When Does Teacher Specialization Work?

Evidence from a Cluster-Randomized Experiment in Elementary Schools

Makiko NAKAMURO Chien Quang LE
Kengo IGEI Tomoya MURAKAWA

Abstract

We estimate the causal effect of subject specialization in elementary schools using a cluster-randomized experiment in Japan. Assigning part-time science specialists increases science achievement by 0.18 SD and mathematics by 0.13 SD, while mathematics specialization yields no gains. The asymmetry reflects comparative advantage. Student achievement is strongly linked to homeroom teachers' subject expertise, and reallocating instruction away from high-expertise teachers attenuates returns. We find no evidence of increased coordination costs or deterioration in teacher–student relationships. The intervention is substantially more cost-effective than class size reduction. Targeted specialization improves learning when aligned with teacher comparative advantage and subject-specific skill constraints.

Keywords: Teacher Specialization; Departmentalized Instruction

JEL Codes: I20; I21; I28

This study is conducted as a part of the Research Project of the undertaken at the Research Institute of Economy, Trade and Industry (RIETI), titled "Comprehensive Study to Promote Evidence-Based Policy Making (EBPM) in Japan." We thank Fumio Ohtake, Stephan Litschig, Jun Goto, Taiyo Fukai, Keisuke Kawata, Hiromi Hara for valuable comments. We thank seminar participants at seminars at RIETI and National Graduate Institute for Policy Studies, and Tokyo Labor Economics Workshop. We are especially grateful to the Chiba Prefectural Board of Education for their cooperation. This study was approved by the SFC Research Ethics Committee at Keio University (Approval No. 475).

1 Introduction

In recent years, the shortage of teachers and the problem of excessive working hours in Japanese elementary schools have become urgent policy concerns. In response, the government has implemented a range of reforms aimed at improving teachers' workload and securing a sustainable teaching workforce. Among these reforms, the introduction of teacher specialization (or departmentalized instruction) has attracted particular attention[1], although Japanese elementary schools have traditionally operated under a *homeroom-based system* in which a single teacher is responsible for teaching all subjects to one class.

Following the introduction of teacher specialization in Grades 5 and 6 in 2022, there has been a policy movement to extend this model to Grades 3 and 4 based on recommendations from the Central Council for Education (CCE) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan. In its August 2024 report, the CCE proposed expanding teacher specialization to lower upper-elementary grades, arguing that the annual instructional hours in Grade 4 are approximately 1,015 hours and nearly equivalent to those in Grades 5 and 6. Given this comparable workload, the council emphasized that introducing teacher specialization in Grades 3 and 4 could help alleviate teachers' workload. Reflecting this recommendation, the MEXT announced in December 2024 to expand the number of subject-specialist teachers at these grade levels, with budgetary approval secured for fiscal year 2025 and corresponding regulatory revisions underway.

The Japanese government has strongly promoted teacher specialization, primarily to reduce teacher workload. This policy rationale, however, leaves a critical question unanswered: whether specialization also benefits the students themselves. Empirical evidence on teacher specialization has generally been cautious, with limited support for positive achievement effects. Most notably, Fryer Jr. (2018), using a cluster randomized controlled trial (cluster RCT) in Texas elementary schools, finds that teacher specialization reduced student achievement and increased disciplinary

[1]Departmentalized instruction refers to a school-level organizational structure in which different teachers teach different subjects to the same group of students, whereas teacher specialization more broadly describes the narrowing of subject assignments at the individual teacher level, which may or may not involve a full restructuring of the homeroom system.

incidents. Similarly, subsequent studies have reported null or negative effects, particularly in settings where specialization substantially alters teacher–student relationships or increases coordination demands. In contrast, more recent quasi-experimental studies in the United States suggested that the effects may depend on institutional settings (Backes et al., 2024; Hwang & Kisida, 2022). Thus, rather than simply asking whether teacher specialization is effective, this paper examines when specialization improves students and teacher outcomes and how its effectiveness depends on institutional design, teacher comparative advantage, and subject-specific production technologies.

Japan offers a unique setting for this question: a historically strong homeroom-based system now coexists with policy-driven efforts toward specialization. This allows a clean test of whether subject-specific expertise can improve outcomes without dismantling the pedagogical continuity characteristic of elementary education. Moreover, Japan's centralized system, such as nationally standardized curricula, uniform instructional guidelines, and fixed instructional hours, enhances internal validity: observed differences in outcomes can be credibly attributed to changes in teacher assignment rather than institutional heterogeneity.

In the present study, we conducted a cluster RCT across 60 public elementary schools identified as relatively low-performing in Chiba Prefecture, Japan, to evaluate the impact of introducing subject-specialist teachers on both students and teachers. Unlike a complete shift to departmentalized instruction, the intervention in this study adopted partial specialization that preserved the homeroom structure while assigning retired licensed teachers as part-time specialist teachers specifically for mathematics and science classes. In the experiment, 20 schools were randomly selected to receive part-time mathematics specialist teachers, another 20 to receive part-time science specialist teachers, while the remaining 20 schools served as the control group.

Three findings emerge. First, assigning part-time science specialists raises science achievement by 0.18 SD, with a positive spillover of 0.13 SD in mathematics; mathematics specialization yields no detectable gains. Second, this asymmetry reflects comparative advantage: student achievement is strongly linked to homeroom teachers' subject expertise, and specialists generate gains only where such expertise is scarce. In mathematics, specialists crowd out homeroom teachers who are themselves math-specialized; in science, where fewer than 8 percent of homeroom teachers report science expertise, specialists fill a genuine skill gap. Third, these gains arise without the costs documented for full departmentalization: we find no deterioration in teacher–student relationships, no increase in behavioral problems, and no measurable rise in coordination burdens.

Our findings refine and extend prior evidence on departmentalization. Fryer Jr. (2018) and related studies document that full departmentalization in elementary schools can reduce achievement, potentially due to coordination costs and weakened student–teacher relationships. We show that specialization need not be harmful when institutional design mitigates these risks. Unlike full departmentalization, our intervention maintains the homeroom structure and introduces specialists only in selected subjects. More importantly, we demonstrate that outcomes depend on the distribution of subject expertise among existing teachers. Student achievement is strongly associated with homeroom teachers' subject specialization, and reallocating instruction away from highly specialized homeroom teachers attenuates gains. In mathematics, specialists crowd out high-quality homeroom teachers; in science, where subject expertise among homeroom teachers is scarce, specialization generates large returns.

The broader implication is that specialization is not inherently beneficial or harmful; its effectiveness depends on alignment between institutional design, teacher comparative advantage, and subject-specific production technologies.

We also contribute to the literature on teacher labor markets and policy design by showing that rehiring experienced retired teachers can be a cost-effective response to subject-specific skill constraints. Science instruction at the elementary level requires substantial preparation and experimental expertise, and surveys repeatedly document low teacher confidence in science (Nakajima & Kusaka, 2020). Deploying experienced specialists in such domains yields large achievement gains at a fraction of the cost of class size reduction.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 describes the institutional background and experimental design. Sections 4 and 5 present the data and empirical strategy. Section 6 reports the effects of the intervention on students and teachers, including a comparison with the nationally promoted departmentalization model. Section 7 assesses cost-effectiveness relative to class size reduction, and Section 8 presents robustness checks. Section 9 discusses mechanisms and policy implications, and Section 10 concludes.

2 Literature Review

A line of literature has argued that, in theory, specialization could enhance instructional productivity by exploiting teachers' comparative advantages. However, specialization substantially alters teacher–student relationships and/or increases coordination demands. Whether specialization yields positive or negative net effects is thus an empirical question that must be evaluated within specific institutional contexts. This is consistent with the theoretical framework of Becker and Murphy (1992), which conceptualizes specialization as a trade-off between productivity gains from narrower task focus and coordination costs among complementary workers.

In empirical work on teacher specialization (or departmentalized instruction), particularly at the primary education level, improvements in student achievement are rarely observed, and in some cases, negative effects have been documented (Bastian & Fortner, 2020; Fryer Jr., 2018; Hwang & Kisida, 2022). Among the existing studies, Fryer Jr. (2018) is particularly relevant to the present research. The author conducts a cluster RCT in low-performing public elementary schools in Houston, Texas. In 23 randomly selected schools, departmentalized instruction is introduced, while the remaining 23 schools continue with the conventional homeroom-based system. Pooled across two years, students in the treatment schools score 0.10 (SD) lower than their peers in the control group. The impact is especially large among students in special education programs and those taught by inexperienced teachers. Furthermore, students in treatment schools have 1.13 times more major behavioral infractions and attended 0.36 fewer days per year than students in control schools. Despite teaching fewer subjects, teachers are responsible for a larger number of students, making it more difficult to tailor instruction to individual learning proficiency. Teacher satisfaction in the treatment group is also significantly lower than in the control group.

Following Fryer's experimental evidence, subsequent quasi-experimental research has reached similar conclusions. Using longitudinal administrative data from Indiana, Hwang and Kisida (2022) shows that teachers are less effective when they specialize, with larger negative effects for vulnerable students, and found no evidence that greater specialization improves school-level outcomes.

Recent work by Backes et al. (2025) provides a more nuanced perspective on departmentalized instruction. Using administrative data from Massachusetts, Backes et al. (2024) find that while individual teachers tend to be less effective when they specialize, schools that adopt departmentalized models can nonetheless improve student achievement by reallocating teachers to their stronger subjects, thereby exploiting comparative advantage. Extending this analysis to longer-run outcomes, Backes et al. (2025) shows that exposure to departmentalized instruction at elementary schools modestly improves student achievement during the transition to middle school, although many of these effects fade out over time. Together, these studies suggest that the net effects of specialization depend not only on individual teacher productivity but also on how schools organize and deploy their instructional workforce.

The effects of specialization also have been found to vary by subjects and grades. Bastian and Fortner (2020) find that although relatively more effective teachers are assigned to specialize, specialization in mathematics and reading is associated with lower teacher value-added and does not improve aggregate achievement. In contrast, science specialization shows more positive patterns. These findings suggest that specialization may entail important trade-offs, particularly in core subjects.

Specialization also has significant impact on teachers, especially teachers' retention. Using statewide administrative data from North Carolina, Bastian et al. (2023) find that teachers are more likely to remain in the same school after becoming specialists, particularly in upper elementary grades, although specialization does not significantly reduce exits from the profession.

The surveys conducted in Japan also presents a mixed picture of specialization. Some surveys report improvements in student motivation, comprehension, and reduced anxiety toward middle school transition, as well as perceived instructional benefits from specialization (Asada & Nakanishi, 2018; Tokoro et al., 2022). However, other studies note concerns about increased teacher workload, particularly for preparation-intensive subject, such as science (Kajita & Doho, 2022; Tokoro et al., 2022). Most existing studies rely on teachers' subjective perceptions rather than counterfactual comparisons.

3 Experimental Design

3.1 Background and Intervention

Chiba Prefecture provides a large and demographically diverse educational setting for evaluating the impact of teacher specialization. The prefecture operates over 800 public elementary schools serving approximately 300,000 elementary school students, making it one of the largest school systems in Japan. Schools vary substantially in size and socioeconomic composition, spanning dense urban districts, suburban communities, and rural municipalities. Average class sizes are approximately 30 students, consistent with Japan's nationally regulated class-size standards. At the same time, all public elementary schools follow nationally standardized curricula and instructional guidelines, ensuring institutional uniformity across schools.

The intervention on teacher specialization in Chiba prefecture was explicitly designed to leverage comparative advantages while mitigating the risks identified in prior studies, such as Fryer Jr. (2018) and Hwang and Kisida (2022), including coordination costs and weakened teacher–student relationships. Unlike full departmentalized instruction, Chiba's teacher specialization intentionally maintained the homeroom-based system and selectively introduced subject-specialist teachers only in preparation-intensive subjects, particularly mathematics and science. Homeroom teachers continued to oversee student guidance and teach remaining subjects, thereby limiting potential disruptions while leveraging subject-specific expertise.

Prior research suggests that departmentalization staffed by inexperienced teachers can produce larger negative achievement effects (Fryer Jr., 2018). To avoid these risks, Chiba prefecture recruited former full-time teachers holding valid teaching licenses as part-time subject-specialist teachers in mathematics and science. This strategy was particularly important because the applicant-to-position ratio in Chiba's public school teacher recruitment examinations had reached historically low levels in 2024, limiting the prefecture's ability to recruit teachers with both subject expertise and classroom experience.[2]. By relying on experienced, licensed educators to serve as mathematics and science specialists, the intervention sought to strengthen instructional quality while mitigating the potential downsides documented in prior studies.

The part-time subject-specialist teachers hired through this intervention had an average age of 55.76 years, and they taught an average of 9.68 classes per week. A single teacher could be assigned to multiple schools, multiple grades, and multiple classes. In total, 38 specialists accepted posts for fourth-grade classes.

This intervention shares many similarities with the United Kingdom's supply teacher system, which is originally designed to address temporary staffing shortages due to absences or parental leave. In Japan, MEXT is reportedly considering the introduction of a similar regionally administered mechanism to employ retired teachers as part-time staff to fill temporary vacancies (Yomiuri Shimbun, 2025). However, evidence from the United Kingdom suggests that such systems may entail unintended educational costs. A 2024 report by the UK Department for Education documents that while most supply teachers are highly educated and many are experienced former full-time teachers, 64% of school principals reported adverse effects on student behavior and 61% reported negative effects on learning, although the evaluation is not based on causal inference.

3.2 Randomization

To rigorously evaluate the effect of this intervention on students' and teachers' outcomes, we run a cluster RCT. Schools were selected and randomly assigned using a stratified design to ensure balance across school districts. Using data from the National Assessment of Academic Ability in 2022, we first identified public elementary schools in each of Chiba Prefecture's five school districts with

[2]This approach was also consistent with the institutional characteristics of the Japanese teacher labor market, where mid-career hiring is limited and public schools rely primarily on newly graduated candidates.

relatively low academic performance. Within each district, we then randomly assigned four schools to receive a part-time mathematics specialist teachers, four schools to receive a part-time science specialist teachers, and four schools to serve as controls without specialist deployment, resulting in 12 schools per district. In total, the sample comprised 20 mathematics treatment schools, 20 science treatment schools, and 20 control schools. Importantly, mathematics and science specialist teachers were deployed in separate schools, and no school received specialist teachers in both subjects simultaneously. Note that the schools included in the intervention were relatively low-performing compared with both the Chiba Prefecture and national averages. Based on the National Assessment of Academic Ability, the average achievement level of the treatment sample was approximately 0.3–0.4 SD below the prefectural and national averages. Thus, our results estimated are likely more applicable to relatively low-performing schools.

The deployment of subject-specialist teachers was largely completed during the 2023 academic year, although implementation was not immediate at the start of the academic year. While deployment formally began in April 2023, only 55 % of treatment schools had their assigned subject-specialist teachers in place at the beginning of the academic year. By the time at the baseline assessment, part-time subject specialists had been successfully deployed in 65 % of mathematics schools (13 schools) and 75 % of science schools (15 schools). By the end of the academic year, implementation rates had risen to 93 % in mathematics schools and 98 % in science schools (The deployment of subject-specialist teachers was delayed until October in the two schools with the latest). We examine the sensitivity of our estimates to the timing of specialist deployment in Appendix Table A3.

4 Data

4.1 Baseline and Endline Assessments

The intervention targeted fourth-grade studentsChiba Prefecture deployed subject-specialist teachers in not only fourth-grade but also third-grade classrooms. However, the impact evaluation focuses on fourth-grade students, for whom comparable standardized achievement measures were available.. Student-level outcomes are drawn from standardized achievement tests and surveys administered to fourth-grade students during the 2023 academic year. Standardized achievement tests in reading, mathematics, and science, along with student surveys, were conducted during regular class hours. The survey includes validated psychological scales, allowing for the assessment of non-cognitive skills.

Fifth-grade students in the same schools, who were not part of the experiment, took exactly the same standardized achievement tests and student questionnaires as fourth graders. This design serves two purposes: to conduct placebo tests and to benchmark the intervention against alternative policies, such as class size reduction (see, Section 8).

Teacher-level information was obtained through a questionnaire developed by the authors and distributed to homeroom teachers of the 111 classes included in the evaluation; 91 teachers responded, yielding a response rate of 82 %.

Baseline and endline assessments were conducted approximately one year apart to measure the impact of the intervention. The baseline standardized achievement tests were administered in May 2023, immediately prior to the deployment of part-time subject-specialist teachers, and the endline assessments took place in May 2024. Data from two schools could not be used because student identifiers were inconsistently assigned across the two waves due to administrative errors, preventing accurate longitudinal matching. Appendix Table A1 compares baseline achievement between the two excluded schools and the remaining sample. We find no statistically significant differences in baseline test scores, suggesting that attrition at the school level is unlikely to bias the main results. After excluding these schools, the response rate was 98.5 % at baseline and 95.2 % at endline, calculated as the number of participating students divided by total enrollment as of April 2, 2023. Table 1 reports the final analytic sample by treatment status.

To complement the quantitative data, we conducted semi-structured qualitative interviews with three pairs of homeroom teachers and subject-specialist teachers, who agreed to participate among those randomly selected for interviews. Each interview lasted approximately one hour and was conducted on June 21, September 25, and October 30, 2023, by representatives of the Chiba Prefectural Board of Education and three of the authors (Igei, Le, and Nakamuro). The interviews covered the division of instructional responsibilities, coordination between homeroom and specialist teachers, and perceived changes in workload and classroom practice. We use these interviews to contextualize the quantitative estimates of teacher outcomes and coordination costs (see in Section 6).

4.2 Outcomes

Student achievements are the primary outcome of interest and is measured using standardized test scores. Mathematics, reading, and science scores are standardized to have a mean of zero and a standard deviation of one. Although subject-specialist teachers were deployed only in mathematics and science, reading outcomes are included to examine potential spillover effects, as reduced instructional burdens on homeroom teachers may have increased preparation time for other subjects.

In addition to cognitive outcomes, this study examines learning strategies and non-cognitive skills. Learning strategies, defined as deliberate actions that enhance learning effectiveness, are measured using scales developed by Sato and Arai (1998), with higher scores indicating more mature strategies. Non-cognitive skills are assessed through self-control and self-efficacy scales, each consisting of eight items. The self-control scale is based on Tsukayama et al. (2013), and the self-efficacy scale is based on Pintrich et al. (1991). Higher scores indicate lower self-control and higher self-efficacy.

Teacher outcomes are measured using a survey administered by the Chiba Prefectural Board of Education at the endline assessment. The survey collected information on demographic characteristics, teaching experience, working hours, lesson preparation time, perceived workload, instructional practices, mental health, and growth mindset.

4.3 Descriptive Statistics and Balance Check

Baseline characteristics are well balanced across treatment and control groups prior to deployment of subject-specialist teachers. Table 2 presents descriptive statistics for student-level variables, reporting mean values for mathematics treatment schools (Column 1), science treatment schools (Column 2), and control schools (Column 3). Columns (4) and (5) display mean differences between each treatment group and the control group, respectively. No statistically significant differences are observed in baseline achievement in mathematics, science, or reading. Likewise, student demographics, learning strategies, and non-cognitive skill measures are comparable across groups. These results support the validity of the random assignment and suggest that pre-treatment differences are unlikely to confound the estimated effects.

5 Empirical Strategy

We estimate the causal effect of subject-specialist deployment on student achievement using a value-added (VA) framework. The primary specification estimates the Intention-to-Treat (ITT) effect by regressing endline test scores on baseline achievement, treatment assignment indicators, and student-level controls, as shown in Equation 1.

$$Y_{ist} = \alpha + \beta\, Y_{ist-1} + \tau_1\, D_{is}^{T1} + \tau_2\, D_{is}^{T2} + \boldsymbol{\gamma}\mathbf{X}_{is} + \varepsilon_{ist} \tag{1}$$

The dependent variable, Y_{ist}, denotes student i's achievement in school s at the endline assessment, while Y_{ist-1} captures the baseline achievement. The indicators D_{is}^{T1} and D_{is}^{T2} represent assignment to mathematics and science specialist teachers, respectively, and $\mathbf{X}_{is}$ includes students demographic characteristics, such as gender, class size (the total numbers of enrolled students divided by the number of classes in the fourth grade), and randomization strata fixed effects.

Because deployment of subject-specialist teachers was not fully implemented at baseline, we additionally estimate Local Average Treatment Effects (LATE) to account for partial compliance. As reported in Table 1, only 13 mathematics schools and 15 science schools had specialists in place by baseline testing. We therefore use random assignment to treatment as an instrumental variable for actual deployment, implementing a standard two-stage least squares (2SLS) approach. In the first stage, actual deployment indicators are regressed on assignment dummies and control variables (Equations 2–3).

$$D_{is}^{T1} = \pi_1 \, Z_{is}^{T1} + \pi_2 \, Z_{is}^{T2} + \boldsymbol{\pi}_3 \mathbf{X}_{is} + u_{is} \tag{2}$$

$$D_{is}^{T2} = \rho_1 \, Z_{is}^{T1} + \rho_2 \, Z_{is}^{T2} + \boldsymbol{\rho}_3 \mathbf{X}_{is} + v_{is} \tag{3}$$

The second stage replaces actual deployment with its predicted value to estimate Equation 4. By instrumenting actual specialist deployment with randomized assignment, the 2SLS estimates recover the causal effect of the intervention under imperfect compliance.

$$Y_{ist} = \alpha + \beta \, Y_{ist-1} + \gamma_1 \, \widehat{D}_{is}^{T1} + \gamma_2 \, \widehat{D}_{is}^{T2} + \boldsymbol{\gamma} \mathbf{X}_{is} + \varepsilon_{ist}. \tag{4}$$

6 Results

6.1 Intention-to-Treat Effect

The ITT estimates indicate that the deployment of science specialist teachers significantly improved student achievement, while mathematics specialist deployment produced no detectable effects. Table 3 presents the value-added (VA) estimates, controlling for strata fixed effects and clustering standard errors at the school level. The coefficient on D_{is}^{T2} is positive and statistically significant for both science and mathematics achievements. Students in schools assigned a science specialist teachers experienced gains of 0.17 SD in science and 0.12 SD in mathematics. The latter suggests a spillover effect from science to mathematics, consistent with prior evidence that improvements in science proficiency can reinforce mathematical performance (Judson & Sawada, 2000). No spillover effects are observed for reading achievement.

The positive effects of science specialization extend modestly to certain learning strategies. In particular, flexible learning strategies defined as students' ability to adjust their learning approaches to changing circumstances show statistically significant improvements. However, no significant changes are detected in non-cognitive skills. In contrast, assignment to a mathematics specialist (D_{is}^{T1}) yields no statistically significant effects on achievement in any subject, nor on learning strategies or non-cognitive outcomes. Taken together, these findings suggest that the benefits of specialization are subject-specific and do not generalize uniformly across subjects.

6.2 Local Average Treatment Effects

The LATE estimates confirm and slightly amplify the positive effects observed in the ITT analysis for science specialist deployment. As shown in Table 4, schools assigned to the science treatment experienced gains of 0.18 SD in science achievement and 0.13 SD in mathematics. In addition to cognitive outcomes, students in science treatment schools exhibited a 0.11 SD increase in flexible learning strategies and a 0.14 SD improvement in self-control. No comparable effects are observed in schools assigned to the mathematics specialist treatment. Consistent with the ITT results, neither treatment group shows statistically significant effects on reading achievement. The first-stage relationship between random assignment and actual deployment is strong: first-stage F-statistics exceed 35 in all specifications, well above conventional thresholds for instrument relevance, reflecting the high compliance rates. Overall, the LATE estimates closely mirror the ITT findings in both magnitude and statistical significance, and the discussion that follows focuses primarily on the LATE results.

6.3 Effects on Science Achievement by Domain

The domain-level results clarify how science specialists improved achievement. Table 5 decomposes science scores into five domains: two difficulty-based categories (basic and applied) and three competency-based categories corresponding to the evaluation criteria of Japan's national Course of Study.

Two patterns stand out. First, the gains are concentrated in cognitive domains. Science-specialist deployment significantly improves basic and applied items as well as the knowledge-and-skills domain, where the effect is largest (0.20 SD), and inquiry, reasoning, and expression (0.13 SD), indicating that the intervention enhanced not only basic proficiency but also higher-order reasoning skills. These improvements in underlying competencies may have contributed not only to higher science achievement but also to the observed gains in mathematics: reasoning and inquiry skills cultivated through expert science instruction are plausibly transferable to mathematical problem solving, consistent with prior evidence that improvements in science proficiency reinforce mathematical performance (Judson & Sawada, 2000).

Second, and in contrast, the intervention leaves students' proactive attitude toward learning unchanged. This null result is consistent with the broader pattern in Tables 3 and 4: with the exception of a modest improvement in flexible strategies, specialist deployment does not shift students' learning strategies.

Taken together, these results indicate that the intervention operated through the transmission of subject content and reasoning skills rather than through changes in students' learning attitudes or strategies. This channel also accounts for the cross-subject spillover: had the intervention worked by altering students' learning attitudes or strategies, we would expect gains across all subjects, including reading; instead, the gains are confined to science and its cognitively adjacent subject, mathematics, while reading is unaffected.

Experienced specialists appear to have improved the quality of instruction itself, including the delivery of experimental and inquiry-based content that generalist homeroom teachers find difficult to teach. This interpretation is consistent with the view that the returns to specialization are governed by subject-specific production technologies: in science, expert instruction directly raises the acquisition of knowledge, skills, and reasoning, and these competencies transfer to nearby domains without requiring any change on the students' behavior.

6.4 Heterogeneity

The positive effects of science specialist deployment are broadly distributed and do not generate major equity concerns. We first examine heterogeneity by baseline achievement, gender, and socioeconomic status (SES). In science treatment schools, initially high-achieving students experience larger gains: although top-tercile students exhibit mean reversion, the interaction between treatment assignment and top baseline achievement is positive and statistically significant, offsetting the decline. This pattern suggests that specialization helped sustain performance among students who were already strong in science. See the result in Appendix Table A5.

By contrast, we find little systematic heterogeneity by gender or SES in science schools. Treatment effects do not differ significantly between boys and girls, nor across socioeconomic backgrounds. In mathematics schools, no meaningful heterogeneity emerges in achievement, although some shifts in learning strategies are observed for lower-SES students. Teacher experience is positively but modestly associated with science gains. Overall, the absence of pronounced differential effects—particularly in science treatment schools—suggests that the intervention's benefits are broadly shared and unlikely to exacerbate existing inequalities. See the results in Appendix Tables A4, A6, and A7.

6.5 Effects on Homeroom Teachers

The intervention had limited effects on homeroom teachers' workload and instructional practices, but modest positive effects on teachers' mindsets. We use data from the February 2024 teacher survey to estimate these impacts. Balance tests (Table 6) indicate some pre-treatment differences: teachers in treatment schools tend to have longer teaching experience, and science schools have a higher share of teachers with a university degree. These characteristics are included as controls in all subsequent specifications.

Table 7 reports the estimated effects. The deployment of subject-specialist teachers did not significantly change homeroom teachers' working hours, lesson preparation time, or perceived workload. Qualitative interviews conducted with homeroom and specialist teachers provide context for these findings. Several science specialists reported that their part-time roles allowed them to devote more time to lesson planning and inquiry-based activities than during full-time employment. At the same time, homeroom teachers noted that coordination, particularly aligning assessments and scheduling, required additional effort, offsetting potential workload reductions. These field observations are consistent with the quantitative results, which show no measurable decline in homeroom teachers' time at school or perceived workload.

The intervention did not significantly affect homeroom teachers' mental well-being (WHO-5), but did significantly increase their growth mindset. Some evidence indicates that reducing the number of subjects taught may strengthen teachers' subject focus and professional alignment, with potential implications for retention and instructional quality (Backes et al., 2024, 2025; Bastian & Fortner, 2020). Our null effect on WHO-5 is consistent with the view that specialization does not automatically alleviate stress, particularly when evaluation responsibilities and coordination remain with homeroom teachers.

In contrast, the positive and statistically significant effect on teachers' growth mindset is notable. Yeager et al. (2022) show that teachers' growth-oriented beliefs are associated with stronger student achievement gains following interventions. Qualitative interviews conducted with homeroom teachers suggested that interaction with experienced specialist teachers may strengthen homeroom teachers' beliefs about the malleability of student ability, functioning as informal professional development. Thus, while partial specialization did not improve short-run well-being, it may have fostered a more growth-mindset, with potential longer-run implications for student outcomes.

On the other hand, instructional practices show minor adjustments: repetitive instruction declined in mathematics schools, and lecture-based instruction decreased in science schools. However, there is no corresponding increase in inquiry-based instruction. Overall, the presence of specialists did not substantially alter homeroom teachers' observable instructional behavior.

6.6 Effects on Teacher−Student Relationships and Classroom Behavior

A central concern in prior work is that specialization weakens teacher−student relationships and worsens classroom behavior. The evidence suggests that the partial specialization implemented in this study mitigates several risks associated with full departmentalization. Prior research, most notably Fryer Jr. (2018) and Hwang and Kisida (2022), highlights coordination costs and weakened student−teacher relationships as key mechanisms through which full departmentalization may reduce student achievements. We examine whether similar risks arise under our intervention.

First, we find no evidence that the intervention weakened student−teacher relationships or increased behavioral problems. Table 8 reports LATE estimates for survey-based measures of teacher−student relationships, behavioral problems, and twelve indicators of classroom discipline; none shows a significant deterioration in either treatment arm. Additionally, the self-control scale, which captures indicators such as forgetfulness, inattentiveness, and disruptive speech, shows improvement (Table 4). Measures of discipline, including punctuality, classroom organization, and respectful greetings, show no systematic deterioration.

Second, we find limited evidence that subject-specialist deployment reduced homeroom teachers' workload. Quantitative estimates show no statistically significant effects on working hours, lesson preparation time, or perceived burden, although the estimates for working hours in science schools are relatively precisely estimated and may suggest the presence of non-negligible coordination costs. Qualitative interviews further indicate that coordination costs persisted. In particular, because homeroom teachers remained responsible for grading and student evaluation, alignment between specialist and homeroom teachers required ongoing communication. Taken together, these findings suggest that the intervention did not substantially alleviate teachers' workload and that coordination costs may have remained considerable, although they did not escalate to a degree that undermined instructional outcomes.

7 Mechanism

7.1 Why Did Effects Emerge Only in Science?

The results so far establish that specialist deployment raised achievement in science but not in mathematics. An important question raised by our findings is why subject specialization improved achievement in science but not in mathematics. This subsection discusses several potential explanations for this asymmetry.

A first possibility is that science specialists were simply more qualified than mathematics specialists. However, this explanation is not supported by the data. Mathematics and science specialists did not differ significantly in any observable predetermined characteristics, including gender, age, years of teaching experience, educational attainment, teaching experience outside the schools, or possession of a secondary-school teaching license (see, Table 6). Therefore, differences in observable teacher characteristics are unlikely to account for the differential impacts across subjects.

A second possibility is that the estimated effects reflect differences in the quality of teacher-subject matching rather than specialization itself. In our setting, however, this concern is limited because specialists were recruited explicitly based on their expertise in science or mathematics. Subject mismatch was therefore minimized ex ante.

However, reallocating instruction to specialists may crowd out high-quality homeroom teachers if they themselves possess strong subject-specific expertise. Table 11 directly tests this mechanism by interacting specialist deployment with homeroom teachers' reported field of specialization.

Before interpreting the interaction term, one result from this specification deserves attention. Homeroom teachers' subject specialization is positively associated with student achievement: students assigned to science-specialized homeroom teachers perform significantly better in science, and a similar (though smaller) pattern is observed in mathematics. The science coefficient is notably larger, underscoring the role of comparative advantage in instructional assignment.

The interaction term suggests evidence of crowd-out in mathematics. When a mathematics specialist is deployed to a classroom led by a mathematics-specialized homeroom teacher, the positive effect associated with the homeroom teacher's expertise is offset. This is consistent with the prior research that specialization generates smaller gains when it replaces teachers who already possess strong subject-specific comparative advantage (Fryer Jr., 2018).

In contrast, the number of science-specialized homeroom teachers was very small (only 7.7% in our sample)[3], and no science specialists were assigned to classrooms led by science-specialized homeroom teachers. As a result, science expertise among homeroom teachers was not displaced. This distinction is critical: specialist deployment appears most effective when targeted to classrooms where homeroom teachers lack subject-specific expertise, thereby leveraging comparative advantage rather than displacing it.

7.2 Benchmarking Against Alternative Policies

Establishing that targeted specialization works in this setting leaves open how it compares with the alternatives actually available to policymakers. This section benchmarks the intervention against two: the nationally promoted departmentalization model currently implemented in Grade 5, and class size reduction, the most widely studied margin of school resource policy.

7.2.1 The Nationally Promoted Departmentalization Model in Grade 5

First, we compare our experimentally evaluated partial specialization with the nationally promoted departmentalization in Grade 5. The teacher specialization currently implemented in Grades 5 and 6 in Chiba Prefecture differs in important respects from our intervention.

Since 2022, Japan has formally promoted subject specialization in upper elementary grades, with implementation left to local discretion. Nationally promoted specialization is limited to a subset of subjects, such as foreign language, science, mathematics, and physical education, and schools retain discretion over whether to implement specialization in each subject. In our sample of

[3]These figures refer to Grade 4 homeroom teachers, who constitute the main analysis sample. We also collected teacher data for Grade 5, for which science specialists were similarly rare: only 7.5% homeroom teachers reported science as their field of specialization, suggesting that the scarcity of science specialists is representative rather than sample-specific.

60 schools, 55.2 % of fifth-grade classes employed a science specialist and 10.3 % a mathematics specialist.[4] Unlike our intervention, however, specialization is typically delivered by full-time teachers who teach the same subject across multiple classes and grades, substantially expanding the number of students each teacher is responsible for.

The Grade 5 model is closer to full departmentalization studied in Fryer Jr. (2018) and Hwang and Kisida (2022) than to the partial specialization evaluated in our experiment. As emphasized in Fryer Jr. (2018), the effectiveness of specialization depends not only on subject expertise but also on the institutional structure. The contrast between Grades 4 and 5 therefore provides a natural setting to examine how the effectiveness of teacher specialization depends on institutional design.

The results suggest that nationally implemented departmentalization does not improve academic achievement and may even generate negative spillovers across subjects. Using Grade 5 students who were not part of the randomized intervention, Table 9 reports OLS estimates of departmentalization on academic outcomes. We find no evidence of gains in overall achievement. Instead, in science-departmentalized schools, mathematics achievement declines by 0.14 SD[5]. These findings align with prior evidence that departmentalization alone does not guarantee improvements in student performance.

Consistent with this interpretation, the effects on learning processes and non-cognitive outcomes are limited. Table 10 shows that mathematics departmentalization increases students' use of planning strategies, but no systematic improvements are observed across other learning strategies or non-cognitive measures. Taken together, these results highlight that the effectiveness of specialization depends critically on institutional design, particularly the balance between gains in subject expertise and the organizational costs induced by reallocating instruction.

[4]Because Grade 5 departmentalization was adopted at schools' discretion rather than randomly assigned, schools with and without specialists differ in some predetermined characteristics. Tables 9 and 10 therefore control for these characteristics. The results of the balance tests are available upon request.

[5]Tables 9 and 10 use the combined sample of the 60 experimental schools and the 39 additional schools that voluntarily participated in the survey, thereby increasing the sample size to 99 schools. As a robustness check, we re-estimate the specifications using only the 60 experimental schools. The results are quite similar. In particular, the coefficient on Science-departmentalization schools for mathematics achievement remains negative and statistically significant (-0.172, s.e. = 0.052).

7.2.2 Cost-Effectiveness Relative to Class Size Reduction

Second, we show that the intervention is substantially more cost-effective than class size reduction, which is one of the most extensively studied and widely implemented education reforms. Using the same data with Table 9, we find that reducing class size by ten students increases science achievement by approximately 0.09 SD, consistent with prior Japanese evidence (Ito et al., 2020). This comparison underscores that the relevant policy question concerns not only whether an intervention works, but how efficiently it translates resources into achievement gains.

Appendix B.1 quantifies these differences. The cost of achieving a one-standard-deviation improvement in science achievement is approximately 79,413 JPY (539 USD) under our intervention, compared with 966,446 JPY (6,555 USD) under a class size reduction policy. These estimates indicate that our intervention delivers achievement gains at a fraction of the cost of reducing class size.

More generally, Kraft (2020) proposes a scheme for categorizing effect size and cost based on nearly 747 randomized controlled trials. The LATE estimate of a 0.18 SD increase in science test scores through deploying science part-time specialists falls into the upper bound of a ”medium” effect size, which is particularly meaningful for a large-scale randomized controlled trial. Additionally, the cost per student was 14,135 JPY (96 USD) for deploying part-time specialists. Under the framework by Kraft (2020), interventions costing under 500 USD per student are classified as low cost. Hence, our intervention yields an upper bound of a medium effect size through an exceptionally low cost.

8 Robustness Check

This section addresses four threats to the interpretation of our estimates: differential attrition, the timing of specialist deployment, unobserved school-level shocks, and the possibility that the estimated effects reflect additional resources or specialist teacher quality rather than specialization itself.

8.1 Attrition

Attrition does not threaten the internal validity of the estimates. Appendix Table A2 reports results from a model in which the dependent variable equals one if a student attrited at baseline or endline and zero otherwise. Although lower baseline achievement predicts attrition, assignment to a subject-specialist teacher is not statistically associated with dropout from the sample. These findings suggest that differential attrition across treatment groups is unlikely to bias the main results.

8.2 Timing of Deployment

The timing of specialist deployment may matter. Appendix Table A3 estimates a LATE specification using an instrument equal to one if the specialist was assigned at the begining of academic year and zero if assignment occurred later. The estimated effects for science treatment schools are larger than those reported in Table 3, although the difference is statistically insignificant (Wald t test = 0.61, $p > 0.05$).

8.3 Unobserved School-level Shocks

The placebo test provides no evidence of spurious school-level effects. Although the intervention targeted fourth-grade students, standardized tests were also administered to fifth-grade students in the same schools who were not exposed to the treatment. We estimate the effect of treatment assignment at the school level on fifth-grade outcomes by regressing achievement on a dummy equal to one if the school's fourth-grade cohort had been assigned to either the mathematics or science treatment group.

Table 13 shows no statistically significant effects for either treatment. These results suggest that the estimated gains for fourth-grade students are not driven by unobserved school-level characteristics or concurrent reforms, reinforcing the internal validity of the main findings.

8.4 Resources or Specialization?

A potential concern is that the estimated treatment effects may simply reflect an increase in school resources rather than the effect of subject specialization itself, because treatment schools received an additional teacher. However, in the Japanese public elementary school system, instructional hours are centrally regulated by the national curriculum guidelines and are highly rigid across schools and grades. Even when additional teachers are assigned, the total number of instructional hours received by students does not increase. The intervention therefore did not expand students' exposure to classroom instruction, but instead altered the allocation of instructional responsibilities across teachers. In this sense, the estimated effects should be interpreted as arising from changes in instructional organization and subject-specific teacher assignment rather than from increased instructional time or additional school resources per se.

8.5 Specialist Experience

Another potential concern is that the observed effects may reflect differences in teacher quality rather than the organizational benefits of subject specialization. Specialist teachers were considerably more experienced than regular homeroom teachers, with average teaching experience of 27.08 years (s.e. = 2.421) versus 11.56 years (s.e. = 0.770), respectively. If experience improves instructional effectiveness, the estimated treatment effects could overstate the contribution of specialization. To address this concern, we re-estimated the main specifications including homeroom teachers' years of teaching experience as an additional control variable (Table 12). The result of Science schools remain positive and statistically significant, and if anything somewhat larger after controlling for teacher experience, suggesting that the results are not driven solely by the greater experience of specialist teachers. Rather, the findings are more consistent with the interpretation that aligning instruction with subject-specific expertise and comparative advantage improves student achievement.

9 Conclusion and Policy Implications

This paper provides experimental evidence that subject specialization in elementary schools improves student achievement when its design preserves teacher comparative advantage. In a cluster-randomized experiment across 60 Japanese public elementary schools, assigning experienced part-time science specialists raised science achievement by 0.18 SD and mathematics by 0.13 SD, while mathematics specialization generated no measurable gains. These effects emerged without increasing coordination costs, raising teacher workload, or weakening teacher-student relationships, the costs that prior studies attribute to full departmentalization.

The pattern of effects identifies the channel through which specialization operates. The intervention changed remarkably little on the surface: homeroom teachers' working hours, lesson preparation time, and instructional practices were essentially unchanged, and students' learning strategies, attitudes toward learning, and non-cognitive skills showed little movement. What improved was achievement itself, concentrated in knowledge, skills, and reasoning within science and its cognitively adjacent subject, mathematics. This configuration is difficult to reconcile with explanations based on motivation, study behavior, or reduced teacher burden. It points instead to subject-specific production technologies: in science, where instruction requires specialized preparation and experimental expertise, expert instruction directly raises the acquisition of knowledge and skills without requiring any change on the part of students or homeroom teachers. Two further results discipline this interpretation. First, the gains are not explained by teacher experience alone: controlling for homeroom teachers' years of experience leaves the estimates essentially unchanged, and mathematics specialists, who were observationally similar to science specialists, produced no gains. Second, the asymmetry between subjects tracks the pre-existing distribution of expertise. Mathematics specialists displaced homeroom teachers who were themselves math-specialized, offsetting the returns; science specialists filled a genuine gap, since science expertise among homeroom teachers was scarce. Comparative advantage, rather than specialist quality or formal specialization per se, governs the returns to reallocation.

Three policy implications follow. First, teacher assignment should preserve comparative advantage. Specialists should be deployed to classrooms where homeroom teachers lack subject-specific expertise, rather than replacing instruction by teachers whose expertise already aligns with the subject. Designing specialization policies without regard to the existing distribution of teacher strengths risks crowding out effective instruction, as the null results for mathematics specialization and for the nationally promoted Grade 5 model illustrate. Second, institutional design matters as much as expertise. The partial specialization evaluated here preserved the homeroom structure and introduced specialists in a single subject per school; the Grade 5 model, which is closer to full departmentalization, produced no gains and some evidence of negative cross-subject spillovers. Third, rehiring experienced retired teachers is a cost-effective response to subject-specific skill constraints. This staffing model delivered a one-standard-deviation gain in science achievement at roughly one-twelfth the cost of achieving the same gain through class size reduction, and appears to have generated positive professional spillovers to homeroom teachers' growth mindsets. Amid deepening teacher shortages, rehiring experienced educators aligned with subject expertise offers a margin of adjustment that conventional recruitment channels cannot.

Several limitations qualify these conclusions. First, we estimate short-run achievement effects; impacts on longer-run outcomes such as persistence in STEM, secondary school performance, and teacher retention remain unknown. Second, while coordination costs did not offset gains in this setting, they may bind under broader or differently structured forms of departmentalization, as the null-to-negative results for the Grade 5 model suggest. Third, the comparative-advantage interpretation rests on the variation our experiment happened to generate, and two counterfactual cells are unobserved. No science specialist was assigned to a classroom led by a science-specialized homeroom teacher, so we cannot directly verify that science specialization would also crowd out expertise where it exists; conversely, we do not observe mathematics specialists deployed exclusively to classrooms lacking math expertise, so we cannot confirm that targeted mathematics specialization would have generated gains. The prediction that deployment targeted on comparative advantage maximizes returns is therefore an extrapolation, albeit one consistent with all the observed cells. Testing it directly through deployment rules conditioned on homeroom teacher expertise is a natural next experiment.

The policy question, then, is not whether elementary schools should specialize, but when: in which subjects, and in whose classrooms. Targeted deployment of experienced specialists, in subjects marked by low teacher confidence and high preparation demands, and in classrooms where homeroom teachers lack subject expertise, can generate meaningful achievement gains without organizational frictions. Aligning instructional assignments with teacher comparative advantage emerges as the central design principle for school workforce reform.

References

Asada, E., & Nakanishi, K. (2018). *Chuuichi gyappu no kaishou ni muketa shougakkou ni okeru kyoukatanninsei no kouka no kentou - gakkyuu tanninsei de manabu jidou tono ishikichousa no hikaku wo tooshite. A Study on the Effects of Departmentalization in Elementary Schools to Resolve the "Grade 7 Gap": A Comparison by Attitude Survey with Pupils Taught under the Homeroom-Teacher System* [Doctoral dissertation, International Pacific University] [In Japanese. Available at https://cir.nii.ac.jp/crid/1390853649310018816].

Backes, B., Cowan, J., & Goldhaber, D. (2024). Departmentalized instruction and elementary school effectiveness. working paper no. 298-0424. *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*.

Backes, B., Cowan, J., & Goldhaber, D. (2025). Building bridges to middle school? elementary school departmentalization and academic achievement in the upper grades. working paper no. 315-0325. *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*.

Bastian, K. C., & Fortner, C. K. (2020). Is less more? subject-area specialization and outcomes in elementary schools. *Education Finance and Policy*, *15*(2), 357–382.

Bastian, K. C., Fortner, C. K., & Caton, K. (2023). Subject-area specialization and teacher retention: An elementary school story. *The Elementary School Journal*, *124*(2), 343–366.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly journal of economics*, *107*(4), 1137–1160.

Chiba Prefecture. (2022, December). *Chiba kenmin dayori reiwa 4-nen 12 gatsu-gō. chiba prefectural newsletter december 2022* (tech. rep.) (https://www.pref.chiba.lg.jp/kouhou/kenmindayori/r4/r4-12/index.html). Chiba Prefecture. Chiba, Japan.

Chiba Prefecture Board of Education. (2025, October). *Kouritsu gakkou rinjiteki nin'yō kyōshokuin / kaikei-nen'yō nin'yō shokuin / ninki-tsuki kyōshokuin no tōroku ni tsuite —taigū ni tsuite. conditions of employment for temporary and fixed-term teaching staff at public schools in chiba prefecture* (tech. rep.) (https://www.pref.chiba.lg.jp/kyouiku/syokuin/saiyou/koushi/taiguu.html). Chiba Prefecture Board of Education. Chiba, Japan.

Fryer Jr., R. G. (2018). The "pupil"factory: Specialization and the production of human capital in schools. *American Economic Review*, *108*(3), 616–656.

Hwang, N., & Kisida, B. (2022). Spread too thin: The effect of specialization on teaching effectiveness. *Educational Evaluation and Policy Analysis*, *44*(4), 593–607.

Ito, H., Nakamuro, M., & Yamaguchi, S. (2020). Effects of class-size reduction on cognitive and non-cognitive skills. *Japan and the World Economy*, *53*, 100977.

Judson, E., & Sawada, D. (2000). Examining the effects of a reformed junior high school science class on students' math achievement. *School Science and Mathematics*, *100*(8), 419–425.

Kajita, H., & Doho, A. (2022). Shogakkou ni okeru kyoukatanninsei ni muketeno kongo - jissenkou ni okeru ishikichousa kara. toward the introduction of the departmentalization in elementary schools: From a survey of mindsets in practicing schools [In Japanese. Available at https://hijiyama-u.repo.nii.ac.jp/records/371]. *Review of Teacher Education Curriculum, Hijiyama University / Hijiyama University Junior College*, *8*, 24–33.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, *49*(4), 241–253.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2021, June). *Hyōjun jugyō jisū ni tsuite. overview of standard class hours* (tech. rep.) (https://www.mext.go.jp/content/20210629-mxt_kyoiku01-000016453_4.pdf). Ministry of Education, Culture, Sports, Science and Technology (MEXT). Tokyo, Japan.

Nakajima, A., & Kusaka, T. (2020). Shogakkou kyouin no rika shidou ni taisuru fuan, kyoushi kouryokukan, gakushudouki no kentou. a study of elementary school teachers' anxiety about science instruction, teacher efficacy, and learning motivation [In Japanese. Available at: https://doi.org/10.15077/jjet.43044]. *Japan Journal of Educational Technology*, *44*(1), 1–9.

Pintrich, P., Smith, D., Garcia, T., & McKeachie, W. (1991). *A manual for the use of the motivated strategies for learning questionnaire (mslq)* (Technical Report No. NCRIPTAL-91-B-004) (ERIC Document Reproduction Service No. ED338122 / Contract OERI-86-0010). National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan. Ann Arbor, MI. https://files.eric.ed.gov/fulltext/ED338122.pdf

Sato, J., & Arai, H. (1998). Gakushu houryaku no shiyou to tassei mokuhyou oyobi genin kizoku tono kankei. the relation between the use of learning strategies learning goals and causal attributions [In Japanese]. *Tsukuba Psychological Research*, *20*, 115–124. https://cir.nii.ac.jp/crid/1050282677523203456

Tokoro, S., Suzuki, K., Tamura, T., & Kato, M. (2022). *Shogakkou kyoukatanninsei no unyou ni kansuru ichiteiann: Iwateken ni okeru kyoukatanninsei no genjyouhaaku to kadaikaiketsu no tedaten no tansaku kara. A Proposal for the Operation of the Departmentalization in Elementary Schools: From Surveying the Current State and Exploring Solutions in Iwate Prefecture* [Doctoral dissertation, Iwate University] [In Japanese. Available at https://iwate-u.repo.nii.ac.jp/records/15669].

Tsukayama, E., Duckworth, A. L., & Kim, B. (2013). Domain-specific impulsivity in school-age children. *Developmental Science*, *16*(6), 879–893.

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., Muller, C., Murray, J., Mhatre, P., Kersting, N., et al. (2022). Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychological science*, *33*(1), 18–32.

Yomiuri Shimbun. (2025). Byoki ya kazoku no kanbyou nado de yasumu kyouin wo kaba-, kyouinobra wo tanki haken suru shikumi dukuri he... eikoku wo sankou ni. creating a system to hire retired teachers short-term to cover teachers who are absent due to illness or caring for family members —inspired by the system in the uk [https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20250813-OYT1T50004/]. *The Yomiuri Shinbun*.

Table 1: Sample Composition and Treatment Implementation

	Mathematics	Science	Control
Assigned Schools	20	20	20
Classrooms (Homeroom Teachers)	40	36	35
Schools with Specialist Assigned by April	8	14	–
Schools with Specialist Assigned by Baseline (May)	13	15	–
Schools with Baseline Data	19	20	20
Students with Baseline Data	989	852	913
Schools with Endline Data	18	19	20
Students with Endline Data	977	764	913
Students Enrolled (as of April 2)	1,021	852	913

Table 2: Baseline Characteristics and Mean Differences

	Mathematics		Science		Control		Difference	
	Mean	N	Mean	N	Mean	N	Math–Ctrl	Sci–Ctrl
Panel A. Academic Achievement								
Science (std.)	0.138	983	0.239	854	0.154	910	-0.016	0.085
Mathematics (std.)	0.002	988	0.063	861	-0.013	914	0.015	0.076
Japanese (std.)	-0.037	986	0.077	861	0.002	913	-0.040	0.074
Panel B. Learning Strategies								
Flexible	0.033	975	0.031	854	0.038	897	-0.005	-0.007
Planning	0.029	974	0.007	857	0.022	896	0.007	-0.014
Task-Oriented	0.099	973	0.016	855	0.027	897	0.072	-0.011
Cognitive	-0.000	975	0.000	859	0.014	902	-0.015	-0.014
Effort Regulation	-0.006	975	0.023	858	0.029	897	-0.036	-0.006
Panel C. Non-Cognitive Skills and Demographics								
Self-Efficacy	0.048	979	0.094	859	0.103	905	-0.055	-0.009
Self-Control	-0.032	986	0.040	859	-0.003	901	-0.028	0.043
Male (share)	0.488	997	0.503	869	0.467	920	0.021	0.035
Panel D. School Characteristics								
Students per School	51.050	20	42.600	20	45.650	20	5.400	-3.050
3rd Grade Classes	1.900	20	1.850	20	1.800	20	0.100	0.050
4th Grade Classes	1.950	20	1.650	20	1.650	20	0.300	0.000

Notes.

Mean differences and inference. Columns (1)–(3) report mean values for mathematics treatment, science treatment, and control schools, respectively. The last two columns report mean differences relative to the control group (Mathematics–Control and Science–Control). Mean difference tests control for strata fixed effects and the number of classes per school, with standard errors clustered at the school level. ***, **, and * denote statistical significance at the 0.1 percent, 1 percent, and 5 percent levels, respectively.

Learning strategies. Learning strategies capture deliberate behaviors that enhance learning effectiveness and include five domains: (i) **Flexible**—adapting study methods to one's circumstances; (ii) **Planning**—systematic approaches to studying; (iii) **Task-oriented**—engagement in concrete learning activities (e.g., note-taking); (iv) **Cognitive**—efforts to deepen conceptual understanding; and (v) **Effort regulation**—managing emotions to sustain motivation.

Non-cognitive measures. Non-cognitive skills are assessed through self-control and self-efficacy indices, each consisting of eight items. Each index is constructed by averaging responses on a 5-point Likert scale and standardizing the resulting score to have mean 0 and variance 1 in the analytic sample. Higher values indicate lower self-control and higher self-efficacy, respectively. The self-control scale is based on the Japanese translation of Tsukayama et al. (2013), and the self-efficacy measure corresponds to the "Self-Efficacy for Learning and Performance" subscale from Pintrich et al. (1991).

Table 3: Intention-To-Treat (ITT) Estimates on Student Outcomes

	Achievement			Learning Strategies					Non-cognitive Skills	
	Science (std.)	Math (std.)	Reading (std.)	Flexible strategy	Planning strategy	Task-oriented	Cognitive strategy	Effort regulation	Self-efficacy	Self-control
D^{T1} (Math schools)	-0.038	0.015	0.022	0.070	0.046	0.057	0.027	-0.063	-0.019	0.014
	(0.071)	(0.064)	(0.046)	(0.046)	(0.065)	(0.058)	(0.071)	(0.078)	(0.058)	(0.072)
D^{T2} (Science schools)	0.166**	0.119*	0.052	0.103*	0.069	0.110	0.126	-0.043	0.032	-0.132
	(0.060)	(0.053)	(0.047)	(0.051)	(0.066)	(0.056)	(0.064)	(0.066)	(0.061)	(0.067)
Y_{it-1}	0.214***	0.722***	0.721***	0.421***	0.421***	0.428***	0.407***	0.449***	0.481***	0.241***
	(0.003)	(0.027)	(0.034)	(0.028)	(0.024)	(0.024)	(0.026)	(0.023)	(0.024)	(0.029)
Strata fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covariates	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	2,395	2,394	2,402	2,334	2,334	2,333	2,203	2,335	2,350	2,353
R^2	0.096	0.513	0.531	0.183	0.184	0.200	0.167	0.203	0.238	0.098

Note: ***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively. Standard errors are clustered at the school level. In addition to strata fixed effects, the regressions control for student gender, number of classes, and total student enrollment as covariates.

Learning strategies capture deliberate behaviors that enhance learning effectiveness and include five domains: (i) **Flexible**—adapting study methods to one's circumstances; (ii) **Planning**—systematic approaches to studying; (iii) **Task-oriented**—engagement in concrete learning activities (e.g., note-taking); (iv) **Cognitive**—efforts to deepen conceptual understanding; and (v) **Effort regulation**—managing emotions to sustain motivation.

Non-cognitive measures are assessed through self-control and self-efficacy indices, each consisting of eight items. Each index is constructed by averaging responses on a 5-point Likert scale and standardizing the resulting score to have mean 0 and variance 1 in the analytic sample. Higher values indicate lower self-control and higher self-efficacy, respectively. The self-control scale is based on the Japanese translation of Tsukayama et al. (2013), and the self-efficacy measure corresponds to the "Self-Efficacy for Learning and Performance" subscale from Pintrich et al. (1991).

Table 4: Local Average Treatment Effect (LATE) Estimates on Student Outcomes

	Achievement			Learning Strategies					Non-cognitive Skills	
	Science (std.)	Math (std.)	Reading (std.)	Flexible strategy	Planning strategy	Task-oriented	Cognitive strategy	Effort regulation	Self-efficacy	Self-control
D^{T1} (Math schools)	-0.047	0.018	0.023	0.080	0.053	0.066	0.030	-0.074	-0.022	0.018
	(0.082)	(0.074)	(0.054)	(0.052)	(0.075)	(0.068)	(0.081)	(0.088)	(0.067)	(0.084)
D^{T2} (Science schools)	0.178**	0.127*	0.051	0.110*	0.073	0.117	0.134	-0.045	0.034	-0.142*
	(0.059)	(0.053)	(0.050)	(0.055)	(0.071)	(0.062)	(0.071)	(0.069)	(0.064)	(0.070)
Y_{it-1}	0.274***	0.722***	0.726***	0.421***	0.421***	0.428***	0.407***	0.448***	0.481***	0.241***
	(0.054)	(0.026)	(0.033)	(0.028)	(0.023)	(0.023)	(0.026)	(0.023)	(0.023)	(0.028)
Strata fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covariates	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
First-stage F	35.37	39.48	37.98	44.71	44.83	44.34	44.68	43.73	43.94	43.70
Diff-in-Sargan statistic for exogeneity (p-value)	0.646	0.642	0.668	0.620	0.530	0.412	0.457	0.340	0.225	0.979
N	2,395	2,394	2,399	2,334	2,334	2,333	2,203	2,335	2,350	2,353
R^2	0.098	0.514	0.532	0.183	0.183	0.199	0.166	0.204	0.238	0.099

Note: ***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively. Standard errors are clustered at the school level. In addition to strata fixed effects, the regressions control for student gender, number of classes, and total student enrollment as covariates.

Learning strategies capture deliberate behaviors that enhance learning effectiveness and include five domains: (i) **Flexible**—adapting study methods to one's circumstances; (ii) **Planning**—systematic approaches to studying; (iii) **Task-oriented**—engagement in concrete learning activities (e.g., note-taking); (iv) **Cognitive**—efforts to deepen conceptual understanding; and (v) **Effort regulation**—managing emotions to sustain motivation.

Non-cognitive measures are assessed through self-control and self-efficacy indices, each consisting of eight items. Each index is constructed by averaging responses on a 5-point Likert scale and standardizing the resulting score to have mean 0 and variance 1 in the analytic sample. Higher values indicate lower self-control and higher self-efficacy, respectively. The self-control scale is based on the Japanese translation of Tsukayama et al. (2013), and the self-efficacy measure corresponds to the "Self-Efficacy for Learning and Performance" subscale from Pintrich et al. (1991).

Table 5: LATE Estimates on Science Achievement by Domain

	Difficulty-Based Domains		**Competency-Based Evaluation**		
	Basic	Applied	Knowledge and Skills	Inquiry, Reasoning, and Expression	Proactive Attitude toward Learning
D^{T1} (Math Schools)	-0.035	-0.071	-0.039	-0.045	-0.010
	(0.084)	(0.072)	(0.085)	(0.074)	(0.072)
D^{T2} (Science Schools)	0.172**	0.157*	0.201***	0.131*	0.096
	(0.059)	(0.064)	(0.058)	(0.065)	(0.057)
Y_{it-1}	0.260***	0.170***	0.256***	0.214***	0.150***
	(0.050)	(0.039)	(0.048)	(0.047)	(0.036)
Strata fixed effects	YES	YES	YES	YES	YES
Covariates	YES	YES	YES	YES	YES
F-statistic	35.35	35.47	35.27	35.47	35.23
P-value	0.805	0.837	0.462	0.710	0.578
N	2,395	2,395	2,395	2,395	2,395
R^2	0.093	0.046	0.093	0.060	0.027

Note: Columns are grouped into (i) difficulty-based domains (Basic and Applied), which classify items by level of difficulty, and (ii) competency-based domains (Knowledge and Skills; Inquiry, Reasoning, and Expression; Proactive Attitude toward Learning), which correspond to the competency framework in the national curriculum guidelines. ***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively. Standard errors are clustered at the school level. All regressions include strata fixed effects and control for student gender, number of classes, and total enrollment.

Table 6: Baseline Teacher Characteristics and Mean Differences

	Mathematics		Science		Control		Difference	
	Mean	N	Mean	N	Mean	N	Math–Ctrl	Sci–Ctrl
Female (%)	0.559	34	0.409	22	0.457	35	0.102	-0.048
Age (years)	37.588	34	38.909	22	34.686	35	2.903	4.223
Teaching experience (years)	12.647	34	14.364	22	8.743	35	3.904*	5.621**
Non-teaching work experience (yes = 1)	0.147	34	0.136	22	0.229	35	-0.082	-0.092
Bachelor's degree or higher (yes = 1)	0.941	34	1.000	22	0.914	35	0.027	0.086*
Secondary teaching license (yes = 1)	0.353	34	0.545	22	0.514	35	-0.161	0.031

Note: Columns report mean values for homeroom teachers in mathematics treatment, science treatment, and control schools, respectively. The last two columns report mean differences relative to the control group (Mathematics–Control and Science–Control). ***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table 7: LATE Estimates on Homeroom Teachers Outcomes

	Work Conditions			**Instructional Practices**			**Psychological Measures**	
	Working Hours	Lesson Preparation	Perceived Workload	Lecture-Based	Repetitive Practice	Inquiry-Based	WHO-5 Mental Health	Growth Mindset
D^{T1} (Math Schools)	0.082	-0.232	-0.047	-0.272	-0.412*	-0.066	0.995	0.836**
	(0.273)	(0.423)	(0.090)	(0.148)	(0.161)	(0.135)	(1.663)	(0.290)
D^{T2} (Science Schools)	0.602	0.189	0.065	-0.370*	-0.159	0.041	-0.434	0.747**
	(0.353)	(0.389)	(0.072)	(0.160)	(0.154)	(0.113)	(1.629)	(0.271)
First-stage F	43.59	43.50	40.67	92.70	43.70	43.59	42.78	42.78
Diff-in-Sargan statistic for exogeneity (p-value)	0.168	0.152	0.993	0.403	0.109	0.090	0.970	0.722
N	91	90	88	90	89	91	90	90
R^2	0.137	0.103	0.181	0.084	0.147	0.108	0.113	0.317

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. All specifications include strata fixed effects and control for teacher gender, age, education, years of experience, possession of a secondary teaching license, prior non-teaching work experience, number of classes, and total enrollment.

Perceived workload is an index constructed from eight items (4-point Likert scale) aligned with the 2019 TIMSS teacher questionnaire. *Instructional practices* are measured across three domains: lecture-based instruction, repetitive practice, and inquiry-based instruction, following the TIMSS 2019 framework (4-point scale). *Mental health* is measured using the WHO-5 Well-Being Index. *Growth mindset* is measured using two items adapted from Yeager et al. (2022). Higher values indicate stronger endorsement of a growth mindset. ***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table 8: Effects on Student–Teacher Relationship, Behavioral Problems, and Classroom Discipline

	Student–Teacher Relationship	Behavioral Problems	Punctuality	Attendance Discipline	Proper Shoe Arrangement	Organization & Orderliness	Proper Greetings	Responsiveness	Clarity of Speech	Polite Language Use	Preparedness of Learning Materials	Attentive Listening	Quiet Behavior	Participation in Cleaning
D^{T1} (Math schools)	0.000	0.017	0.042	0.011	0.027	0.043	0.005	0.015	-0.022	-0.012	0.000	-0.003	-0.006	0.026
	(0.030)	(0.024)	(0.028)	(0.039)	(0.026)	(0.025)	(0.033)	(0.028)	(0.034)	(0.028)	(0.038)	(0.029)	(0.037)	(0.042)
D^{T2} (Science schools)	-0.001	0.011	0.005	0.024	0.002	0.011	0.020	-0.009	-0.002	-0.006	-0.010	-0.023	-0.018	0.034
	(0.022)	(0.025)	(0.028)	(0.037)	(0.024)	(0.029)	(0.022)	(0.022)	(0.033)	(0.023)	(0.035)	(0.026)	(0.032)	(0.029)
Y_{it-1}	0.361***	0.411***	0.325***	0.285***	0.253***	0.320***	0.299***	0.242***	0.243***	0.327***	0.253***	0.251***	0.251***	0.268***
	(0.026)	(0.040)	(0.025)	(0.021)	(0.023)	(0.026)	(0.021)	(0.024)	(0.019)	(0.022)	(0.018)	(0.022)	(0.025)	(0.021)
Strata fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covariates	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	2,143	1,895	2,183	2,142	2,109	2,103	2,093	2,208	2,139	2,124	2,113	2,097	2,211	2,129
R^2	0.130	0.226	0.115	0.091	0.094	0.136	0.114	0.076	0.069	0.119	0.077	0.075	0.078	0.081

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. All specifications include strata fixed effects and the same covariates as in Table 3.

Student-teacher relationship is constructed from three student self-reported responses on a four-point Likert scale regarding whether teachers recognized students' strengths, listened to their concerns, and provided explanations until understanding was achieved.

Behavioral problem is constructed from as the sum of the 12 items listed from “punctuality”onward. Each of these items is based on student self-reported responses on a four-point Likert scale.

Table 9: Impact of Nationally Implemented Departmentalization on Student Achievements (Grade 5)

	Science Achievement						Math	Reading
	Overall	Basic	Applied	Knowledge and Skills	Inquiry, Reasoning and Expression	Proactive Attitude toward Learning		
Math Departmentalization	0.021	0.051	-0.037	0.056	0.004	0.071	-0.019	0.036
	(0.084)	(0.080)	(0.090)	(0.054)	(0.117)	(0.051)	(0.066)	(0.054)
Science Departmentalization	0.065	0.075	0.038	0.096	0.043	0.034	-0.135*	-0.006
	(0.046)	(0.046)	(0.046)	(0.051)	(0.047)	(0.051)	(0.055)	(0.037)
Class Size	-0.009*	-0.009*	-0.006	-0.008	-0.010*	-0.006	-0.006	-0.003
	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.003)
Y_{it-1}	0.318***	0.297***	0.252***	0.271***	0.274***	0.147***	0.777***	0.780***
	(0.048)	(0.045)	(0.036)	(0.042)	(0.042)	(0.029)	(0.017)	(0.016)
Covariates	YES	YES	YES	YES	YES	YES	YES	YES
N	3,451	3,451	3,451	3,451	3,451	3,451	3,591	3,601
R^2	0.130	0.121	0.080	0.105	0.100	0.045	0.606	0.621

Note: Entries report OLS estimates using Grade 5 students. Standard errors clustered at the school level are in parentheses. All specifications include strata fixed effects and control for student gender, number of classes, and total enrollment.
Grade 5 students were not part of the randomized intervention in this study, but they took the same standardized achievement tests as Grade 4 students (All variable definitions corresponds to those provided in Table 3). Unlike Grade 4, Grade 5 is subject to the nationally promoted departmentalization policy implemented since 2022. This institutional contrast allows us to benchmark the effects of national departmentalization against the experimental intervention evaluated in this paper.
***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table 10: Impact of Nationally Implemented Departmentalization on Learning Strategies and Non-cognitive Outcomes (Grade 5)

	Learning Strategies					Non-cognitive Skills	
	Flexible	Planning	Task-oriented	Cognitive	Effort Regulation	Self-Efficacy	Self-Control
Math Departmentalization	0.014	0.156*	-0.045	0.006	-0.032	0.018	-0.058
	(0.061)	(0.064)	(0.050)	(0.074)	(0.050)	(0.064)	(0.058)
Science Departmentalization	0.064	0.041	0.056	0.024	-0.036	-0.011	-0.065
	(0.048)	(0.043)	(0.049)	(0.043)	(0.036)	(0.044)	(0.045)
Class Size	-0.003	-0.001	0.000	0.001	0.000	-0.002	-0.001
	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Y_{it-1}	0.494***	0.467***	0.499***	0.471***	0.471***	0.594***	0.502***
	(0.015)	(0.018)	(0.015)	(0.017)	(0.020)	(0.014)	(0.018)
Covariates	YES	YES	YES	YES	YES	YES	YES
N	3,555	3,552	3,551	3,376	3,560	3,560	3,559
R^2	0.261	0.232	0.280	0.246	0.235	0.376	0.265

Note: Entries report OLS estimates using Grade 5 students. Standard errors clustered at the school level are in parentheses. All specifications include strata fixed effects and control for student gender, number of classes, and total enrollment.

Grade 5 students were not part of the randomized intervention in this study, but they took the same standardized surveys as Grade 4 students (All variable definitions corresponds to those provided in Table 3). Unlike Grade 4, Grade 5 is subject to the nationally promoted departmentalization policy implemented since 2022.

***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table 11: Crowding-Out by Homeroom Teacher Subject Expertise

	Achievement		
	Math (std.)	Science (std.)	Reading (std.)
D^{T1} (Math schools)	0.010	-0.090	0.031
	(0.072)	(0.067)	(0.063)
D^{T2} (Science schools)	0.149	0.233*	0.082
	(0.081)	(0.090)	(0.082)
Y_{it-1}	0.729***	0.274***	0.749***
	(0.019)	(0.056)	(0.022)
Math-specialized homeroom teacher	0.253*	0.103	0.162
	(0.101)	(0.098)	(0.114)
Science-specialized homeroom teacher	-0.007	0.427***	0.049
	(0.100)	(0.091)	(0.086)
$D^{T1}\times$ Math-specialized homeroom teacher	-0.261*	0.153	-0.122
	(0.116)	(0.110)	(0.138)
$D^{T1}\times$ Science-specialized homeroom teacher	0.086	-0.097	0.000
	(0.137)	(0.135)	(0.098)
$D^{T2}\times$ Math-specialized homeroom teacher	-0.199	-0.283	-0.154
	(0.144)	(0.152)	(0.159)
Strata fixed effects	YES	YES	YES
Covariates	YES	YES	YES
N	1,786	1,808	1,777
R^2	0.527	0.105	0.565

Note: Entries report LATE estimates with standard errors clustered at the school level in parentheses. All specifications include strata fixed effects and the same set of covariates as in Table 3). ***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.
Interaction terms capture potential crowding-out effects when the homeroom teacher is specialized in mathematics or science. Homeroom teacher subject specialization is based on self-reported expertise in the teacher questionnaire.
Student-teacher relationship is constructed from three student-reported items regarding whether teachers recognized students' strengths, listened to their concerns, and provided explanations until understanding was achieved.

Table 12: Robustness: Controlling for Age Differences between Homeroom and Specialist Teachers

	Achievement		
	Math (std.)	Science (std.)	Reading (std.)
D^{T1} (Math schools)	-0.032	0.006	-0.023
	(0.104)	(0.111)	(0.077)
D^{T2} (Science schools)	0.090	0.266**	-0.032
	(0.109)	(0.096)	(0.093)
Y_{it-1}	0.730***	0.274***	0.730***
	(0.019)	(0.055)	(0.019)
Difference in age (homeroom − specialist in science)	0.002	0.000	0.004
	(0.003)	(0.003)	(0.002)
Difference in age (homeroom − specialist in math)	0.003	-0.005	0.006
	(0.003)	(0.003)	(0.003)
Age (homeroom teacher)	0.000	-0.002	0.000
	(0.003)	(0.004)	(0.003)
Strata fixed effects	YES	YES	YES
Covariates	YES	YES	YES
N	1,786	1,808	1,790
R^2	0.525	0.099	0.568

Note: Entries report regression estimates with standard errors clustered at the school level in parentheses. All specifications include strata fixed effects and the same set of covariates as in Table 3. ***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively. *Difference in age* is defined as the homeroom teacher's age minus the specialist teacher's age in the corresponding subject.

Table 13: Placebo Test Using Grade 5 Students

	Academic Achievement			Learning Strategies					Non-cognitive Skills	
	Science (std.)	Math (std.)	Japanese (std.)	Flexible strategy	Planning strategy	Task-oriented	Cognitive strategy	Effort regulation	Self-efficacy	Self-control
D^{T1} (Math schools)	-0.019	-0.117	-0.041	-0.072	-0.029	-0.066	-0.057	-0.122*	-0.089	-0.151**
	(0.071)	(0.094)	(0.059)	(0.062)	(0.062)	(0.061)	(0.058)	(0.057)	(0.054)	(0.055)
D^{T2} (Science schools)	-0.009	-0.040	-0.032	-0.011	0.041	-0.095	-0.002	-0.091	0.032	-0.072
	(0.051)	(0.055)	(0.043)	(0.064)	(0.066)	(0.062)	(0.059)	(0.052)	(0.060)	(0.056)
Y_{it-1}	0.362***	0.771***	0.773***	0.511***	0.480***	0.505***	0.488***	0.474***	0.611***	0.517***
	(0.061)	(0.022)	(0.021)	(0.018)	(0.023)	(0.020)	(0.022)	(0.025)	(0.016)	(0.023)
Strata fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covariates	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	2,360	2,375	2,389	2,348	2,343	2,344	2,228	2,348	2,350	2,349
R^2	0.153	0.573	0.597	0.261	0.229	0.271	0.244	0.228	0.379	0.266

Note: Entries report estimates drawn from the same specification of 3 using Grade 5 students, who were not exposed to the intervention. Standard errors clustered at the school level are in parentheses. All specifications include strata fixed effects and control for student gender, number of classes, and total student enrollment. ***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.

A Appendix Tables

Table A1: Baseline Scores and Attrition

		Non-dropped(a)	Dropped from Sample(b)	(a)-(b)
Math Schools（T1）	Math	-0.022	0.140	-0.162
	N	977	11	
	Science	-0.045	0.340	-0.385
	N	970	11	
	Reading	-0.049	-0.071	0.022
	N	975	11	
Science Schools（T2）	Math	0.064	-0.149	0.214
	N	777	84	
	Science			
	N			
	Reading	0.088	-0.149	0.238*
	N	777	84	

Note: Achievement tests and student questionnaires were in some cases administered on the same day, but in most cases the science test was administered on a separate day because it was conducted by a different testing provider; in all science treatment schools, the science test was administered on a day different from the mathematics and reading tests. As described in Section 4, data from two schools could not be used because student identifiers were inconsistently assigned across the two waves due to administrative errors, preventing accurate longitudinal matching; these were schools that administered all assessments on the same day. Because all science treatment schools administered the science test separately, this mismatch was avoided for science outcomes, which accounts for the differences in sample size across columns.

Table A2: Attrition and Treatment Assignment

	Science	Math	Reading
D^{T1} (Math School)	-0.107	-0.125	-0.124
	(0.080)	(0.082)	(0.082)
D^{T2} (Science School)	-0.137	-0.034	-0.034
	(0.085)	(0.114)	(0.114)
Y_{it-1}	-0.010	-0.026***	-0.020**
	(0.005)	(0.007)	(0.007)
Strata FE	YES	YES	YES
Covariates	YES	YES	YES
N	2,631	2,732	2,732
R^2	0.109	0.126	0.123

Note: The dependent variable equals one if a student attrited at the baseline or endline assessment and zero otherwise. A majority of attritted sample belongs to the treatment schools: whereas attrition in RCTs typically concentrates in the control schools, in this study attrition occurs in the treatment group, as reflected in the uniformly negative coefficients on the assignment indicators. This attrition originates from administrative errors in identifier assignment rather than from responses to the treatment itself, and is therefore unlikely to be systematic. Entries report regression estimates with standard errors clustered at the school level in parentheses. In addition to strata fixed effects, the regressions control for the number of classes and total student enrollment as covariates. ***, **, and * denote statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table A3: Timing of Subject-Specialist Teachers

	Science	Math	Reading
D^{T1} （Math School）	-0.083	0.028	0.041
	(0.149)	(0.130)	(0.095)
D^{T2} （Science School）	0.247*	0.177*	0.071
	(0.098)	(0.077)	(0.068)
Y_{it-1}	0.274***	0.724***	0.727***
	(0.053)	(0.026)	(0.033)
F-statistic	8.911	8.617	8.435
P-value	0.221	0.766	0.917
Strata FE	YES	YES	YES
Covariates	YES	YES	YES
N	2,395	2,394	2,399
R^2	0.093	0.514	0.532

Note: Entries report LATE estimates using an instrument equal to one if the specialist was assigned at the beginning of academic year and zero if assignment occurred later. ***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively. Standard errors are clustered at the school level. In addition to strata fixed effects, the regressions control for number of classes and total student enrollment as covariates.

Table A4: Effect of Subject-Specialist Teacher Deployment: Heterogeneity by Gender

	Academic Achievement			Learning Strategies					Non-cognitive Skills	
	Science	Math	Reading	Flexible	Planning	Task-oriented	Cognitive	Effort Regulation	Self-Efficacy	Self-Control
D^{T1} (Math schools)	-0.066	0.020	0.078	0.048	-0.008	0.064	0.005	-0.036	-0.064	0.011
	(0.091)	(0.067)	(0.075)	(0.068)	(0.098)	(0.078)	(0.106)	(0.104)	(0.080)	(0.097)
D^{T2} (Science schools)	0.180**	0.117*	0.070	0.074	0.037	0.148*	0.135	-0.080	-0.046	-0.176
	(0.064)	(0.058)	(0.067)	(0.066)	(0.095)	(0.072)	(0.096)	(0.087)	(0.072)	(0.098)
Female	-0.026	-0.020	0.048	-0.159**	-0.161**	0.108	-0.136	0.022	-0.190***	0.242***
	(0.045)	(0.046)	(0.044)	(0.052)	(0.062)	(0.061)	(0.072)	(0.057)	(0.051)	(0.065)
$D^{T1} \times$ Female	0.040	-0.005	-0.112	0.067	0.126	0.002	0.051	-0.078	0.087	0.013
	(0.132)	(0.085)	(0.085)	(0.094)	(0.087)	(0.092)	(0.113)	(0.103)	(0.082)	(0.109)
$D^{T2} \times$ Female	-0.003	0.018	-0.038	0.071	0.071	-0.060	-0.000	0.066	0.157*	0.066
	(0.084)	(0.059)	(0.064)	(0.077)	(0.094)	(0.101)	(0.102)	(0.101)	(0.080)	(0.117)
F-statistic	15.01	17.74	16.41	21.12	21.45	21.16	23.53	20.29	20.33	19.78
P-value	0.645	0.637	0.649	0.628	0.534	0.415	0.459	0.335	0.229	0.983
N	2,395	2,394	2,399	2,334	2,334	2,333	2,203	2,335	2,350	2,353
R^2	0.098	0.514	0.532	0.183	0.185	0.199	0.166	0.204	0.240	0.099

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. In addition to strata fixed effects, the regressions control for baseline test scores, number of classes, and total student enrollment.
***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table A5: Effect of Subject-Specialist Teacher Deployment: Heterogeneity by Baseline Test Score

	Academic Achievement			Learning Strategies					Non-cognitive Skills	
	Science	Math	Reading	Flexible	Planning	Task-oriented	Cognitive	Effort Regulation	Self-Efficacy	Self-Control
D^{T1} (Math School)	-0.034	0.109	0.155	0.190	0.114	0.178	0.054	-0.063	-0.007	0.112
	(0.177)	(0.127)	(0.134)	(0.116)	(0.128)	(0.109)	(0.130)	(0.150)	(0.109)	(0.130)
D^{T2} (Science School)	-0.036	0.056	0.069	0.159	0.071	0.110	0.083	-0.043	-0.015	-0.194
	(0.139)	(0.061)	(0.061)	(0.092)	(0.110)	(0.087)	(0.111)	(0.123)	(0.093)	(0.129)
Bottom tercile	-0.125	-0.177*	-0.202**	0.011	0.005	-0.173	0.059	-0.010	-0.055	-0.468***
	(0.141)	(0.074)	(0.068)	(0.121)	(0.093)	(0.096)	(0.099)	(0.098)	(0.075)	(0.088)
Middle tercile	-0.209	-0.207*	-0.059	0.153	-0.022	-0.198*	-0.002	-0.021	-0.104	-0.331***
	(0.147)	(0.088)	(0.073)	(0.100)	(0.073)	(0.101)	(0.109)	(0.085)	(0.078)	(0.082)
Top tercile	-0.316*	0.168	0.062	0.052	-0.105	-0.163	0.207	0.080	-0.021	0.163
	(0.154)	(0.110)	(0.098)	(0.109)	(0.086)	(0.116)	(0.109)	(0.083)	(0.098)	(0.096)
$D^{T1}\times$ middle tercile	-0.021	-0.097	-0.187	-0.224	-0.085	-0.176	0.090	0.017	-0.004	-0.152
	(0.198)	(0.129)	(0.143)	(0.161)	(0.120)	(0.117)	(0.127)	(0.143)	(0.094)	(0.127)
$D^{T1}\times$ top tercile	-0.016	-0.229	-0.251	-0.123	-0.109	-0.205	-0.190	-0.064	-0.052	-0.177
	(0.289)	(0.192)	(0.247)	(0.210)	(0.163)	(0.161)	(0.172)	(0.153)	(0.153)	(0.156)
$D^{T2}\times$ middle tercile	0.157	0.127	-0.067	-0.134	-0.030	-0.038	0.180	-0.037	0.072	0.058
	(0.200)	(0.082)	(0.080)	(0.110)	(0.098)	(0.115)	(0.126)	(0.119)	(0.082)	(0.152)
$D^{T2}\times$ top tercile	0.565*	0.122	0.051	-0.007	0.053	0.085	-0.032	0.035	0.083	0.153
	(0.270)	(0.080)	(0.077)	(0.138)	(0.106)	(0.112)	(0.139)	(0.146)	(0.080)	(0.157)
F-statistic	10.46	11.85	9.602	11.85	14.06	16.14	16.02	14.24	14.91	15.56
P-value	0.639	0.659	0.734	0.608	0.545	0.411	0.466	0.329	0.220	0.996
N	2,395	2,394	2,399	2,334	2,334	2,333	2,203	2,335	2,350	2,353
R^2	0.111	0.527	0.532	0.183	0.184	0.202	0.165	0.205	0.239	0.134

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. In addition to strata fixed effects, the regressions control for baseline test scores, number of classes, and total student enrollment.
***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table A6: Effect of Subject-Specialist Teacher Deployment: Heterogeneity by Household SES

	Academic Achievement			Learning Strategies					Non-cognitive Skills	
	Science	Math	Reading	Flexible	Planning	Task-oriented	Cognitive	Effort Regulation	Self-Efficacy	Self-Control
D^{T1} (Math schools)	-0.044	-0.090	0.123	0.399*	0.466*	0.329	0.221	0.255	0.069	0.442**
	(0.224)	(0.197)	(0.141)	(0.185)	(0.196)	(0.177)	(0.173)	(0.158)	(0.171)	(0.153)
D^{T2} (Science schools)	0.083	-0.100	-0.046	0.181	0.239	0.179	0.188	-0.092	-0.072	0.002
	(0.207)	(0.138)	(0.123)	(0.193)	(0.211)	(0.203)	(0.191)	(0.186)	(0.158)	(0.181)
Low SES	0.000	0.001	0.036	0.011	0.051	0.011	0.031	0.052	0.007	0.064***
	(0.029)	(0.029)	(0.042)	(0.022)	(0.027)	(0.024)	(0.021)	(0.029)	(0.021)	(0.017)
$D^{T1}\times$ Low SES	0.002	0.036	-0.030	-0.105	-0.126*	-0.086	-0.057	-0.092	-0.027	-0.125**
	(0.068)	(0.059)	(0.042)	(0.054)	(0.056)	(0.054)	(0.051)	(0.047)	(0.049)	(0.044)
$D^{T2}\times$ Low SES	0.030	0.074	0.012	-0.024	-0.048	-0.021	-0.016	0.026	0.036	-0.034
	(0.061)	(0.042)	(0.017)	(0.054)	(0.063)	(0.058)	(0.054)	(0.061)	(0.050)	(0.054)
F-statistic	16.49	20.30	17.55	30.73	31.46	30.51	35.92	30.24	30.22	27.67
P-value	0.392	0.344	0.494	0.357	0.237	0.228	0.208	0.280	0.280	0.374
N	2,395	2,394	2,402	2,334	2,334	2,333	2,203	2,335	2,350	2,353
R^2	0.097	0.515	0.532	0.185	0.187	0.202	0.168	0.207	0.239	0.103

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. In addition to strata fixed effects, the regressions control for number of classes and total student enrollment. Low SES is defined based on the share of students receiving school attendance support.
***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively.

Table A7: Effect of Subject-Specialist Teacher Deployment: Heterogeneity by Homeroom Teacher Experience

	Academic Achievement			Learning Strategies					Non-cognitive Skills	
	Science	Math	Reading	Flexible	Planning	Task-oriented	Cognitive	Effort Regulation	Self-Efficacy	Self-Control
D^{T1} (Math schools)	-0.005	-0.007	-0.002	0.146	0.087	0.194	0.089	-0.067	-0.013	0.099
	(0.155)	(0.107)	(0.097)	(0.123)	(0.163)	(0.146)	(0.159)	(0.156)	(0.141)	(0.147)
D^{T2} (Science schools)	-0.023	0.172*	0.119	0.250*	0.188	0.168	0.314*	-0.036	-0.067	-0.255
	(0.102)	(0.085)	(0.086)	(0.107)	(0.151)	(0.125)	(0.133)	(0.144)	(0.141)	(0.180)
Year	-0.009	-0.006	-0.008	-0.004	-0.003	0.003	0.012	-0.006	-0.002	-0.010
	(0.007)	(0.009)	(0.007)	(0.007)	(0.013)	(0.010)	(0.011)	(0.012)	(0.009)	(0.011)
$D^{T1}\times$ Year	-0.000	0.004	0.004	-0.006	-0.004	-0.015	-0.011	-0.001	-0.002	-0.006
	(0.014)	(0.011)	(0.010)	(0.012)	(0.014)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
$D^{T2}\times$ Year	0.025**	0.002	0.001	-0.007	-0.008	-0.007	-0.021	0.002	0.008	0.016
	(0.009)	(0.010)	(0.009)	(0.010)	(0.013)	(0.011)	(0.013)	(0.013)	(0.012)	(0.014)
F-statistic	11.46	13.10	12.06	16.80	16.93	16.59	17.75	16.60	16.49	15.96
P-value	0.607	0.992	0.451	0.558	0.493	0.614	0.472	0.358	0.0752	0.658
N	2,180	2,168	2,173	2,114	2,113	2,112	1,994	2,115	2,129	2,133
R^2	0.104	0.504	0.527	0.179	0.176	0.205	0.160	0.202	0.240	0.107

Note: Entries report LATE estimates. Standard errors clustered at the school level are in parentheses. In addition to strata fixed effects, the regressions control for student gender, number of classes, and total student enrollment. The variable *Year* indicates the homeroom teacher's years of experience.
***, **, and * indicate statistical significance at the 0.1%, 1%, and 5% levels, respectively.

B Appendix

B.1 Cost–Effectiveness Calculation

This section presents the basis for calculating the cost-effectiveness of deploying specialist science teachers and reducing class size.

Cost-Effectiveness of Deploying Specialist Science Teachers The total annual cost of deploying specialist science teachers was calculated as: 「Total Cost = (Number of Schools with Science Specialists) × (Number of Classes per School) × (Standard Instruction Hours) × (Hourly Wage Rate)」. In this calculation, the number of schools refers to the actual number of schools where specialist science teachers were assigned (15 schools), and the number of classes refers to the average number of fourth grade classes in those schools (1.85 classes). The standard number of instruction hours (105 hours per year) and the hourly wage rate for teachers (3,100 JPY per hour) were taken from guidelines issued by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2021) and the Chiba Prefecture Board of Education (2025), respectively. Based on these figures, the total cost of deploying specialist science teachers across the 15 schools was calculated to be 9,032,625 JPY.

Given that the average number of fourth grade students in these schools was 42.6, the total number of students who received instruction from specialist science teachers was 639. Accordingly, the cost per student was calculated to be 14,135.56 JPY. Since the deployment of specialist science teachers increased science test scores by 0.18 SD (see Table 4, LATE estimate), the cost of achieving a one standard deviation increase in test scores is estimated to be 79,413.28 JPY.

Cost-Effectiveness of Class Size Reduction The cost of class size reduction is estimated based on the expenditure required to reduce the number of students per teacher by one. Since the marginal cost of reducing class size increases as class sizes become smaller, this approach yields a conservative estimate which is likely to be lower than the actual cost of implementing a small class size policy.

Given that the average number of fifth grade classes per school is 1.65 and the average number of students per class is 26.6, reducing the class size by one student would require the equivalent of hiring an additional 0.064 teachers per school. With the average teacher salary in Chiba Prefecture being 5,923,000 JPY per year (Chiba Prefecture, 2022), the cost of reducing class size by one student is estimated to be 381,755.86 JPY per school. As the average number of fifth grade students per school is 43.9, the per-student cost of implementation is calculated to be 8,698.01 JPY. Since reducing class size by one student increases test scores by 0.009 SD (see Table 9), the cost of achieving a one standard deviation increase in test scores is estimated to be 966,446.06 JPY.

Taken together, these calculations indicate that the deployment of specialist science teachers is substantially more cost-effective than reducing class size. Using the estimates derived from the Chiba Prefecture data, the specialist teacher intervention is approximately 12.2 times more cost-effective. Considering that the cost estimate for class size reduction is conservative, the true difference in cost-effectiveness is likely to be even larger.

Two caveats apply to this comparison. First, the class size effect is an OLS estimate based on non-experimental variation among Grade 5 students and may therefore be biased. We note, however, that our class size estimate (0.009 SD per student) is consistent with quasi-experimental evidence from Japan (Ito et al., 2020), which mitigates this concern. Second, the cost estimate for class size reduction is conservative: because the marginal cost of reducing class size rises as classes become smaller, the true cost of a class size reduction policy is likely higher than our calculation implies. Both caveats work in the same direction, and if anything strengthen the conclusion that targeted specialization is the more cost-effective policy.