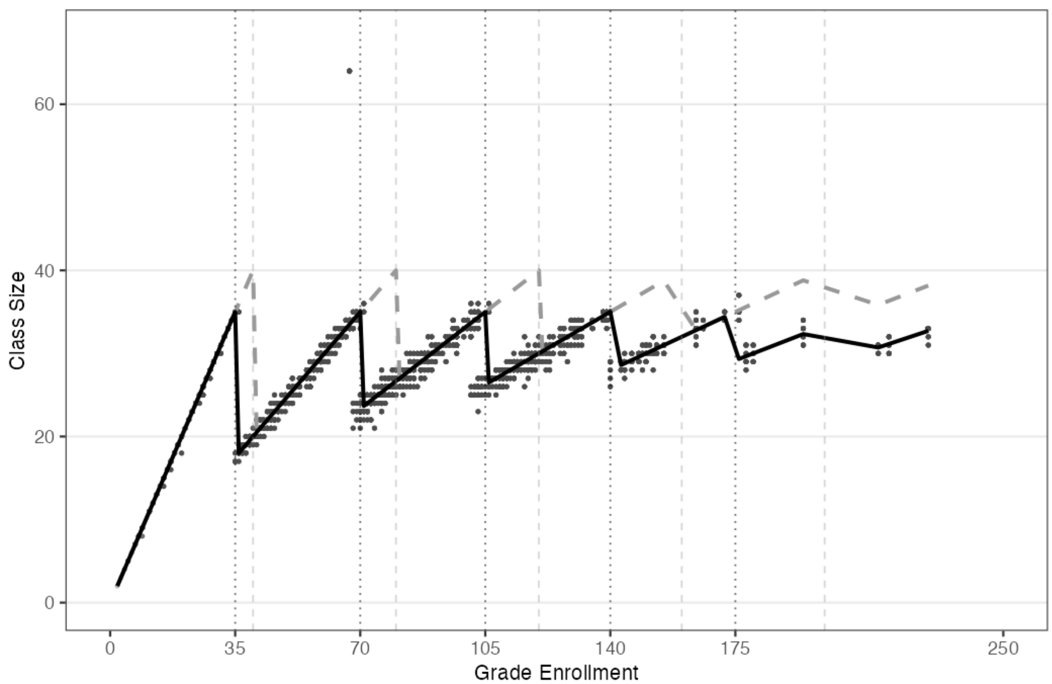


Figure 9: Class Size and Grade Enrollment: Grade 5 (2024) under the 35-Student Cap

Notes: Each dot represents one school–grade–year cell. The solid line plots predicted class size under the standard 40-student Maimonides rule ($PS = GS/(\lfloor (GS-1)/40 \rfloor + 1)$, where GS denotes lagged grade enrollment). The sample is restricted to Grade 5 students in 2024—the first cohort subject to the revised 35-student statutory cap. Because the legal maximum has shifted to 35 students, the 40-student rule cutoffs no longer bind: schools split classes before reaching 40, flattening the discontinuity and rendering the instrument irrelevant for this cohort. This is consistent with the first-stage F-statistic of 0.26 reported in Table 21.

Table 21: Results on 6 Individual Items Included in Composite Indices (Grade 5, 2024: 35-Person Class Rule)

Variable	Teacher–Student Relationship			Classroom Climate		
	(1) Recognized	(2) Listened	(3) Explained	(4) Satisfaction	(5) Calm Environment	(6) Cooperation
Class Size	-0.094	-0.201	-0.066	-0.213	-0.083	-0.116
	(0.193)	(0.403)	(0.136)	(0.410)	(0.196)	(0.227)
Grade Enrollment	0.014	0.030	0.009	0.034	0.012	0.017
	(0.030)	(0.063)	(0.021)	(0.064)	(0.030)	(0.035)
Grade Enrollment^2	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (female=1)	0.173***	0.115***	0.096***	0.095***	0.063***	0.092***
	(0.011)	(0.017)	(0.010)	(0.016)	(0.011)	(0.011)
Private Tutoring	-0.019*	-0.034**	-0.037***	-0.037**	-0.084***	-0.060***
	(0.011)	(0.016)	(0.010)	(0.016)	(0.011)	(0.012)
Books at Home	-0.193***	-0.110***	-0.131***	-0.086***	-0.041***	-0.136***
	(0.018)	(0.023)	(0.015)	(0.024)	(0.016)	(0.017)
Observations	41,200	40,499	42,473	44,017	42,471	42,105
First-stage Coef. (PS_{ts})	0.0147					
	(0.0289)					
First-stage F-statistic	0.26					

Note: Sample restricted to Grade 5 students in 2024, the first cohort exposed to the 35-person class size cap under the revised Act on Standards for Compulsory Education (2021). Instrument: predicted class size under the standard 40-person Maimonides Rule ($PS_{ts} = GS_{ts}/(\text{int}((GS_{ts} - 1)/40) + 1)$, where GS_{ts} denotes lagged grade enrollment). No school or year fixed effects (single grade-year cross-section). Standard errors clustered at the school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

constant annual income stream L/T over T working years, with t_0 years between intervention and entry into the labor market and discount rate r:

$$\text{PDV}(L) = \frac{L}{T} \cdot \frac{1 - (1 + r)^{-T}}{r} \cdot (1 + r)^{-(t_0 - 1)} \tag{5}$$

The cost is defined as in Section 6: denoting the annual employment cost per teacher as C and the baseline class size as N, the marginal per-student annual cost of a 10-student reduction is:

$$\text{Cost}(N) = C\left(\frac{1}{N - 10} - \frac{1}{N}\right) \tag{6}$$

The benefit-cost ratio is $\text{BCR}_k = \text{Benefit}_k/\text{Cost}(N)$. Since this study estimates the effect of approximately one school year of exposure per child, benefits are interpreted as arising from one year of class-size reduction, and costs are expressed on a per-year basis.

A.8.2 Parameters

Effect sizes (this study's estimates). For channel A, we use a conservative central estimate that lies midway between the two main specifications. From Table 5, a 10-student reduction improves reading and mathematics scores by approximately 0.02 SD each under the school-FE specification, while the effects are indistinguishable from zero under the student-FE specification. We therefore set $\Delta\mathrm{SD}_A = 0.01$—the midpoint between the school-FE estimate (0.02) and the student-FE estimate (0)—as the central value. $\Delta\mathrm{SD}_A = 0$ is retained as the lower bound in sensitivity analysis. For channel B, $\Delta\mathrm{SD}_B = 0.04$ is taken from the composite index estimates (Table 3; both the teacher–student relationship index and the classroom climate index equal 0.04 SD). As an upper-bound reference, the 0.26 SD effect on classroom discipline from the school questionnaire (Table 4) is noted, but treated as supplementary given that it is estimated from school-level aggregated data and subject to upward bias (Hanushek et al., 1996).

Wage returns (ρ). For channel A, studies of U.S. youth estimate that a 1 SD increase in test scores at the end of high school is associated with a 10–15% increase in annual earnings (Mulligan, 1999; Murnane et al., 2000). More broadly, Hanushek et al. (2015) report an average wage return of 18% per 1 SD in numeracy skills across 23 countries using PIAAC. Because Japan has a relatively compressed domestic skill distribution, the within-country SD return may differ from international estimates (Hanushek et al., 2015). Since the test scores in this study are standardized within year–grade cells, the relevant parameter is the within-country SD return; we use a central value of $\rho_A = 0.13$, with a range of 0.10–0.18.

For channel B, Lindqvist and Vestman (2011) synthesize 13 studies and find that a 1 SD improvement in noncognitive skills is associated with a 4–8% wage increase, while Heckman et al. (2006) estimate an 11.2% return per 1 SD in latent noncognitive ability. We use a central value of $\rho_B = 0.06$, with a range of 0.04–0.11. As an alternative monetization of channel B, the long-run evidence already cited in this paper—showing that disruptive classroom environments reduce adult earnings (Carrell et al., 2018), that student absence reduces adult earnings and employment (Cattan et al., 2023; Gershenson et al., 2017), and that early classroom quality affects earnings beyond the fade-out of test score effects (Chetty et al., 2011b)—could also be used to directly monetize behavioral outcomes (discipline, violence, absence).

Other parameters. Total teacher employment cost is C = 8,113,304 JPY (Ministry of Education, Culture, Sports, Science and Technology 2019; see Table 24 for details), with an exchange rate of 150 JPY/USD. Lifetime earnings are set conservatively at L = 200 million JPY based on Japan Institute for Labour Policy and Training (JILPT) (ranging from approximately 200–270 million JPY by educational attainment). Working years are $T = 38$, years to labor market entry $t_0 = 12$ (assuming intervention at age 10 and labor market entry at age 22). Sensitivity analysis is conducted over discount rates of 2%, 3%, and 4%, and baseline class sizes of 25, 30, 35, and 40 students.

A.8.3 Results

Table 22 reports the per-child lifetime benefit (discounted present value) for each channel. At a 3% discount rate with central returns, channel A yields a benefit of approximately 111,000 JPY and channel B yields approximately 205,000 JPY—a ratio of approximately 1.85 in favor of channel B. This ratio is invariant to the discount rate, since PDV(L) cancels out between channels.

Table 23 reports benefit-cost ratios (BCR) at central returns across combinations of baseline class size and discount rate. Because costs are lower for larger classes, BCRs are higher for reductions from larger class sizes. At a 3% discount rate, reducing class size from 30 to 20 students (close to the sample mean) yields $\text{BCR}_A = 0.82$ and $\text{BCR}_B = 1.52$; channel B exceeds the breakeven threshold, while channel A does not at this discount rate. At a 2% discount rate, $\text{BCR}_A = 1.07$ also exceeds 1. Channel B's BCR exceeds channel A's across all combinations of discount rate and class size.

Channel B's BCR exceeds channel A's across all parameterizations. The primary driver of $\text{BCR}_B > \text{BCR}_A$ is the larger effect size for classroom environment ($\Delta\text{SD}_B = 0.04$ vs. $\Delta\text{SD}_A = 0.01$), despite a higher per-SD wage return for channel A ($\rho_A = 0.13$ vs. $\rho_B = 0.06$). Channel B rests on the assumption that improvements in classroom environment measures can be mapped to a 1 SD improvement in noncognitive skills, which is an approximation.

Under the student-FE specification, the academic effect approaches zero, implying $\text{BCR}_A \to 0$.

Taken together, while channel B's BCR consistently exceeds channel A's under the current parameterization, the relative magnitudes depend strongly on the estimated effect sizes and assumptions about wage returns. Given that the two channels are not independent—improvements in classroom climate may also affect academic achievement—the benefits of the two channels should not be summed, but rather treated as separate illustrative benchmarks for the policy relevance of class-size reduction through each pathway.

Table 22: Lifetime Benefits by Channel (Per Child, Discounted Present Value of Lifetime Earnings)

	Lifetime Earnings (PV)	Channel A: Academic Achievement (10,000 JPY)			Channel B: Relations & Climate (10,000 JPY)		
Discount rate r	(100 million JPY)	Low	Central	High	Low	Central	High
2%	1.12	11.2	14.5	20.1	17.9	26.9	49.2
3%	0.86	8.5	11.1	15.4	13.7	20.5	37.6
4%	0.66	6.6	8.6	11.9	10.6	15.9	29.1

Note: Benefits are per-child lifetime benefits from one school year of exposure, calculated using Equation (4). Effect sizes are $\Delta\text{SD}_A = 0.01$ (school-FE specification; Table 5) and $\Delta\text{SD}_B = 0.04$ (composite indices; Table 3). "Low," "Central," and "High" correspond to the range of wage returns per 1 SD: Channel A: 0.10/0.13/0.18; Channel B: 0.04/0.06/0.11. Lifetime earnings $L = 200$ million JPY, working years $T = 38$, years to labor market entry $t_0 = 12$.

Table 23: Benefit-Cost Ratios by Channel (Central Returns, Baseline Class Size × Discount Rate)

Baseline class size (Before → After)	Cost (10,000 JPY)	BCR_A: Academic Achievement			BCR_B: Relations & Climate		
		r=2%	r=3%	r=4%	r=2%	r=3%	r=4%
25 → 15	21.6	0.67	0.51	0.40	1.24	0.95	0.73
30 → 20	13.5	1.07	0.82	0.63	1.99	1.52	1.18
35 → 25	9.3	1.57	1.20	0.93	2.90	2.21	1.71
40 → 30	6.8	2.15	1.65	1.27	3.97	3.04	2.35

Note: Central wage returns ($\rho_A = 0.13$, $\rho_B = 0.06$) are used. Cost is the per-child per-year marginal cost calculated using Equation (6), based on teacher employment cost $C =$ 8,113,304 JPY. BCR > 1 indicates that lifetime benefits exceed one year of per-child costs.

Table 24: Teacher Employment Cost Assumptions

Item	JPY (per year)
Monthly salary (incl. allowances)	5,160,000
Annual bonus	1,913,500
Employer social insurance	1,039,804
Total employment cost	8,113,304

Note:
Based on Ministry of Education, Culture, Sports, Science and Technology (2019). Monthly salary includes base pay and all regular allowances for public elementary and junior high school teachers. Bonus is calculated at 4.45 months per the National Personnel Authority recommendation (fiscal year 2019). Social insurance rate of 14.7% covers health insurance, employees' pension, and employment insurance (employer share).

Table 25: Cost of a 10-Student Class-Size Reduction: Sensitivity Analysis

	Annual additional cost per student		Cost per 1 SD improvement in well-being	
Class size (before → after)	JPY	USD	JPY	USD
25 → 15	216,355	1,442	5,408,870	36,059
30 → 20 (baseline)	135,222	901	3,380,544	22,537
35 → 25	92,723	618	2,318,087	15,454
40 → 30	67,611	451	1,690,272	11,268

Note: A 10-student reduction in class size (from N to $N-10$) requires an additional $\frac{1}{N-10} - \frac{1}{N}$ teacher per student. Annual additional cost per student $= C \times \left(\frac{1}{N-10} - \frac{1}{N}\right)$, where $C =$ JPY 8,113,304 is the total employment cost per teacher (see Table 24). Cost per 1 SD improvement divides the annual cost by the estimated effect size of 0.04 SD. USD figures use an exchange rate of JPY 150/USD.