RIETI Discussion Paper Series 23-E-066

The Impact of the COVID-19 Pandemic on the Academic Achievement of Elementary and Junior High School Students: Analysis using administrative data from Amagasaki City

ASAKAWA, Shinsuke
Saga University / Osaka University

OHTAKE, Fumio
RIETI

SANO, Shinpei
Kobe University

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

RIETI Discussion Paper Series 23-E-066

September 2023

The Impact of the COVID-19 Pandemic on the Academic Achievement of Elementary and Junior High School Students: Analysis using administrative data from Amagasaki City[*†]

Shinsuke ASAKAWA
Saga University and Osaka University

Fumio OHTAKE
Osaka University and Research Institute of Economy, Trade and Industry

Shinpei Sano
Kobe University

Abstract

This study examines the effects of the COVID-19 pandemic on academic achievements in mathematics and Japanese language among public elementary and junior high school students from grades 1 to 8. Using data from the Amagasaki City Survey of Academic Achievement and Life Conditions from 2018 to 2021, this study compares the growth in the academic achievement of COVID-19 and non-COVID-19 cohorts 7 and 19 months after the school closure using the difference-in-differences method. The findings indicate that the negative impact of the pandemic on academic achievement was more pronounced in math compared to Japanese language, both at 7 months and 19 months after the closure. Math scores showed a considerable decline of 0.129 standard deviations (SD) and 0.251 SD at 7 and 19 months after the closure, respectively, while Japanese language scores only worsened slightly by 0.006 SD and 0.062 SD during the same periods. Further, the negative effects on Japanese language scores were more significant in younger grades, whereas math scores were consistently affected across all grades.

Keywords: COVID-19, school closure, academic achievement, mathematics, Japanese language
JEL classification: I21, I24, I28

[*]This study is a part of the Project "Implementing Evidence-Based Policy Making in Japan" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). We would like to thank Mayuko Abe, Charles Yuji Horioka, Keisuke Kawata, Yuta Kuroda, Masayuki Morikawa, Kazuo Yamaguchi, Izumi Yokoyama and participants of Kansai Labor Workshop, Japanese Economic Association Spring Meeting 2023, and Discussion Paper seminar participants at RIETI for helpful comments on this study. This study is part of the research projects of the Amagasaki Manabitosodachi Research Institute, and administrative data were used with the approval of the Ethics Committee of Amagasaki City. This work was supported by Saga University and a Grant-in-Aid for Scientific Research (S) (Grant number [20H05632]) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. This study uses data from Amagasaki City, which cannot be publicly shared. We used R version 4.2.2 to estimate the results. The source code used for the analysis is available upon request. The authors are solely responsible for any errors made in this study.

[†] Asakawa: Assistant Professor, Faculty of Economics, Saga University, 1 Honjo-machi, Saga, 840-8502, Japan. (Corresponding author. Email: asakawas@cc.saga-u.ac.jp) and Visiting Researcher, Graduate School of Economics, Osaka University, Japan. Ohtake: Specially Appointed Professor, Center for Infectious Disease Education and Research (CiDER), Osaka University. Sano: Associate Professor, Graduate School of Economics, Kobe University, Japan

1. Introduction

The spread of the coronavirus disease 2019 (COVID-19) caused temporary school closures in many countries and regions, resulting in a decline in student academic achievement (Australia: Gore et al. 2021; Belgium: Gambi et al. 2021; Maldonado and De Witte 2022; Brazil: Lichand et al. 2022; Colombia: Vegas 2022; Denmark: Birkelund and Karlson 2022; Germany: Schult et al. 2022a, 2022b; Italy: Bazoli et al. 2022; Contini et al. 2022; Japan: Asakawa and Ohtake 2022; Mexico: Hevia et al. 2022; Netherlands: Engzell, Frey, and Verhagen 2021; Haelermans et al. 2022; Spain: Arenas and Gortazar 2022; South Africa: Ardington, Wills, and Kotze 2021; Switzerland: Tomasik, Helbling, and Moser 2021; UK: Blainey and Hannay 2021; US: Jack et al. 2021; Kuhfeld et al. 2020; Kuhfeld, Lewis, and Peltier 2022; Kuhfeld et al. 2022).

Some studies used the meta-analysis and systematic review methods to show that the negative impact of the COVID-19 pandemic on academic performance varies widely by country/region (Betthäuser, Bach-Mortensen, and Engzell 2023; Donnelly and Patrinos 2022). Qualitatively, the negative impact of COVID-19 on students' test scores was confirmed in several countries and regions. However, quantitatively, the magnitude of the impact varies across countries and regions. For example, Betthäuser et al. (2023) conducted a meta-analysis using estimates from 42 previous studies in 15 countries, and found that COVID-19 reduced students' test scores by an average of 0.14 SD. They also confirmed that the students with a less advantaged socioeconomic status were more likely to deteriorate their test scores due to COVID-19, and that test scores declined more in middle-income countries than in high-income countries due to the COVID-19.

However, evidence on the impact of COVID-19 on students' academic performance has not been fully accumulated in Japan. Therefore, this study estimates the effects of the COVID-19 pandemic, accompanied by the declaration of a state of emergency and temporary school closures to prevent the outbreak,[1] on Japanese language and math test

[1] On January 16, 2020, the first case of COVID-19 infection was confirmed in Japan. Subsequently, the Japanese government requested that all elementary, junior high, high, and special-needs schools be temporarily closed from March 2, 2020 until spring break to prevent the nationwide spread of infection among students. After the spring break ends on April 6, 2020, schools remained closed in municipalities with serious COVID-19 outbreaks. As of May 11, 2020 (2 months and 10 days after the closure), the school closure rate was approximately 88%. In addition, 33% of schools remained closed for more than 2 months and the closures were lifted for all schools on June 1, 2020. For the the state of emergency, on April 7, 2020,

scores for students in grades 1–8 in all public elementary and junior high schools in Amagasaki City, Japan. Amagasaki City is a large municipality with a population of approximately 460,000 in 2019, just before COVID-19, yet the average household income in 2019 was approximately 4.62 million JPY, lower than the average of 5.58 million JPY in Hyogo Prefecture and the national average of 5.61 million JPY.

To compare test scores for the same school grade across different cohorts, we use the Amagasaki City achievement test, which is vertically and horizontally equalized according to item response theory (IRT, Embretson and Reise 2013). We standardize the test scores to compare effect sizes across grades and previous studies. In Amagasaki City, the COVID-19 school closure period (March 2–May 31, 2020) was 12 weeks, so the period between the end of the closure and the 2020 and 2021 achievement tests used in the analysis was 7 and 19 months, respectively.

We use the difference-in-differences (DID) method to estimate the impact of the COVID-19 pandemic on students' Japanese and math 7 months and 19 months after the school closure. To estimate the impact 7 months after the school closure, we compare the growth from 2019 to 2020 for the COVID-19 experience cohort (treatment group 1) with the growth from 2018 to 2019 for the COVID-19 non-experience cohort (control group 1). Furthermore, to analyze the impact 19 months after the school closure, we compare the growth from 2019 to 2021 for the cohort who took the test twice after the COVID-19 school closure (treatment group 2) and the growth from 2018 to 2020 (control group 2). However, because control group 2 includes the effect of school closure in 2020, we remove this effect by subtracting the effects 7 months after the school closure estimated between treatment group 1 and control group 1. To facilitate interpretation, we analyze the data at 7 months after school closure for three grade groups: lower elementary for grade groups 1–3, upper elementary for grade groups 4–5, and junior high school for grade groups 6–7 in the first of the two analysis periods. In the analysis, 19 months after

the first state of emergency was issued for Saitama, Chiba, Tokyo, Kanagawa, Osaka, Hyogo, and Fukuoka prefectures. In addition, on April 16, 2020, the government issued a state of emergency for the remaining prefectures. Subsequently, the state of emergency declaration was continued in only eight prefectures (Hokkaido, Saitama, Chiba, Tokyo, Kanagawa, Kyoto, Osaka, and Hyogo) on May 14, 2020, lifted in Kyoto, Osaka, and Hyogo on May 21, 2020, and lifted in all prefectures on May 25, 2020. In 2021, states of emergency were also declared in some prefectures for periods of up to January 8 to March 21, April 25 to June 20, and July 12 to September 30.

school closure, only grade 6 is considered junior high school due to a lack of data.

We further perform quantile-DID and DID with interaction terms to check for heterogeneity of effects across quantiles of test scores and pre-determinant variables, including gender and socio-economic status of the student's household. In addition, we conduct a triple-difference (DDD) estimation to identify heterogeneity in the effects between students in schools that significantly reduced athletic events after school closure and others.

There are five main findings. First, the negative effects of the COVID-19 pandemic were greater for math than for the Japanese language at 7 months, and the deterioration in math scores was more pronounced than in the Japanese language at 19 months after school closure. Specifically, Japanese language scores worsened slightly by 0.006 SD and 0.062 SD on average at 7 and 19 months after school closure, respectively. On the other hand, math scores worsened considerably by 0.129 SD and 0.251 SD at 7 and 19 months after school closure, respectively. Second, Japanese language scores were negatively affected only in the lower grade groups, but the negative effects on math scores did not differ by grade group. Third, Japanese language scores declined more in the upper quartile only in elementary school students, while math scores declined more in the lower quartile in all grades. Fourth, the negative effects of the COVID-19 pandemic varied little depending on the living condition before the COVID-19 pandemic and the gender of the student. However, female students had a smaller negative impact on the Japanese language scores than male students in junior high school. Finally, reducing athletic events after school closure contributed little to the recovery of academic performance.

There are four significant contributions that this study makes to the literature, as follows. First, it examines the impact of the COVID-19 pandemic on academic achievement by using vertically and horizontally equated tests based on IRT to account for changes in test difficulty. In Japan, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT 2021a, 2021b) showed that school closures did not affect the mean or variance of student achievement by using the annual National Assessment of Educational Progress and Learning for sixth-grade and ninth-grade students in Japan. However, this is not an IRT test, so these results may not adequately identify the impact of the COVID-19 pandemic and the changes in test difficulty. Therefore, our study attempts to remove the effects of the changes in test difficulty using the IRT test of

Amagasaki City. This approach aims to offer a more accurate identification of the effects of the COVID-19 pandemic on students' academic performance.

Second, we examine the effects of the COVID-19 pandemic on academic achievement for the grades 1–8 of public elementary and junior high schools. We then compare the effect sizes across grades in Amagasaki to those in existing studies by standardizing the Japanese and math grades at the time of school closure by school grade. This comparison helps clarify whether the effect of school closure differs by grade level, as in previous studies where the effect differed by country/region.

Third, our study examines the medium- and long-term effects of the COVID-19 pandemic on academic achievement. Some existing studies analyzed test data from 2021, more than a year after the school closures (Asakawa and Ohtake 2022; Blainey and Hannay 2021; Kuhfeld, Lewis, and Peltier 2022). By comparing academic performance 7 and 19 months after the closure, we can provide insights into the duration of the impact of the COVID-19 pandemic on academic achievement.

Finally, we analyze the heterogeneity of the impact of the COVID-19 pandemic based on the differences in achievement quartiles, pre-determinant variables, and reduced hours of athletic events after the closure. No consensus has yet been reached as to which academic levels were most affected by the school closures. For example, some studies showed that the lower the academic achievement level was, the greater was the negative impact of the COVID-19 pandemic on academic achievement (Ardington, Wills, and Kotze 2021; Asakawa and Ohtake 2022; Kuhfeld et al. 2022; Schult et al. 2022a, 2022b). Other studies showed that the top academic groups are negatively affected (Contini et al. 2022; Gambi et al. 2021). Moreover, while many countries/regions reduced athletic events after the school closures, no study examined the measure for recovering lost lecture time. Therefore, we determine whether the effects of the COVID-19 pandemic on academic achievement differ by achievement level and the reduction of athletic events. This heterogeneity analysis provides insights into the grades and subjects for which the achievement gap became wider.

This paper consists of seven sections. Section 2 reviews previous studies. Section 3 describes the school closures of elementary and junior high schools in Amagasaki City. Sections 4 and 5 explain the data and the estimation method, respectively. Section 6 presents the results of the empirical analysis. Section 7 concludes the paper.

2. Existing research

In existing research, the countries, grades, subjects, and timing and methods of tests vary from study to study. Here, we focus only on related studies comparable to ours. Specifically, we review the studies in which standardized test scores and national language or math test results are analyzed.

In Australia, where the first wave of school closures lasted as little as 8 weeks, Gore et al. (2021) used matching estimation to analyze the impact on language and math scores of grade 3 and 4 students in New South Wales 6–8 months after the school closure. They found that neither national language nor math scores were significantly affected.

In Germany, where the school closure period was also 8 weeks, Schult et al. (2022a) estimated the effects of the school closure on the national language and math scores of grade 5 students in the state of Baden-Württemberg, 5 months after the school closure. Their pre-post analysis found a negative effect of -0.07 SD for national language and -0.06 SD for math. In a pre-post analysis, Schult et al. (2022b) estimated the effects on the same cohort's national language and math scores 17 months after the school closure. They found that the national language scores recovered slightly (-0.045 SD), while math hardly did so (-0.063 SD).

In the Netherlands, where schools were closed for 8 weeks in the first wave and the hybrid teaching period was 4 weeks, Haelermans et al. (2022) used DID to estimate the impact on the national language and math scores of students in grades 1–5, 2 to 3 months after closure. They found the effects of the COVID-19 pandemic on national language and math scores to be negative (-0.096 to -0.190 SD for National language and -0.129 to -0.326 SD for math). Additionally, when comparing the effect size by grade level, the negative effect was larger in the higher grades, especially for math.

In Belgium, where the first wave of school closures lasted 9 weeks, Maldonado and De Witte (2021) used a fixed effects model to estimate the impact of the COVID-19 pandemic on the national language and math scores of grade 4 students in the Flemish region 1 month after the school closure. The results showed a negative effect of -0.19 SD for the national language and -0.17 SD for math. Gambi and De Witte (2021) used a fixed effects model to estimate the impact on the national language and math scores of grade 6 students in the same region 13 months after the school closure, and found a negative impact of -0.14 SD for the national language and -0.05 SD for math. In addition, only the national language scores worsened further over time after the school closure.

Considering the first wave of 10-week school closures in the UK, Blainey and Hannay (2021) estimated the impact on the language and math scores of students in grades 1–6, 4–7 months after the school closure by comparing the post-school closure test scores with those of the same cohort in the previous school year. Their analysis results indicated a negative effect of -0.02 to -0.155 SD for the national language and -0.02 to -0.09 SD for math. Comparing the effect size by grade, the negative effects were particularly large for grades 1–3 in the national language and grades 1, 3, and 5 in math. UK schools were closed in a second wave from December 2020 to January 2021 for 10–14 weeks. Therefore, Blainey and Hannay (2021) also estimated the impact on scores in January–April and April–July 2021, and found that both language and math scores worsened compared to 4–7 months after the first wave of the school closure.

In Japan, the longest school closure period for the first wave was 12 weeks. Asakawa and Ohtake (2022) estimated the DID effect on math scores at the time of school closure and at three other time points (3, 7, and 10 months after the closure) for grade 4 and 5 students in Nara City. The results showed that scores decreased by -0.14 SD at the time of the school closure, but recovered by -0.075 SD (3 months after the school closure), 0.26 SD (7 months after the school closure), and 0.295 SD (10 months after the school closure). Furthermore, the higher was a student's grade at the time of the school closure, the faster proved the recovery of academic achievement after the school closure.

In Italy, where the first wave of school closures lasted for 15 weeks, Contini et al. (2022) estimated the impact of DID on the math scores of grade 2 students in the province of Torino 4 months after the school closure. They found a negative impact of only -0.19 SD. Bazoli et al. (2022) used coarsened exact matching for grades 5, 8, and 13 students randomly sampled class-by-class from the SY2020–2021 national test (INVALSI). They found a negative effect of -0.316 to -0.057 SD for the national language and -0.291 to -0.142 SD for math. Furthermore, comparing the effect size by grade level, the effects of the national language and math on the youngest students, those in grade 5, were small.

In the U.S., where school closure periods varied widely across states and hybrid instruction was often used, Kuhfeld, Lewis, and Peltier (2022) used a national test (NWEA Measures of Academic Progress) administered during August–November 2020 to students in grades 3–7. Based on pre-post analysis to estimate the impact on language arts and math scores, they found that, while the impact on language arts varied by grade (-0.024 to 0.045 SD), the impact on math was negative for all grades (-0.181 to -0.11 SD). Furthermore, in the following year, the effect sizes for language arts and math were -0.095 to -0.173 SD and -0.213 to -0.262 SD, respectively, indicating that academic achievement worsened for all grades. In both language arts and math, the younger the student was, the greater was the deterioration in academic achievement. Kuhfeld et al. (2022) conducted a multilevel growth model analysis of students in grades 3–7 who took the same test during August–November 2020, December 2020–March 2021, and March–June 2021. They showed that language arts and math scores worsened progressively over time after the school closure. Moreover, the younger the students were, the greater was the deterioration of their academic performance in both subjects.

Table 1 and Figure A1 summarize the above results by school grade. Considering the mean effect sizes by grade group, Table 1 shows that, in Japanese language, the negative impact of the COVID-19 pandemic was larger for the lower grade groups (grades 1–3: -0.14 SD, grades 4–5: -0.073 SD, grades 6–7: -0.058 SD). In math, Table 1 also shows that the negative effect of school closure was slightly larger for the lower grades, but the difference between grades was smaller than that for national languages (grades 1–3: -0.147 SD, grades 4–5: -0.14 SD, grades 6–7: -0.129 SD).

(Table 1 around here)

There are four possible reasons for these varying effects of the COVID-19 pandemic on academic performance. The first is that the duration of school closures varied among countries/regions. For example, previous studies showed that some countries had no school closures (e.g., Sweden). By contrast, others closed schools for less than 10 weeks (Australia: 8 weeks, Belgium: 9 weeks, Denmark: 4 weeks for grades 1–5 and 8 weeks for grades 6–9, Germany: 8 weeks, Netherlands: 8 weeks, Switzerland: 8 weeks), 10–20 weeks (Japan: 12 weeks, Spain: 12 weeks, South Africa: around 10 weeks, Italy: 15 weeks), more than 20 weeks or shifted to a hybrid system (UK: 20–24 weeks, Brazil: 35 weeks, Mexico: 48 weeks, Columbia: around 68 weeks). The differences in the length of COVID-19 school closure may have directly affected the magnitude of the effect on academic performance.

The second reason is that the period between the day after the school closure and the test being analyzed varied among countries/regions. For example, the test was administered within 3 months after the school closures in the Flemish region (Belgium), the Netherlands, and Nara City (Japan). However, in New South Wales (Australia), the test was administered for the first time 6–8 months after the school closure (Asakawa and Ohtake, 2022; Engzell et al., 2021; Gambi and De Witte, 2021; Gore et al., 2021; Haelermans et al., 2022; Maldonado and De Witte, 2021). Asakawa and Ohtake (2022) and Jack et al. (2022) showed that academic performance recovered after resuming face-to-face classes. Therefore, the estimated effect size may have been smaller the longer was the interval between the school closure and the test administration period.

The third reason is that the starting month for classes differs among countries/regions. Since most countries covered by previous studies start classes from July to September, they were already in the latter half of the school year in March 2020 when the COVID-19 school closures started. However, in other countries, the COVID-19 school closures started just after the new school year (Australia and Colombia) or just before it (Japan). The new school year includes increased difficulty in learning content and a change of classmates, so students need to adjust to a new environment. If the school closure coincides with the start of a new school year, the academic achievement decline due to the pandemic may be greater.

The fourth reason is that the sudden COVID-19 school closure in many countries and regions imposed considerable limitations on the grades, regions, and subjects tested that could be used in the analysis. Further, some studies used more than five grades in their research (Haelermans et al. 2022: grades 1–5, Blainey and Hannay 2021 2022a: grades 1–6, Kuhfeld et al. 2022a: grades 4–8, Kuhfeld et al. 2022b: grades 4–8), while other studies used only one grade (Arenas and Gortazar 2022; Contini et al. 2022; Gambi and De Witte, 2021; Gore et al. 2021; Schult et al. 2022a; Schult et al. 2022b; Vegas 2022). Most studies use the national language and math as outcomes, but effect sizes vary widely across countries and regions.[2] In Japan, MEXT (2021a, 2021b) showed that school

[2] When we averaged the effect sizes in previous studies using standardized scores as outcomes, the effect of school closures was around -0.085 SD for the national language and -0.151 SD for math (Arenas and Gortazar 2022; Asakawa and Ohtake 2022; Bazoli et al. 2022; Blainey and Hannay 2021; Contini et al.

closures did not worsen Japanese and math achievement scores using cross-sectional data from the National Assessment of Academic Ability for grades 6 and 9. Using panel data from April 2019 to March 2021, Asakawa and Ohtake (2022) showed that the COVID-19 pandemic caused a temporary decline in the math scores of grade 4 and 5 public elementary school students in Nara City, but the students recovered to the pre-closure level after 6 months. However, few studies have comprehensively analyzed the effects of the COVID-19 pandemic on academic achievement across grades.

3. Response to COVID-19 in elementary and junior high schools

3.1 Nationwide response to COVID-19

In Japan, measures against COVID-19 infection began earnestly in elementary and junior high schools in late February 2020.

On February 25, 2020, the MEXT stated that the decision of school closure in the case of COVID-19 infection was left at the discretion of each local government. However, on February 27, 2020, Prime Minister Abe announced a nationwide simultaneous temporary school closure policy for elementary, junior high, high, and special-needs schools. On the same day, the MEXT requested the temporary closure of all schools from March 2, 2020 until the start of the spring break. Due to the spread of COVID-19, a state of emergency was subsequently declared for seven prefectures on April 7, and the declaration became nationwide on April 16.

As a result, temporary school closures were mainly extended in areas where the infection situation was serious until May 31, when the emergency declaration was lifted. Specifically, the school closure rate for elementary, junior high, and high schools in Japan was around 99% as of March 16, 2020 (14 days after the school closure), 95% as of April 22 (1 month and 20 days after the school closure), and 88% as of May 11, 2020 (2 months and 10 days after the school closure).

After the COVID-19 school closure, many schools reduced their events to compensate for the class time loss due to the closure. This reduction in school event

---

2022; Gambi and De Witte 2021; Gore et al. 2021; Haelermans et al 2022; Kuhfeld et al 2022a; Kuhfeld et al 2022b; Maldonado and De Witte 2021; Schult et al. 2022a; Schult et al. 2022b; Vegas 2022). However, Maldonado and De Witte (2021) and Vegas (2022) reported larger negative effects for the national language, suggesting a heterogeneity of effects by country and region.

implementation and preparation time was expected to improve students' academic performance by allowing more class time. For example, the “School Questionnaire of the National Assessment of Academic Ability 2021” showed that school events were reconsidered in 94.4% (elementary schools) and 93.6% (junior high schools) nationwide, and in 97.7% (elementary schools) and 96.9% (junior high schools) in Hyogo Prefecture (excluding designated cities), where Amagasaki City is situated. However, reduction or cancellation policies for school events vary widely among schools and grades. Moreover, the National Assessment of Academic Ability does not provide objective data on which events have been reduced and by how many hours.

3.2 Response to COVID-19 in Amagasaki City

Following the nationwide request for temporary school closure, Amagasaki City closed all schools from March 3 to May 31, 2020, with "dispersed school attendance once a week" during the last week of May, "dispersed school attendance" from June 1 to 12, and "regular school attendance" from June 15.

During this period, Amagasaki City took the following measures based on a notice from the MEXT. First, students were asked to refrain from going out and attending school during the school closure period. Second, schools were to keep in close contact with students staying at home due to the temporary school closure and their parents. Third, during the original class period, the students were provided with paper-based learning materials based on their textbooks for self-study. However, the content and number of paper-based learning handouts could differ among schools. In addition, educational materials and videos were introduced and provided through education board websites to enable children to study by themselves using ICT terminals in their homes.

From June 2020 to 2021, after the school closure, many school events were reduced or canceled in elementary and junior high schools in Amagasaki City. However, since most of the school events were not recorded as objective data, the reduced time of school events at each school and grade is largely unknown. Among them, athletic events, the only events for which Amagasaki City has objective data, showed large variations in the method of holding events, and their time was reduced after the school closure. For example, some schools divided events into separate grades, drastically reduced event content, substituted an event with a regular physical education class, or canceled the event altogether. As a result, the amount of reduced time for athletic events varied among schools and grades, as did the increased class time due to the reduction of athletic events.

4. Data

The data used in this paper are individual data from the "Amagasaki City Survey of Academic Achievement and Life Conditions." This survey was conducted independently by the city of Amagasaki for all students from the first grade of elementary school to the second grade of junior high school in Amagasaki City from FY2018 to FY2021. The survey consists of an academic achievement survey and a questionnaire survey administered to all elementary schools in December each year and all junior high schools on a specific day in January. The data period is from FY2018 to FY2021. The first two fiscal years are pre-COVID-19 and the latter two are post-COVID-19.

The academic achievement tests are as follows. The achievement test is based on content common to all schools, with a private education provider creating the paper test. The paper test consists of two subjects (Japanese and mathematics) and takes 40 minutes for grades 1–6. In comparison, it includes a test of 45 minutes each for five subjects (Japanese, mathematics, science, social studies, and English) for grades 7 and 8 in junior high school. The paper tests were collected after the tests were administered and scored by the private educational providers who wrote the questions. The results were provided to Amagasaki City after equating them based on IRT by a private education provider.

In addition to the achievement survey, we collected information from questionnaires administered on the same day as the achievement survey and administrative data, including Basic Resident Registration data. The questionnaires were administered the same day as the paper test and were about the students and their lives. The Basic Resident Registration data contain a code that the local government hashes to identify student's identity and household information. Using a code that identifies the individual, we can use other administrative information, such as data on public assistance and school attendance assistance.

The following restrictions apply to the data. First, they are limited to students who took the same-day achievement test. Second, the data are limited to students who reside in Amagasaki and attend public schools in the city. Third, we cannot use the information on school districts and addresses of students because the test data and the Basic Resident

Registration data were provided by hashing students' names and address data by Amagasaki City. Therefore, it is impossible to match these data with the number of new COVID-19 cases per school district or the macro data on school districts and addresses. As a result of these restrictions, the maximum sample size available was 21,937.

Additionally, we used data on the schedule of athletic events from FY2019 to FY2021. Elementary and junior high schools in Amagasaki City are required to apply to the city at least two weeks before the scheduled date and time of the event. After the application is made, the events must be held as requested. In 2019, before the school closure, all schools held athletic events with the participation of all students, but from 2020 onward, the time and schedule differed among schools and grades due to the COVID-19 measures. As a proxy variable for reducing school events after the school closure, we used the scheduled hours of physical education events for each school and grade level from 2019 to 2021.

We now define the variables used in the analysis and present their descriptive statistics. First, we describe the outcome variables—the IRT tests for Japanese and math. We use standardized test scores as outcome variables to compare the estimated results of the effects of the COVID-19 pandemic on the scores of Japanese and math with those of existing studies. The test difficulty can be regarded as equivalent across years for the same grade level because the test questions are designed based on the IRT. Therefore, we independently standardized the prime test scores for Japanese and math for the pairs of DID estimation, treatment group 1 (T1, 2019–2020) and control group 1 (C1, 2018–2019), and treatment group 2 (T2, 2019–2021) and control group 2 (C2, 2018–2020). Additionally, we standardized the test score independently for each grade level to compare effect sizes across grades.

Figures A2 and A3 show the histograms of standardized test scores for Japanese and math by cohort and period. From top to bottom, the histograms for each subject are shown for control group 1, control group 2, treatment group 1, and treatment group 2. The left panel shows the distribution of the second period of the two periods (After: After = 1) for each cohort, and the right panel shows the distribution of the first period (Before: After = 0). These figures show no significant difference within subjects before the school closure, but the distribution is extended more to the left in math than in Japanese.

Next, we explain three pre-determined variables. First, among the list of households eligible for school attendance assistance, we create a dummy for households receiving school attendance assistance (1 if the household receives school attendance assistance at After = 0, and 0 otherwise).[3] Second, living with one parent dummy is created (1 if the student is either in a single-parent household or one of the parents lives alone outside of Amagasaki City at After = 0, and 0 otherwise). Third, we create a female dummy (1 if the student is female, and 0 otherwise).

Moreover, to examine whether the reduction of athletic events by school and grade contributed to the recovery of academic achievement after school closure, we first confirm the distribution of reduced hours for athletic events. Figure A4 shows the histogram of the difference between the scheduled hours of events in 2019 and 2020–2021 and Figure A5 the changes in the scheduled hours of events in 2019–2021 by school and grade. These figures include two groups with different event reduction times bordering on 200 minutes both in FY2020 and FY2021. Therefore, we create two athletic event reduction dummies that take 1 if the athletic events are reduced by over 200 minutes from FY2019 to FY2020 and from FY2019 to FY2021. Specifically, we first calculate the difference in the scheduled time of athletic events before and after the school closure for each school and grade. Then, we create two dummy variables, $Reduc20_{c,sg}$ and $Reduc21_{c,sg}$,that take 1 for schools and grades that reduced their athletic events by over 200 minutes from FY2019 to FY 2020 and from FY2019 to FY 2021, respectively.

Table 2 shows descriptive statistics by cohort, period, and grade group for the outcome variables used in this study. For reference, descriptive statistics of the prime test scores are included. Table 3 shows descriptive statistics for the pre-determinant variables and athletic event reduction dummies in the "Before" by cohort and grade group. To examine whether each variable differs between cohorts, the results of the balance test, that is, the p-value of the t-test are also shown.

(Tables 2 and 3 around here)

[3] These are households with the head of household defined as a person in need of public assistance, as prescribed by Article 6, Paragraph 2 of the Public Assistance Act (around 90,000 persons in FY2021) and those whose head of household is recognized by the municipal board of education as being in need of public assistance as prescribed per Article 6, Paragraph 2 of the Public Assistance Act (about 1,210,000 persons in FY2021). The number of households eligible for school attendance assistance is calculated as the total of public assistance households and quasi-necessary public assistance households.

Table 2 shows that treatment groups T1 and T2 had higher test scores before the school closure than control groups C1 and C2, respectively. Specifically, compared to "Before," which does not include the effects of COVID-19 pandemic, grade groups 1–3 had higher scores for both T1 and T2 cohorts than C1 and C2 in both subjects, grade groups 4–5 had higher scores for both T1 and T2 cohorts than C1 and C2 in math only, and grade groups 6–7 had higher scores for both T1 and T2 cohorts than C1 and C2 in both subjects at the 10% level. However, in Table 3, the number of households receiving school attendance assistance is smaller in the treatment group than in the control one. No obvious differences in other variables are observed between the treatment and control groups.

5. Estimation method

5.1 Main analysis

5.1.1 Impact 7 months after the school closure

Using the DID method, we compare the differences between the growth in standardized scores in Japanese and math from 2019 to 2020 for the COVID-19 experienced group 1 (T_1) and from 2018 to 2019 for the COVID-19 non-experienced group (C_1). Figure 1 shows the cohorts and timing of the tests used in the estimation.

(Figure 1 around here)

To facilitate interpretation and identify the impact of COVID-19 on academic performance by school grade groups, we divide students in both cohorts from grades 1−7 at $After = 0$ into three grade groups based on their grade at $After = 0$ (lower elementary: grades 1–3, upper elementary: grades 4–5, and junior high school: grade 6–7). The estimation equation for each grade group $g \in$ {grades 1–3, grades 4–5, grades 6–7} can be written as follows:

$$Y_{it} = \alpha_i + \gamma\, COVID19_i + \lambda\, After_t + \delta\, COVID19_i * After_t + \nu_{sg} + \varepsilon_{it} \qquad (1)$$

Now, let Y_{it} be the standardized test scores of Japanese and math for student i in year

$t \in \{2018 \dots 2020\}$. $COVID19_i$ is a dummy variable that takes 1 if student i is in treatment group T_1 (0 for control group C_1). $After_t$ is a dummy variable that takes 1 if the test for T_1 was conducted in 2020, and the test for C_1 was conducted in 2019 (0 if the test was conducted 1 year earlier). α_i and ν_{sg} are individual fixed effects and school grade fixed effects at $After_t = 0$ (T_1 cohort is for FY2019, C_1 cohort is for FY2018), respectively. ε_{it} denotes the error term. We also assume that $E[\varepsilon_{it}|t] = 0$.

Denoting *G* and *c* as the grade group and cohort, respectively, the ATT of the COVID-19 pandemic on standardized test scores in Japanese and math for grade group g, $\delta_{g,ATT}$ can be written as follows:

$$\begin{aligned}\delta_{g,ATT} = \{&E[Y_{it}|G = g, c = T_1, t = 2020] - E[Y_{it}|G = g, c = T_1, t = 2019]\}\} \\ &- \{E[Y_{it}|G = g, c = C_1, t = 2019] - E[Y_{it}|G = g, c = C_1, t = 2018]\} \quad (2)\end{aligned}$$

We estimate equation (1) for each grade and then average the estimation results for all grade groups for each subject and compare the effect size with those in existing studies.

5.1.2 Impact 19 months after the school closure

Here, we compare the growth in standardized scores in Japanese and math from 2019 to 2021 for the COVID-19 experienced group 2 (T_2) and from 2018 to 2020 for the COVID-19 experienced group 1 (T_1). Unlike the analysis in Section 5.1.1, we divide students in both cohorts from grades 1−6 at $After = 0$ into two grade groups based on their grade at $After = 0$ (lower elementary: grades 1–3, upper elementary: grades 4–6), since the available grades are up to grade 8 as of $After_t = 1$. Moreover, both cohorts experienced the COVID-19 pandemic (T_2: FY2020 and FY2021, T_1: FY2020). Therefore, we remove the impact of COVID-19 included in FY2020 for the T1 cohort using the ATT by grade groups estimated in the equations (1). Here, the T1 and C1 cohorts are each one grade higher than the grade groups defined in Section 5.1.1 (lower elementary: grades 2–4, upper elementary: grades 5–6, junior high school: grade 7) because the interval between $After_t = 0$ and $After_t = 1$ is two years and FY2020, which includes the impact of COVID-19, is the second year. By doing so, we can consider the T_1 cohort as the COVID-19 non-experienced group (control group 2, C_2). The cohorts and timing of the tests used in the estimation are shown in Figure 2.

(Figure 2 around here)

We also divide the students in both cohorts from grades 1−6 at $After = 0$ into three grade groups based on their grade at $After = 0$ (lower elementary: grades 1–3, upper elementary: grades 4–5, and junior high school: grade 6). Here, grade 7 was excluded from the analysis due to a lack of test scores 19 months after school closure. Adding superscript "L" to indicate a long-term effect of the COVID-19 pandemic, the estimation equation for the effect at 19 months after the school closure can be written by three grade groups $g^L \in$ {grades 1–3, grades 4–5, grade 6} as follows:

$$Y_{it}^L = \alpha_i^L + \gamma^L\, COVID19_c^L + \lambda^L\, After_t^L + \delta^L\, COVID19_c^L * After_t^L + \nu_{sg^L}^L + \varepsilon_{it}^L \qquad (3)$$

Now, let Y_{it}^L be the standardized test scores of Japanese and math for student i in year $t \in \{2018 \ldots 2021\}$. $COVID19_c^L$ is a dummy variable that takes 1 if student i is in the treatment group T_2 (0 for the control group C_2). $After_t^L$ is a dummy variable that takes 1 if the test for T_2 was conducted in 2021 and the test for C_2 was conducted in 2020 (0 if the test was conducted 2 years earlier, respectively). α_i^L and ν_{sg}^L are individual fixed effects and school grade fixed effects at $After_t^L = 0$ (T_2 cohort is for FY2019, C_2 cohort is for FY2018), respectively. ε_{it}^L denotes the error term. We assume that $E[\varepsilon_{it}^L | t] = 0$.

The ATT of the COVID-19 pandemic on standardized test scores in Japanese and math for grade group g^L ($\delta_{g^L,ATT}^L$) can be written as follows:

$$\begin{aligned}
\delta_{g,ATT}^L = {} & E[Y_{it} | G = g^L, c = T_2, t = 2021] - E[Y_{it} | G = g^L, c = T_2, t = 2019] \\
& - \{E[Y_{it} | G = g, c = C_2, t = 2019] - E[Y_{it} | G = g, c = C_2, t = 2018]\} \\
& - E[E[Y_{it} | G = g', c = T_1, t = 2020] - E[Y_{it} | G = g', c = T_1, t = 2019] \\
& \quad - \{E[Y_{it} | G = g', c = C_1, t = 2019] - E[Y_{it} | G = g', c = C_1, t = 2018]\}] \\
= {} & \{E[Y_{it} | G = g^L, c = T_2, t = 2021] - E[Y_{it} | G = g^L, c = T_2, t = 2019]\} \\
& - \{[Y_{it} | G = g^L, c = C_2, t = 2020] - E[Y_{it} | G = g^L, c = C_2, t = 2018]\} \\
& + \widehat{\delta_{g',ATT}} \qquad (4)
\end{aligned}$$

Here, the grade group of T1 used to exclude the effects of the COVID-19 pandemic for control group C2 is indicated by $g' \in \{\text{grades 2–4, grades 5–6, grade 7}\}$. Equation (3) is also estimated for each school grade group. We also average the estimated results for three grade groups and compare the effect size with existing studies.

5.1.3 Assumptions for identification

To interpret $\delta_{g,ATT}$ and $\delta^{L}_{g,ATT}$ as causal effects of the COVID-19 pandemic on academic performance, respectively, we need to assume common trends and common shock.

The common trend assumption cannot be directly verified due to insufficient data before FY2018. However, most existing studies have conducted DID using two cohorts over two time periods (Contini et al. 2022; Engzell, Frey, and Verhagen 2021; Haelermans et al. 2022; Lichand et al. 2022). Therefore, to estimate the impact 7 and 19 months after the school closure, we assume the common trends between the T1 and C1 cohorts and the T2 and C2 cohorts for each grade group.

We need two additional assumptions to estimate the impact 19 months after school closure. The first assumption is that the common trend for each grade group holds for the C1 cohort, in addition to the T2 and C2 (= T1) cohorts. This assumption is because the outcome variables for the C2 cohort in FY2020 are post-COVID-19 pandemic, so the impact of COVID-19 in the C2 cohort must be removed using the T1 and C1 cohorts. The second assumption is that the ATT in FY2020 for the C2 cohort is the same within the same grade group. The reason for this assumption is that equations (2) and (4) estimate the ATT 7 months after school closure for each grade group, respectively.

To establish the common shock assumption, it is necessary for no institutional changes affecting outcomes other than COVID-19 to have occurred in both the COVID-19-experienced and non-experienced groups. In Japan, however, the Courses of Study were revised by the MEXT in April 2020. This revision added 26.25 and 52.5 hours of English tuition time per year to grades 3–4 and 5–6, respectively. However, the Japanese and math class times did not change. The common shock assumption seems reasonable, as no other institutional changes affecting the outcomes were implemented between 2018 and 2021.

5.2 Analysis of heterogeneity of effects

5.2.1 Heterogeneity of effects across quartiles of test scores

Existing studies have shown that the negative impact of the COVID-19 pandemic on academic achievement was larger for the lower academic groups (Ardington, Wills, and Kotze 2021; Asakawa and Ohtake 2022; Kuhfeld et al. 2022; Schult et al. 2022a, 2022b). However, the negative impact was larger for the higher academic groups in several countries and regions (Contini et al. 2022; Gambi et al. 2021).

Therefore, we test the hypothesis that "the lower was the students' academic achievement level before the school closure, the greater was the negative impact of the COVID-19 pandemic and the slower the recovery of academic achievement" and compare the results with those of existing studies. Specifically, we use a DID approach using the advantages of the quantile regression (quantile-DID) used by Athey and Imbens (2006) for each subject to test whether the impact of COVID-19 on academic performance differs across quartiles of test scores.

Quantile-DID is performed with quartiles (hereafter, QT) of the outcome excluding the upper and lower 5th percentile (1st QT: 0.05–0.275, 2nd QT: 0.275–0.5, 3rd QT: 0.5––0.725, 4th QT: 0.725–0.95). The equation for estimating the effect of COVID-19 on test scores 7 months after the closure in the q^{th} quartile of standardized test scores for each grade group g can be written as follows according to equation (1):

$$\begin{aligned} Y_{it} = {} & \alpha_i(q) + \gamma(q)COVID19_i + \lambda(q)After_t + \delta(q)COVID19_i * After_t \\ & + \nu_{sg}(q) + \varepsilon_{it}(q) \end{aligned} \tag{5}$$

Denoting Q as the quartile of standardized test scores, the ATT of the COVID-19 pandemic on standardized test scores for grade group g and the q^{th} quartile can be written as follows according to equation (2):

$$\begin{aligned} \delta_{g,ATT}(q) = {} & E[Y_{it}|G = g, c = T_1, Q = q, t = 2020] - E[Y_{it}|G = g, c = T_1, Q = q, t = 2019] \\ & - \{E[Y_{it}|G = g, c = C_1, Q = q, t = 2019] - E[Y_{it}|G = g, c = C_1, Q = q, t = 2018]\} \end{aligned} \tag{6}$$

Using superscript "L" to indicate the long-term effect of the COVID-19 pandemic, the equation for estimating the effect of COVID-19 on test scores 19 months after the closure in the q^{th} quartile of standardized test scores for each grade group g can be written as follows according to equation (3):

$$Y_{it}^{L} = \alpha_{i}^{L}(q) + \gamma^{L}(q)COVID19_{i}^{L} + \lambda^{L}(q)After_{t}^{L} + \delta^{L}(q)COVID19_{i}^{L} * After_{t}^{L} + \nu_{sg^{L}}^{L}(q) + \varepsilon_{it}^{L}(q) \quad (7)$$

Denoting Q as the quartile of standardized test scores, the ATT of the COVID-19 pandemic on standardized test scores for grade group g and the q^{th} quartile can be written as follows according to equation (4):

$$\begin{aligned}
\delta_{g,ATT}^{L}(q) &= E[Y_{it}|G = g^{L}, c = T_2, Q = q, t = 2021] - E[Y_{it}|G = g^{L}, c = T_2, Q = q, t = 2019] \\
&\quad - \{E[Y_{it}|G = g, c = C_2, Q = q, t = 2019] - E[Y_{it}|G = g, c = C_2, Q = q, t = 2018]\} \\
&\quad - E[E[Y_{it}|G = g', c = T_1, Q = q, t = 2020] - E[Y_{it}|G = g', c = T_1, Q = q, t = 2019] \\
&\quad\quad - \{E[Y_{it}|G = g', c = C_1, Q = q, t = 2019] - E[Y_{it}|G = g', c = C_1, Q = q, t = 2018]\}] \\
&= \{E[Y_{it}|G = g^{L}, c = T_2, Q = q, t = 2021] - E[Y_{it}|G = g^{L}, c = T_2, Q = q, t = 2019]\} \\
&\quad - \{[Y_{it}|G = g^{L}, c = C_2, Q = q, t = 2020] - E[Y_{it}|G = g^{L}, c = C_2, Q = q, t = 2018]\} \\
&\quad + \delta_{\widehat{g',ATT}}(q) \qquad (8)
\end{aligned}$$

The standard errors of the parameters of interest, $\delta(q)$ and $\delta^{L}(q)$, in equations (5) and (7) are derived from a nonparametric bootstrap with 300 iterations, respectively. In addition to the assumptions for identification imposed in Section 5.1.3, quantile-DID estimation analysis requires imposing common trend assumptions for the same quartile q of T1 and C1 for analysis 7 months after school closure and T2 and C2 (=T1) and C1 for analysis 19 months after school closure. Additionally, we assume that the distribution of unobserved variables does not change over time, allowing the distribution of unobserved variables to differ between the treatment and control groups.

5.2.2 Heterogeneity of effects by pre-determinant covariates

Previous studies showed that the impact of COVID-19 on academic achievement is highly heterogeneous, not only by country and region, but also by individual characteristics. For example, some studies demonstrated that children from households with lower SES decreased their academic achievement more due to the COVID-19 pandemic (Contini et al. 2022; Gore et al. 2021; Haelermans et al. 2022; Kuhfeld, Lewis, and Peltier 2022; Maldonado and De Witte 2022). Moreover, Haelermans et al. (2022) showed that children from s had lower academic performance due to the COVID-19 pandemic. However, the results of differences in the effects of COVID-19 by child gender have been mixed across countries and regions (Ardington, Wills, and Kotze 2021; Arenas and Gortazar 2022; Birkelund and Karlson 2022; Contini et al. 2022; Hevia et al. 2022).

In line with the previous studies, we examine whether the impact of the COVID-19 pandemic on academic performance differs depending on the pre-pandemic determinant individual characteristics in Amagasaki City. Here, we estimate the following equation, which adds the pre-pandemic pre-determined variables as interaction terms to the full-sample DID defined in equations (1) and (3):

$$\begin{aligned} Y_{it} \; = \; & \gamma^{Hetero}\, COVID19_c + \lambda^{Hetero}\, After_t + \eta^{Hetero}\, Interact_i \\ & + \delta_1^{Hetero}\, COVID19_c * After_t + \delta_2^{Hetero}\, COVID19_c * Interact_i \\ & + \delta_3^{Hetero} After_t * Interact_i + \delta_4^{Hetero}\, COVID19_c * After_t * Interact_i \\ & + \nu_{sg}^{Hetero} + \varepsilon_{it}^{Hetero} \end{aligned} \tag{9}$$

$Interact_i$ is the interaction term of the pre-pandemic pre-determined dummy variables. Specifically, these dummy variables refer to whether the student receives school attendance assistance, whether the student is living with one parent, and whether the student is female. We exclude individual fixed effects from the estimation equation because the dummy variable for the interaction term is constant for individuals over the estimation period. In the analysis 19 months after the school closure, since control group C2 is affected by the COVID-19 pandemic, we add $\widehat{\delta_1^{Hetero}}$ estimated for each grade group $g' \in$ {grades 2–4, grades 5-–6, grade 7} using the T1 and C1 cohorts to the results in equation (9).

5.2.3 Heterogeneity of effects by athletic events reduction

In Amagasaki City, many school events, including athletic events, were reduced to compensate for the lost class time due to the pandemic, as described in Section 3.2. Since most school events in Amagasaki were held before December, the month of the achievement test, the DID coefficient indicating the effect of school closure on academic achievement would be underestimated if academic performance recovered more for the schools and grades that reduced school events and increased class time.

To test whether the COVID-19 pandemic effects on academic performance differ between students with athletic event reductions of more than 200 minutes and others, we perform a DDD estimation by multiplying $COVID19_c * After_t$ by the school- and grade-level athletic event reduction dummies, $Reduc20_{sg}$ and $Reduc21_{sg.}$ These dummy variables take 1 if the reduction time in school s and grade g in 2020 and 2021 is greater than 200 minutes compared to 2019, respectively. Since the reduction hours for athletic events vary across schools and grades, we allow the dummy variables to take different values for different grades rather than grade groups, even within the same school. The estimated equation after 7 months of school closure is as follows:

$$\begin{aligned} Y_{it} \ = \ & \gamma^{DDD}\, COVID19_c + \lambda^{DDD}\, After_t + \eta^{DDD}\, Reduc20_{sg} \\ & + \ \delta_1^{DDD}\, COVID19_c * After_t + \ \delta_2^{DDD}\, COVID19_c * Reduc20_{sg} \\ & + \ \delta_3^{DDD} After_t * Reduc20_{sg} + \ \delta_4^{DDD}\, COVID19_c * After_t * Reduc20_{sg} \\ & + v_{sg}^{DDD} + \ \varepsilon_{it}^{DDD} \end{aligned} \tag{10}$$

We exclude individual fixed effects from equation (10) because $Reduc20_{sg}$ is constant within individuals over the estimation period. We assume that $E\left[\varepsilon_{it}^{DDD} \middle| t\right] = 0$. The impact on students with more than 200 minutes of reduced athletic events from 2019 to 2020 is represented by $\delta_1^{DDD} + \delta_4^{DDD}$, while the impact on students with less than 200 minutes of reduced athletic events is represented by δ_1^{DDD}. Thus, we present only $\widehat{\delta_4^{DDD}}$, the difference between students in schools with Reduc20 = 1 and Reduc20 = 0.

Next, the estimation equation after 19 months of school closure is as follows:

$$\begin{aligned} Y_{it} \ = \ & \gamma^{DDD,L}\, COVID19_c + \lambda^{DDD,L}\, After_t + \eta_1^{DDD,L}\, Reduc20_{sg} + \eta_2^{DDD,L}\, Reduc21_{sg} \\ & + \ \delta_1^{DDD,L}\, COVID19_c * After_t \\ & + \delta_2^{DDD,L}\, COVID19_c * Reduc20_{sg} + \delta_3^{DDD,L}\, COVID19_c * Reduc21_{sg} \\ & + \delta_2^{DDD,L}\, COVID19_c * Reduc20_{sg} + \delta_3^{DDD,L}\, COVID19_c * Reduc21_{sg} \end{aligned}$$

$$+ \delta_4^{DDD,L} After_t * Reduc20_{sg} + \delta_5^{DDD,L} After_t * Reduc20_{sg}$$
$$+ \delta_6^{DDD,L} COVID19_c * After_t * Reduc20_{sg} + \delta_7^{DDD,L} COVID19_c * After_t * Reduc20_{sg}$$
$$+ v_{sg}^{DDD,L} + \varepsilon_{it}^{DDD,L} \quad (11)$$

As in equation (3), since control group C2 has been affected by the COVID-19 pandemic, we add $\widehat{\delta_1^{DDD}}$ estimated in equation (10) for each grade group $g' \in$ {grades 2–4, grades 5–6, grade 7} using the T1 and C1 cohorts to the estimated results in equation (11). We also exclude individual fixed effects from the estimation equation because $Reduc20_{sg}$ and $Reduc21_{sg}$ are constant within individuals over the estimation period. Moreover, we assume that $E\left[\varepsilon_{it}^{DDD,L} | t\right] = 0$.

In equation (11), two different treatment effects are estimated for FY2020 and FY2021, depending on the timing of the reduction in athletic events. Therefore, four estimation results are obtained according to the reduction pattern of athletic events. Since only a few schools had less than 200 minutes of event reduction in 2020 and more than 200 minutes in 2021, we present only $\widehat{\delta_6^{DDD,L}}$, the difference between students in schools with Reduc20 = 1 and Reduc21 = 0 and those with Reduc20 = Reduc21 = 0, and $\widehat{\delta_6^{DDD,L}} + \widehat{\delta_7^{DDD,L}}$, the difference between students in schools with Reduc20 = Reduc21 = 1 and those with Reduc20 = Reduc21 = 0.

6. Estimation results

6.1 Results of the main analysis

Figure 3 and Table 4 present the results of estimating the effect of the COVID-19 pandemic on Japanese language and math standardized test scores. The estimated coefficients 7 months after the school closure are plotted on the left-hand side of the figure and in columns (1)–(5) of the table, and the estimated coefficients 19 months after the closure are plotted on the right-hand side of the figure and columns (6)–(10) of the table.

(Figure 3 and Table 4 around here)

The upper panels of Figure 3 and the first three rows of Table 4 show that, on average, Japanese language scores worsened by 0.006 SD and 0.062 SD at 7 and 19 months after the school closure, respectively. The lower panels of Figure 3 and the last three rows of Table 4 show that, on average, math scores worsened by 0.129 SD and 0.251 SD at 7 and 19 months after the school closure, respectively.

By school grade, test scores for Japanese language in grades 1-3 declined by 0.225 SD at 7 months after school closure and remained unchanged at 0.215 SD decline 19 months after school closure. Their math test scores declined by 0.134 SD 7 months after school closure and declined to -0.205 SD 19 months after school closure. Japanese language test scores in grade groups 4–5 increased by 0.134 SD 7 months after school closure, but no longer had significant positive effects 19 months after school closure. Their math test scores worsened by 0.167 SD 7 months after school closure and further worsened to a 0.321 SD decrease 19 months after school closure. Grade 6 (grade group 6–7 at 7 months after school closure) showed no significant effects on test scores in the Japanese language at both 7 and 19 months after school closure (coefficients of 0.074 SD and 0.033 SD, respectively). Their math scores worsened by 0.086 SD 7 months after school closure and by 0.225 SD 19 months after school closure.

In summary, by subject, the negative effects of the COVID-19 pandemic were greater for math than for the Japanese language at 7 months and the deterioration in math scores was more pronounced than in the Japanese language at 19 months after school closure. By school grade groups, the Japanese language was negatively affected only in the lower grade groups, while the negative impact of math did not differ by grade group.

6.2 Results of the heterogeneity of effects analysis

6.2.1 Heterogeneity of effects across test score quartiles

Figures 4 and 5 and Tables 5 and 6 show the quantile-DID estimates of the effects of the COVID-19 pandemic on standardized test scores in Japanese language and math. For each figure, the upper and lower panels show the effects of school closure 7 and 19 months after the closure, respectively. For each table, columns (1)–(5) and columns (6)–(10) show the effects of school closure 7 and 19 months after the closure, respectively.

(Figures 4 and 5 and Tables 5 and 6 around here)

Figure 4 and Table 5 show that the negative impact of the COVID-19 pandemic on scores in the Japanese language varied by score quartile for grades 1–5. Specifically, grade groups 1–3 had negative and significant coefficients in all quartiles, but the upper quartile was more negatively affected. The coefficients did not change in this group between 7 and 19 months after the school closure. In grade groups 4–5, only the fourth quartile had a negative and significant coefficient, and the negative coefficient increased further from -0.085 SD to -0.338 SD from 7 to 19 months after school closure. However, this group had a positive and significant effect on the 1st–3rd QTs 7 months after school closure and only on the 1st QT 19 months after the closure. In grade 6 (grade group 6–7 for 7 months after school closure), no negative impact was observed in all quartiles of test scores.

Figure 5 and Table 6 show that the negative impact of the COVID-19 pandemic on math test scores varied by score quartile for all grade groups. Specifically, grade groups 1–3 had negative and significant coefficients on the 1st–3rd QTs. Their scores were more negative in the lower quartiles and worsened more from 7 to 19 months after the school closure. However, in this grade group, only the 4th QT had a positive and significant effect. Grade groups 4–5 had negative and significant coefficients in all quartiles, and the negative coefficients increased further from 7 to 19 months after school closure. Grade groups 6–7 had negative and significant coefficients only on the 1st QT 7 months after the closure. Moreover, 19 months after the closure, grade group 6 had negative and significant coefficients on the 1st and 2nd QTs. The negative coefficients on the 1st and 2nd QTs further increased from 7 to 19 months after the closure.

6.2.2 Heterogeneity of effects by pre-determinant covariates

Figures 6–8 and Tables 7–9 show the DID coefficients with interaction terms that include the pre-determinant variables of the SAA receipt dummy, living with one parent dummy, and female dummy, respectively.

(Figures 6–8 and Tables 7–9 around here)

Figures 6 and 7 and Tables 7 and 8 show that no statistically significant differences can be observed in the impact of the COVID-19 pandemic on Japanese language and math scores by living conditions at After = 0, that is, whether the student was receiving

SAA and whether or not the student was living with one parent. However, Figure 8 and Table 9 show that, only in grade 6, female students had a significantly lower negative impact on the Japanese language due to the COVID-19 pandemic than male students. As Figure 3 shows a non-significant, positive effect of 0.074 SD and 0.033 SD for the Japanese language in grade group 6–7 at 7 months after school closure and grade 6 at 19 months after school closure, respectively, this result suggests that females in the upper grades scored significantly higher than males in the Japanese language after the COVID-19 pandemic.

6.2.3 Results of effects of athletic events reduction

Before the analysis, we present the descriptive statistics for students in the schools and grades that reduced athletic events by 200 minutes or more in 2020 and 2021 and others in Table A1. The table shows that students in schools and grade groups that reduced athletic events by more than 200 minutes had a lower academic achievement for treatment group 1 (T1, 2019–2020) and treatment group 2 (T2, 2019–2021). Moreover, for treatment group 1, higher percentages of students in schools and grades of athletic events reduced by over 200 minutes were also in receipt of school attendance assistance and in students living with one parent. Thus, the schools and grades that significantly reduced their athletic events were those with relatively low academic performance and living standards of students before the pandemic. If the negative impact of the COVID-19 pandemic is greater in these schools and grades, the reduction of athletic events may be self-selective.

Therefore, we now use the results of the DDD to determine if the recovery from the negative effects of the pandemic varies between students in the schools and classes with reduced athletic events over 200 minutes and those in other schools and classes. Figure 9 and Tables 10 and 11 show the difference in effects for students in schools with Reduc20 = 1 and Reduc20 = 0 estimated by DDD in equations (10) and (11). The coefficients on the effects 7 months after school closure, estimated using equation (10), are plotted on the left side of Figure 9 and Table 10, and those at 19 months after the closure, estimated using equation (11), are plotted on the right-hand side of Figure 9 and Table 11. To confirm the magnitude of the estimated results, in Figure A6, we also show the treatment effects for students in schools with Reduc20 = 0 estimated by COVID-19×After, that is,

$\widehat{\delta_1^{DDD}}$ in equation (10) and $\widehat{\delta_1^{DDD,L}}$ in equation (11).

(Figure 9 and Tables 10 and 11 around here)

Figure 9 and Tables 10 and 11 show that, for all subjects, periods after the school closure, and grade groups, we find no statistically significant differences in the impact of the COVID-19 pandemic on test scores between students in the schools that reduced athletic events over 200 minutes and others. At 19 months after school closure, several grades and subjects showed differences greater than 0.1 SD, but none were statistically significant at the 5% level. Moreover, Figure A6 shows that the magnitude of these coefficients is smaller than the baseline coefficient of the ATT for students in the school with Reduc20 = 0 for 7 months after the closure and those with Reduc20 = Reduc21 = 0 for 19 months after the closure. Thus, we conclude that schools that drastically reduced the hours of athletic events improved their students' test scores slightly 19 months after school closure, but not enough to counteract the negative effects of the COVID-19 pandemic.

7. Conclusions

We examined whether the COVID-19 pandemic affected the standardized Japanese and math test scores of students in grades 1–7 in all public elementary and junior high schools in Amagasaki using DID estimation.

The analysis compared the 2019–2020 growth of the COVID-19 experienced cohort (treatment group 1) with the 2018–2019 growth of the COVID-19 non-experienced cohort (control group 1). In addition, to analyze the impact 19 months after the school closure, we compared the growth from 2019 to 2021 for the cohort that took the test twice after the school closure (treatment group 2) with the growth from 2018 to 2020 for the cohort that took the test once after the closure (control group 2). Since control group 2 was affected by the school closure in 2020, the effect was removed by subtracting the difference between treatment group 1 and control group 1. To facilitate interpretation, we created three grade groups (lower elementary, upper elementary, and junior high school) and estimated by grade groups. We also performed quantile-DID and DID with interaction terms to check for heterogeneity of effects across test scores and pre-determinant quantiles. Finally, we conducted DDD estimation to identify heterogeneity in effects between students in schools that significantly reduced athletic events after school closure and others.

The results of the main DID estimation showed that, on average, the negative effects of the COVID-19 pandemic were greater for math than for the Japanese language at 7 months and the deterioration in math scores was more pronounced than in the Japanese language at 19 months after school closure. Specifically, Japanese language scores worsened slightly by 0.006 SD and 0.062 SD on average 7 and 19 months after school closure, respectively. Math scores worsened considerably by 0.129 SD and 0.251 SD 7 and 19 months after school closure, respectively. Considering three grade groups, Japanese language scores were negatively affected only in the lower grade groups, but the negative effects on math scores did not differ by grade group. The analysis by the test score quartile revealed that Japanese language scores declined more in the upper quartile, only in elementary school students. By contrast, math scores declined more in the lower quartile in all grades. The results of DID with interaction terms showed that the negative effects of the COVID-19 pandemic varied little depending on the living conditions before school closure and the gender of the students. Finally, the DDD estimation showed that reducing athletic events after school closure contributed little to recovering the scores that declined due to the pandemic.

Compared to the effect sizes of previous studies presented in Table 1, the negative effects 19 months after the school closure in Amagasaki City, shown in Figure 3 and Table 4, are slightly smaller in Japanese (previous study: -0.088 SD, Amagasaki: -0.062 SD) and almost twice as large in math (previous study: -0.139 SD, Amagasaki: -0.251 SD). However, our finding of a larger negative effect in math compared to Japanese language is consistent with previous studies.

By grade group, our finding that the negative effect of Japanese was larger in the lower grade groups and the magnitude of the negative effect on math was similar across grade groups 7 months after the school closure in Amagasaki City is consistent with the evidence presented in Table 1. However, in Amagasaki City, the negative effect of Japanese was larger in grades 1–3 than the extant studies, while no negative effect was observed in grade groups 4–5 and 6–7, unlike in extant studies. In math, the coefficients for the prior study and our findings were almost identical. Specifically, in previous studies, the mean treatment effects for national language scores were -0.14 SD (grades 1–3), -0.073 SD (grades 4–5), and -0.058 SD (grades 6–7). By contrast, in Amagasaki City, 7 months after school closure, the mean treatment effects were -0.225 SD (grades 1–3), 0.135 SD (grades 4–5), and 0.074 SD (grade 6) for national language scores. As for math scores, the mean treatment effects in the previous studies were -0.147 SD (grades 1–3), -0.14 SD (grades 4–5), and -0.129 SD (grades 6–7), while in Amagasaki City, at 7 months after school closure, the mean effects were -0.134 SD (grades 1–3), -0.167 SD (grades 4–5), and -0.086 SD (grade 6).

The negative impact of math in Amagasaki City was significant compared to other municipalities in Japan. For example, Asakawa and Ohtake (2022) show that, in Nara City, Japan, math scores in grades 4 and 5 at the time of school closure had already recovered (0.05 SD for grade 4 and 0.46 SD for grade 5) 7 months after the school closure. However, in Amagasaki, the negative impact remained -0.167 SD for grade group 4–5 during the same period, delaying the recovery by 0.217–0.393 SD compared to Nara City.

Why did the speed of recovery of academic performance differ between the two cities? Since the level of socio-economic activities differs between the two cities, it is not easy to identify the factors contributing to the differences. Therefore, as a discussion, we consider the possibility that the difference in the rate of students attending cram schools, which is directly related to students' educational environment, may have caused the difference in academic achievement recovery between the two cities.

For example, Abe, Ohtake, and Sano (2023) show that the effect of tutoring on standardized math scores in Amagasaki City is 0.37 SD for grade 6. This finding indicates that attending a cram school positively impacts academic achievement. Therefore, in Figure 10, we confirm the difference in the distribution of the tutoring ratio (including the use of private tutors) among grade 6 students in Amagasaki City and Nara City.[4]

(Figure 10 around here)

[4] In both cities, the average rate of attending cram schools is calculated on a school-by-school basis and the number of students is shown as frequency on the vertical axis. Since public data were not available for both cities, we used the data for grade 6 students in all public elementary schools that took the National Assessment of Educational Progress in May 2021 for Nara City and randomly selected grade 6 classes of each school for Amagasaki City in May 2020. Additionally, when calculating the average school attendance rate in Amagasaki City, we assumed that other classes that did not take the survey have the same average school attendance rate and the number of students in each school is estimated by the ratio of the number of students in the class that took the survey to the number of grade 6 students in that school.

This figure shows that, on average, the average rate of students using cram schools is 37.4 percentage points higher in Nara City. Therefore, the difference in out-of-school educational opportunities, such as tutoring, may have mitigated the learning loss due to the COVID-19 pandemic. However, without sufficient data to identify causality, we cannot conduct further analysis and leave this issue for future work.

One possible caveat is the effect of the Courses of Study revision in 2020. As discussed in Section 5.1.3, Japanese and math class time did not change, while the increase in English classes may have affected students' academic performance. Additionally, the Courses of Study revision included curriculum changes and study contents also changed. If test difficulty changes due to the revision of the curriculum, we cannot completely distinguish the effects of the COVID-19 pandemic on academic achievement from the effects of the differences in the curriculum, regardless of the use of the IRT test. Moreover, if the degree of change in the difficulty level differs across grades, it is difficult to accurately estimate the impact of the COVID-19 pandemic on academic achievement for each grade. However, although the revision of the Courses of Study went into full effect in April 2020, at the same time as the COVID-19 pandemic, revisions were likely made in stages since the announcement in 2018. If parents had changed their behavior in advance, the impact of the curriculum revision is expected to be small.

Possible future developments of this study may include analyzing the impact of the pandemic on learning attitudes, such as the learning time and environment. In Japan, there are several studies on the impact of COVID-19 school closures on learning time (Ikeda and Yamaguchi 2021; Nishihata and Kobayashi 2022). Using logs of online learning service use, Ikeda and Yamaguchi (2021) find that students decreased their learning time using these services only during the COVID-19 school closure. They also found that the decline in learning time was heterogeneous across students, with students who had accessed online learning services at home and students in higher-quality schools spending more time learning than others. Nishihata and Kobayashi (2022) show that students in schools with longer COVID-19 closures had less learning time and more screen time, and these effects were more pronounced for students from low-income families, students with lower academic achievement, and elementary school students in single parent households. Due to the differences in the use of extracurricular education, the impact of the COVID-19 pandemic on learning time and environment may differ between Amagasaki City and other municipalities. However, this issue is outside the scope of this study and will be the subject of future research.

References

Abe, M., Ohtake, F., & Sano, S. (2023). The effect of tutoring school on student outcomes. Mimeo.

Ando, M., & Furuichi, M. (2022). The association of COVID-19 employment shocks with suicide and safety net use: An early-stage investigation. *PLoS ONE, 17*(3), 1–26. https://doi.org/10.1371/journal.pone.0264829.

Ardington, C., Wills, G., & Kotze, J. (2021). COVID-19 learning losses: Early grade reading in South Africa. *International Journal of Educational Development, 86*, 102480. https://doi.org/10.1016/j.ijedudev.2021.102480.

Arenas, A., & Gortazar, L. (2022). Learning loss one year after school closures: Evidence from the Basque Country. *Esade Working Paper* No. 277.

Asakawa, S., & Ohtake, F. (2022). Impact of COVID-19 school closures on the cognitive and non-cognitive skills of elementary school students. *The Research Institute of Economy, Trade and Industry (RIETI) Discussion Paper Series* 22-E-075.

Baron, E. J., Goldstein, E. G., & Wallace, C. T. (2020). Suffering in silence: How COVID-19 school closures inhibit the reporting of child maltreatment. *Journal of Public Economics, 190*, 104258. https://doi.org/10.1016/j.jpubeco.2020.104258.

Bazoli, N., Marzadro, S., Schizzerotto, A., & Vergolini, L. (2022). Learning loss and students' social origins during the Covid-19 pandemic in Italy. *FBK-IRVAPP Working*. https://irvapp.fbk.eu/wp-content/uploads/2022/06/WP_IRVAPP_2022-03.pdf.

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2022). A systematic review and meta-analysis of the impact of the COVID-19 pandemic on learning. *SocArXiv*, 1–11.

Birkelund, J. F., & Karlson, K. B. (2022). No evidence of a major learning slide 14 months into the COVID-19 pandemic in Denmark. *European Societies*, 1–21. https://doi.org/10.1080/14616696.2022.2129085.

Blainey, K, & Hannay, T. (2021). The effects of educational disruption on primary school attainment in summer 2021. *RS Assessment from Hodder Education*. August. https://www.risingstars-uk.com/media/Rising-Stars/Assessment/Whitepapers/RSA_Effects_of_disruption_Summer_Aug_2021.pdf.

Contini, D., Di Tommaso, M. L., Muratori, C., Piazzalunga, D., & Schiavon, L. (2022). Who lost the most? Mathematics achievement during the COVID-19 pandemic. *B.E. Journal of Economic Analysis and Policy, 22*(2), 399–408. https://doi.org/10.1515/bejeap-2021-0447.

Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *Prospects, 51*(4), 601–609. https://doi.org/10.1007/s11125-021-09582-6.

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. *Item Response Theory for Psychologists*, 1–371. https://doi.org/10.4324/9781410605269/ITEM-RESPONSE-THEORY-SUSAN-EMBRETSON-STEVEN-REISE.

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences of the United States of America, 118*(17). https://doi.org/10.1073/PNAS.2022376118.

Fukai, T., Ichimura, H., & Kawata, K. (2021). Describing the impacts of COVID-19 on the labor market in Japan until June 2020. *Japanese Economic Review, 72*(3), 439–70. https://doi.org/10.1007/s42973-021-00081-z.

Gambi, L., De Witte, K., De Meyst, M., Derks, A., Groenez, S., Maldonado, J., Chénier, M.-A., Palmaccio, S., & Decoster, A. (2021). The resiliency of school outcomes after the COVID-19 pandemic. Standardised test scores and inequality one year after long term school closures. *FEB Research Report Department of Economics,* Discussion (October).

Gore, J., Fray, L. Miller, A., Harris, J., & Taggart, W. (2021). The impact of COVID-19 on student learning in New South Wales primary schools: An empirical study. *Australian Educational Researcher*, *48*. https://doi.org/10.1007/s13384-021-00436-w.

Haelermans, C., Korthals, R., Jacobs, M., de Leeuw, S., Vermeulen, S., van Vugt, L., Aarts, B., Prokic-Breuer, T., van der Velden, R., van Wetten, S., & de Wolf, I. (2022). Sharp increase in inequality in education in times of the COVID-19-pandemic. *PLoS ONE*, *17*. https://doi.org/10.1371/journal.pone.0261114.

Hevia, F. J., Vergara-Lope, S., Velásquez-Durán, A., & Calderón, D. (2022). Estimation of the fundamental learning loss and learning poverty related to COVID-19 pandemic in Mexico. *International Journal of Educational Development, 88*. https://doi.org/10.1016/j.ijedudev.2021.102515.

Hsu, L. C., & Henke, A. (2021). COVID-19, Staying at home, and domestic violence. *Review of Economics of the Household, 19*(1), 145–155. https://doi.org/10.1007/s11150-020-09526-7.

Jack, R., Halloran, C., Okun, J. C., & Oster, E. (2021). Pandemic schooling mode and student test scores: Evidence from U.S. school districts. *American Economic Review: Insights,* forthcoming, 1–28.

Kuhfeld, M., Lewis, K., & Peltier, T. (2022). Reading achievement declines during the COVID-19 pandemic: Evidence from 5 million U.S. students in grades 3–8. *Reading and Writing*, 0123456789. https://doi.org/10.1007/s11145-022-10345-8.

Kuhfeld, M., Soland, J., Lewis, K., Ruzek, E., & Johnson, A. (2022). The COVID-19 school year: Learning and recovery across 2020–2021. *AERA Open, 8*(1), 1–15. https://doi.org/10.1177/23328584221099306.

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E., & Liu, J. (2020). Projecting the potential impact of COVID-19 school closures on academic achievement. *Educational Researcher, 49*(8), 549–565. https://doi.org/10.3102/0013189X20965918.

Lichand, G., Doria, C. A., Leal-Neto, O., & Cossi Fernandes, J. P. (2022). The impacts of remote learning in secondary education during the pandemic in Brazil. *Nature Human Behaviour, 6*(8), 1079–1086. https://doi.org/10.1038/s41562-022-01350-6.

Maldonado, J. E., & De Witte, K. (2022). The effect of school closures on standardised student test outcomes. *British Educational Research Journal, 48*(1), 49–94. https://doi.org/10.1002/berj.3754.

Ministry of Education, Culture, Sports, Science, and Technology (MEXT). (2021a). Results of the National Survey of School Performance and Learning. https://www.nier.go.jp2.1chousakekkahoukoku/21summary.pdf.

Ministry of Education, Culture, Sports, Science, and Technology (MEXT). (2021b). Results of the National Survey of School Performance and Learning (Supplemental Longitudinal Survey, in Japanese). https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_report.pdf.

Pereda, N., & Díaz-Faes, D. A. (2020). Family violence against children in the wake of COVID-19 pandemic: A review of current perspectives and risk factors. *Child and Adolescent Psychiatry and Mental Health, 14*(1), 1–7. https://doi.org/10.1186/s13034-020-00347-1.

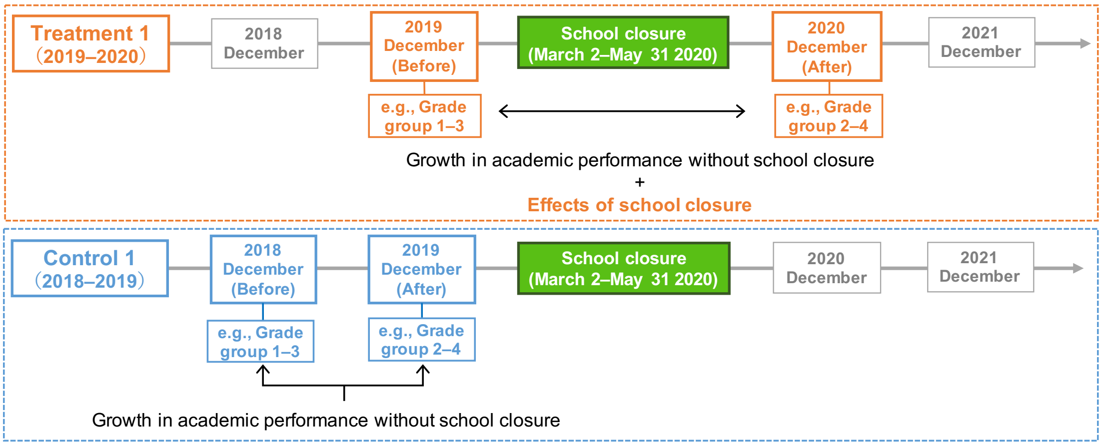
Schult, J., Mahler, N., Fauth, B., & Lindner, M. A. (2022a). Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. *School Effectiveness and School Improvement* 33(4), 544–563. https://doi.org/10.1080/09243453.2022.2061014.

Schult, J., Mahler, N., Fauth, B., & Lindner, M. A. (2022b). Long-term consequences of repeated school closures during the COVID-19 pandemic for reading and mathematics competencies. *Frontiers in Education, 13*, 1–13. https://doi.org/10.3389/feduc.2022.867316.

Tomasik, M. J., Helbling, L. A., & Moser, U. (2021). Educational gains of in-person vs. distance learning in primary and secondary schools: A natural experiment during the COVID-19 pandemic school closures in Switzerland. *International Journal of Psychology*, *56*(4), 566–576. https://doi.org/10.1002/ijop.12728.

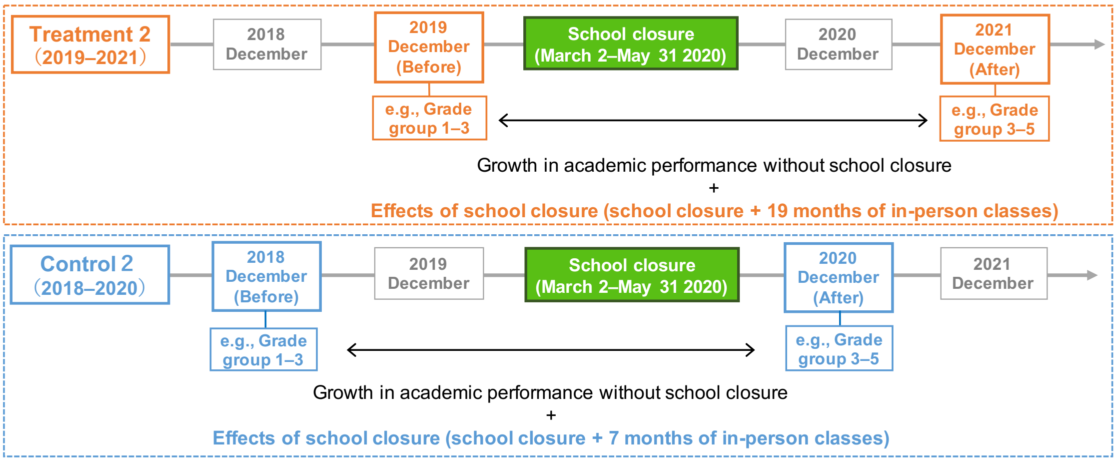
Vegas, E. (2022). COVID-19's impact on learning losses and learning inequality in Colombia. *Center for Universal Education at Brookings, 599*, 1–1.

Figures and Tables



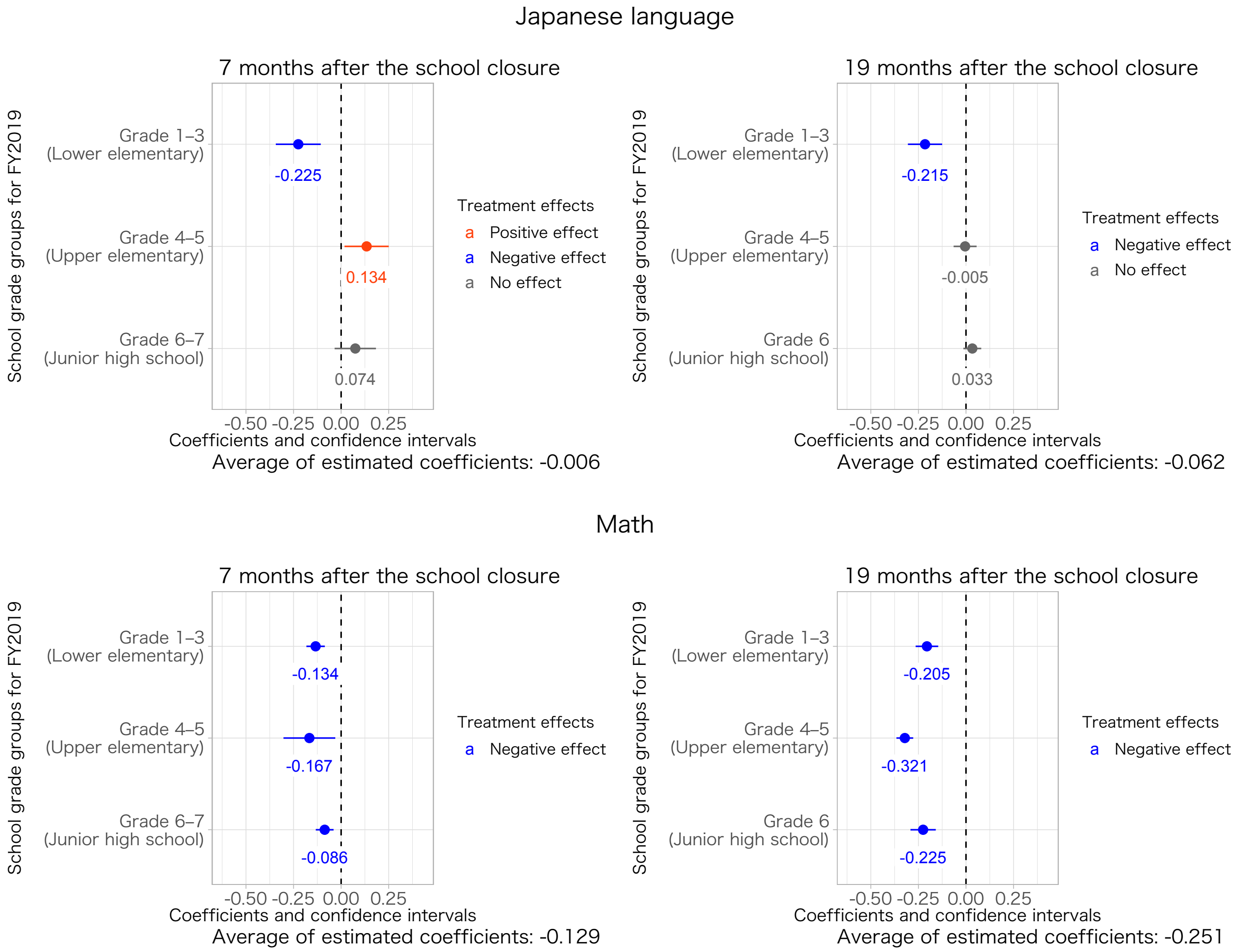
＊ Before (treatment group 1: FY2019, control group 1: FY2018) uses grade groups 1–3, 4–5, 6–7; and After (treatment group 1: FY2020, control group 1: FY2019) uses grade groups 2–4, 5–6, 7–8.

Figure 1 Cohort and test timing for the DID analysis 7 months after school closure



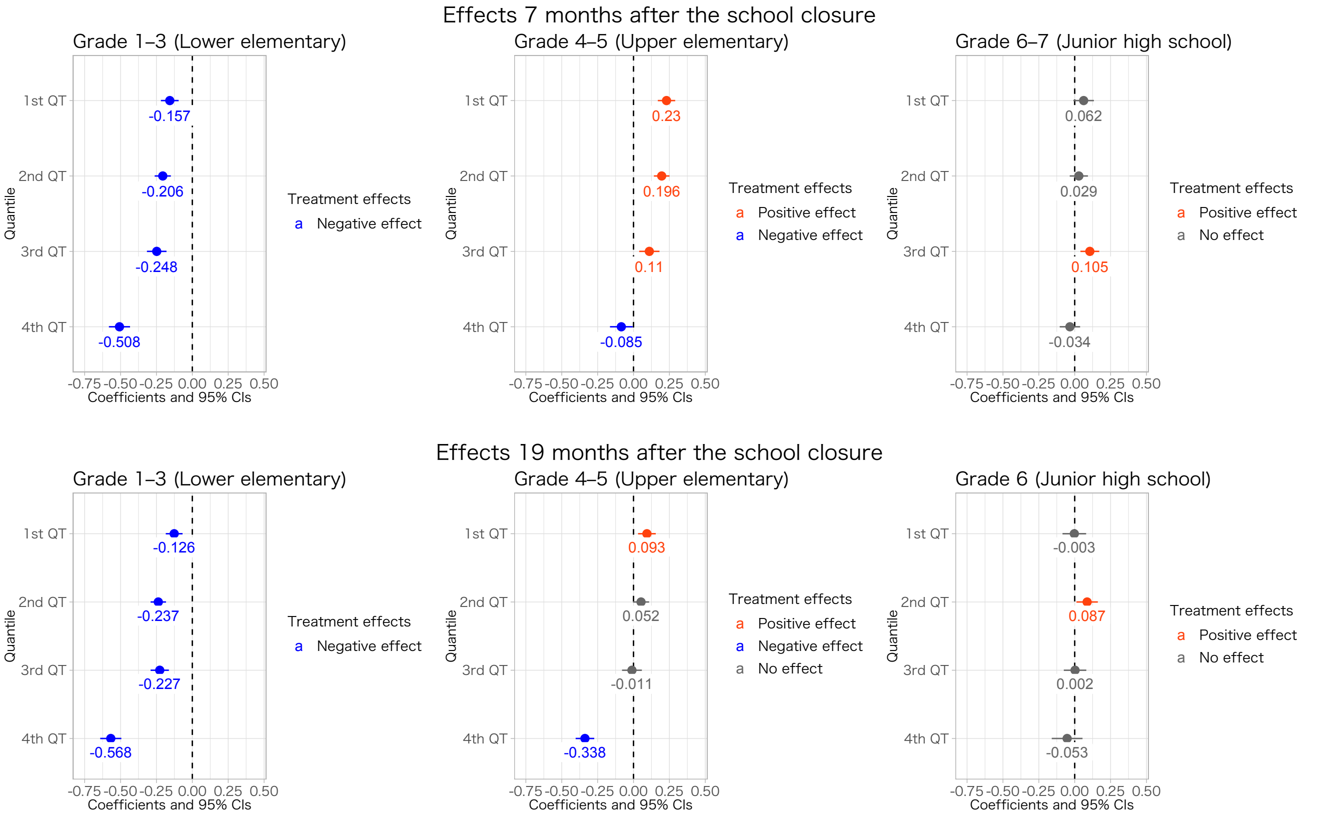
＊ Before (treatment group 2: FY2019, control group 2: FY2018) uses grade groups 1–3, 4–5, 6; and After (treatment group 2: FY2021, control group 2: FY2020) uses grade groups 3–5, 6–7, 8. Subtracting the treatment effect 7 months after school closure expressed in equation (2) from control group 2, we calculate the growth in test scores over the two-year period without school closure.

Figure 2 Cohort and test timing for the DID analysis 19 months after school closure



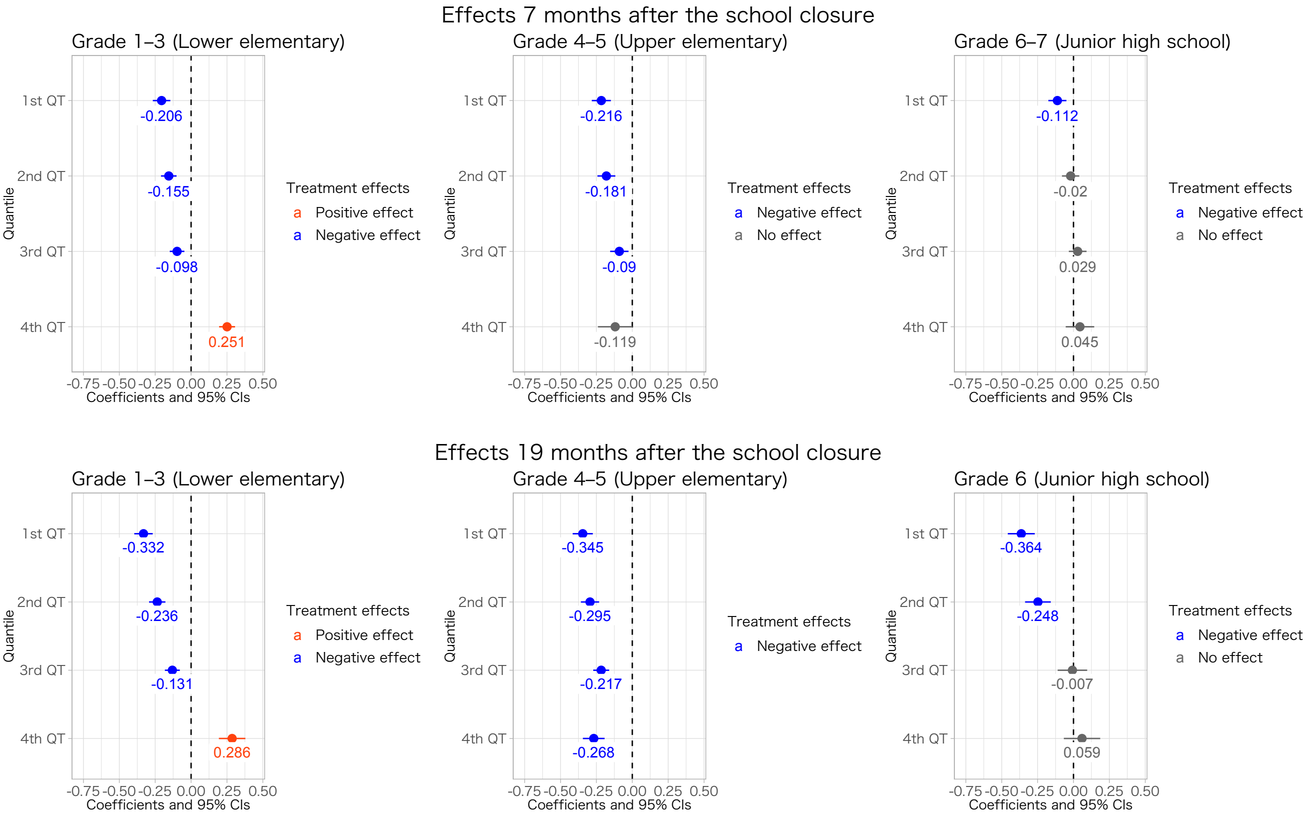
Note: The left and right sides of the figure show ATTs for the three grade groups 7 and 19 months after the school closure, respectively. Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed. "Average of estimated coefficients" represents the averages of ATTs for three school grade groups.

Figure 3 Main results: Full-sample DID



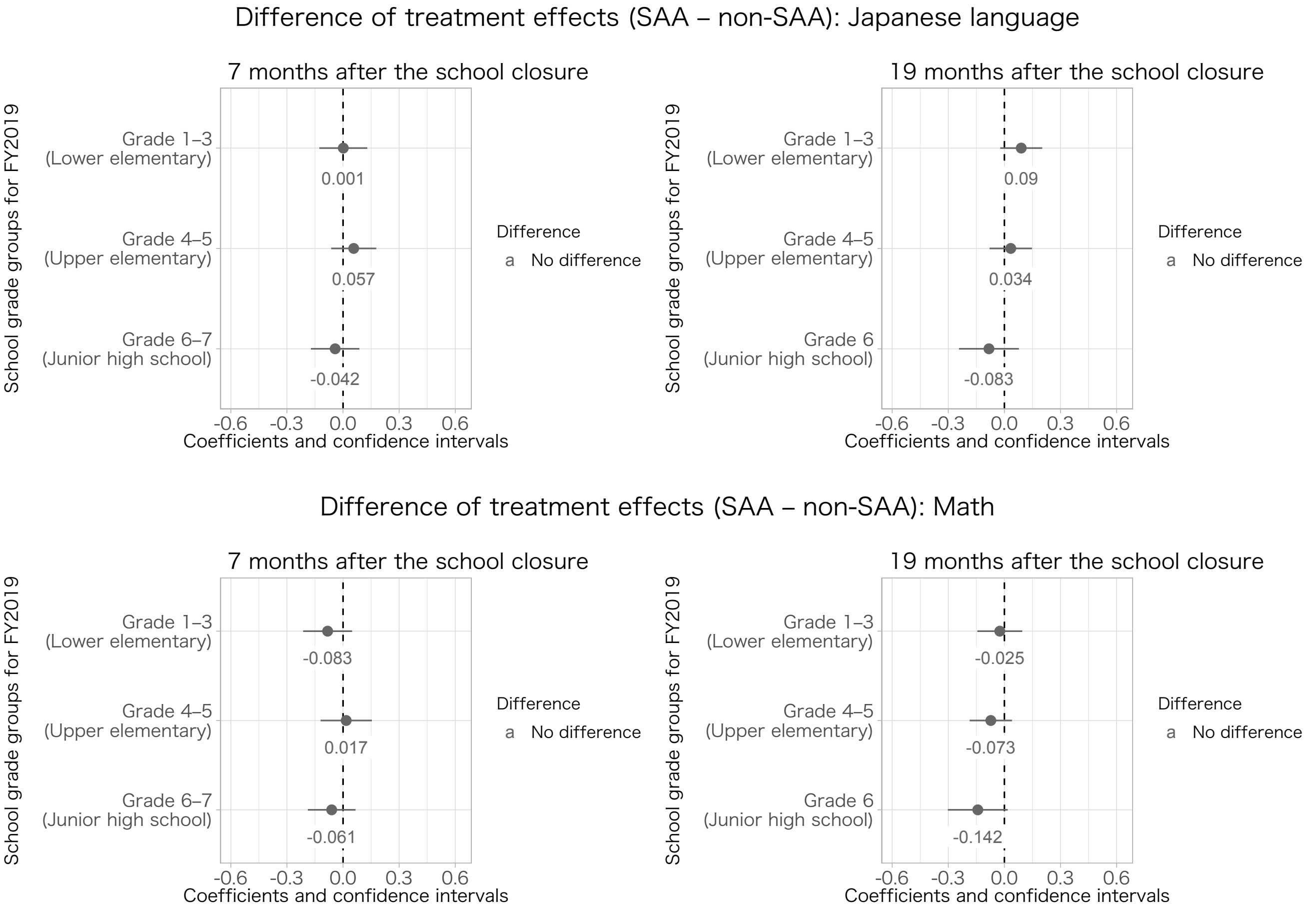
Note: Quartiles were calculated based on 5-95 percentile samples of Japanese language for each grade group and cohort. The upper and lower panels of the figure show ATT by quartile for the three grade groups 7 and 19 months after school closure, respectively. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018) for each quartile, the impacts of COVID-19 pandemic on the control group C2 are removed. The standard errors in the estimates are based on nonparametric bootstrap with 300 replications.

Figure 4 Quantile-DID results: Japanese language



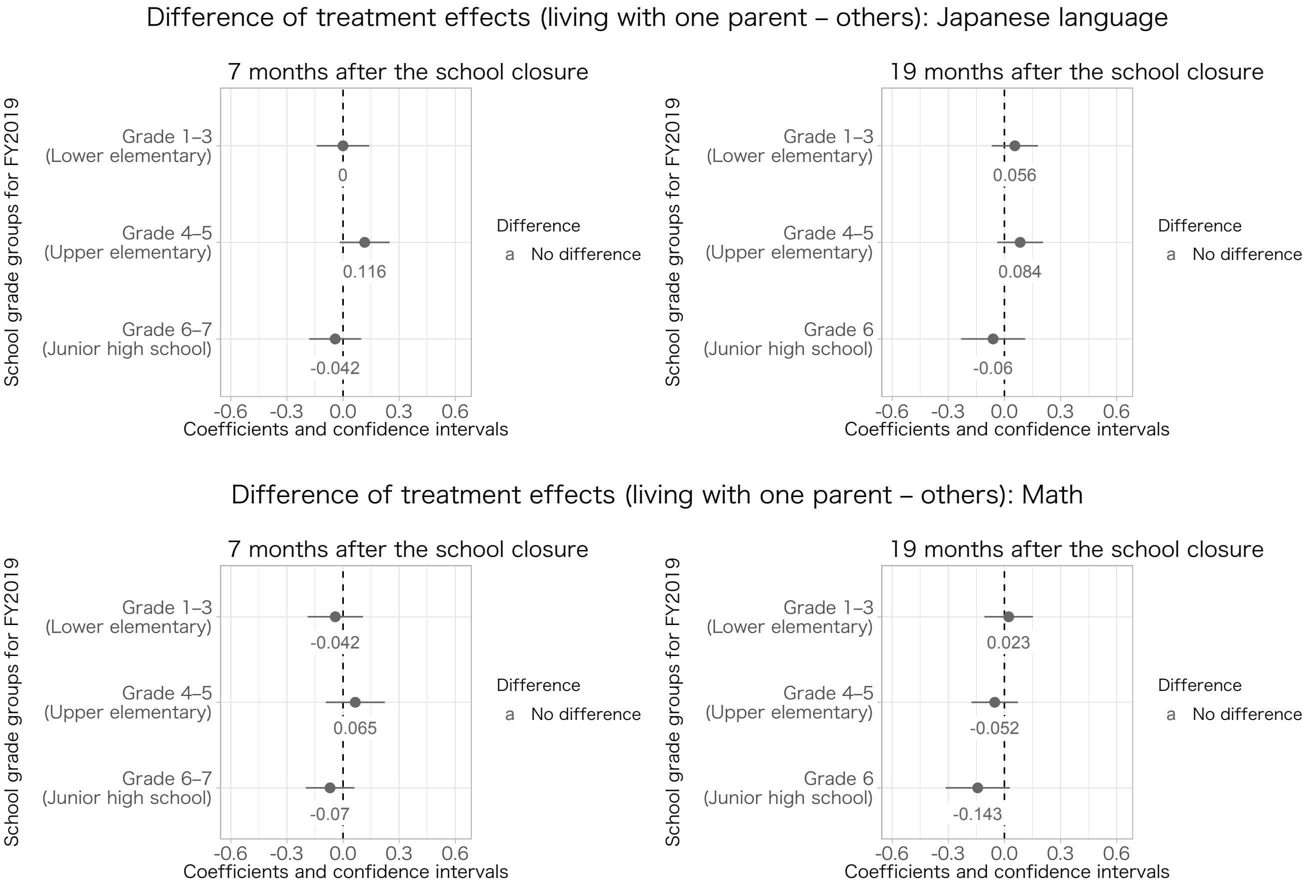
Note: Quartiles were calculated based on 5-95 percentile samples of math for each grade group and cohort. The upper and lower panels of the figure show ATT by quartile for the three grade groups 7 and 19 months after school closure, respectively. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018) for each quartile, the impacts of COVID-19 pandemic on the control group C2 are removed. The standard errors in the estimates are based on nonparametric bootstrap with 300 replications.

Figure 5 Quantile-DID results: Math



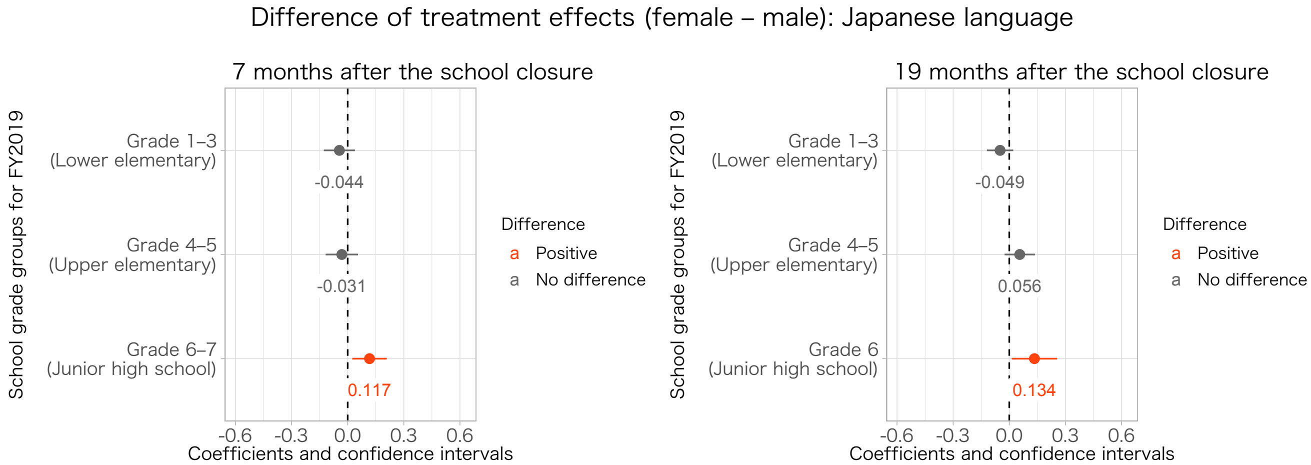
Note: "SAA" means students receiving the school attendance assistance. All figures show the coefficients of "COVID19 x After x SAA", which tests the difference in slope of the treatment effect between SAA = 1 and SAA = 0. If the CIs are significantly away from zero, it means that the treatment effect differs between students with SAA and non-SAA in that grade group. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed.

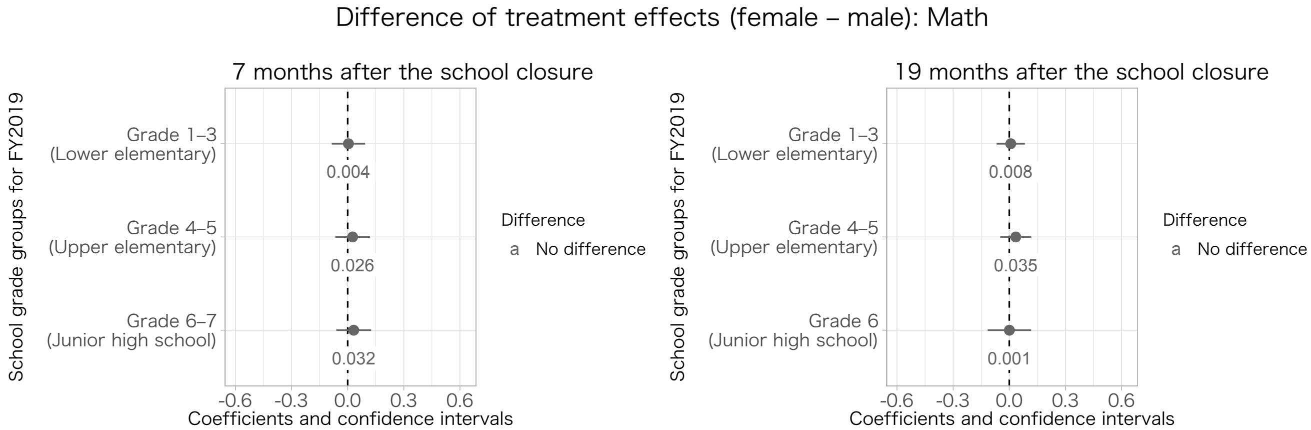
Figure 6 Heterogeneity of treatment effects across groups (students receiving school attendance assistance and others)



Note: "living with one parent" means that the students are either in a single-parent household or one of the parents lives alone outside of Amagasaki City. All figures show the coefficients of "COVID19 x After x living with one parent", which tests the difference in slope of the treatment effect between living with one parent = 1 and living with one parent = 0. If the CIs are significantly away from zero, it means that the treatment effect differs between students living with one parent and others in that grade group. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed.

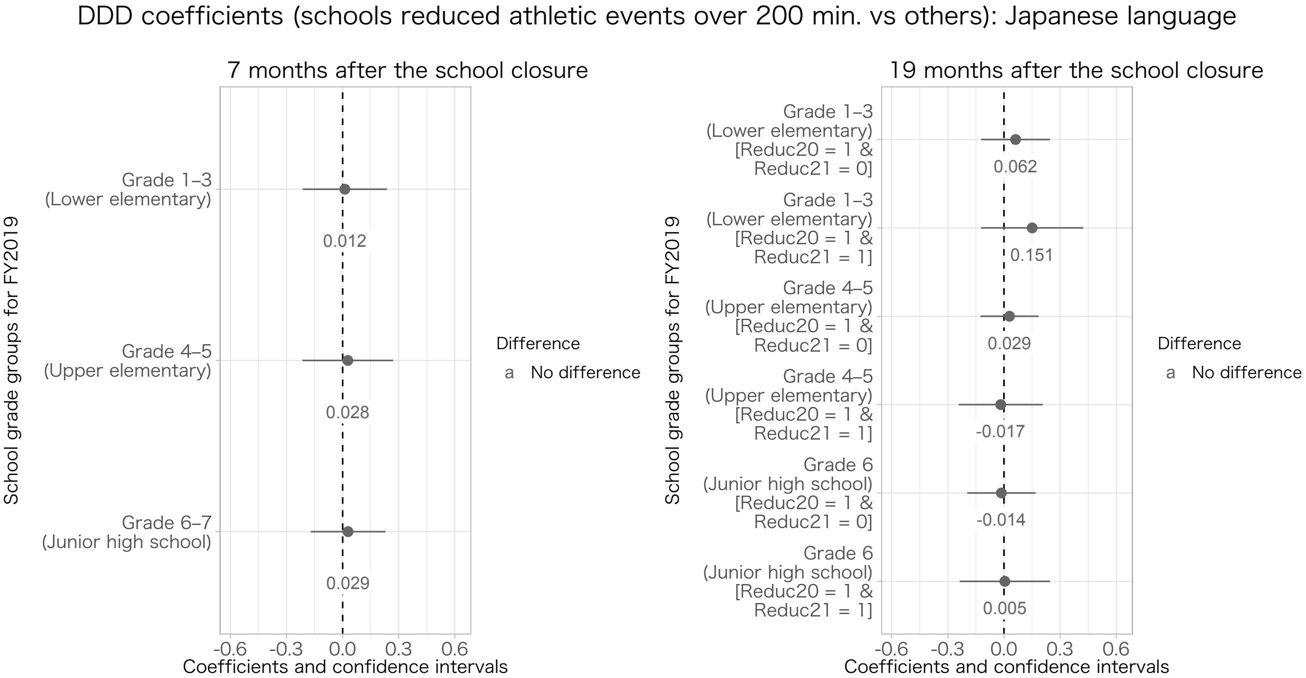
Figure 7 Heterogeneity of effects across groups (students living with one parent and others)

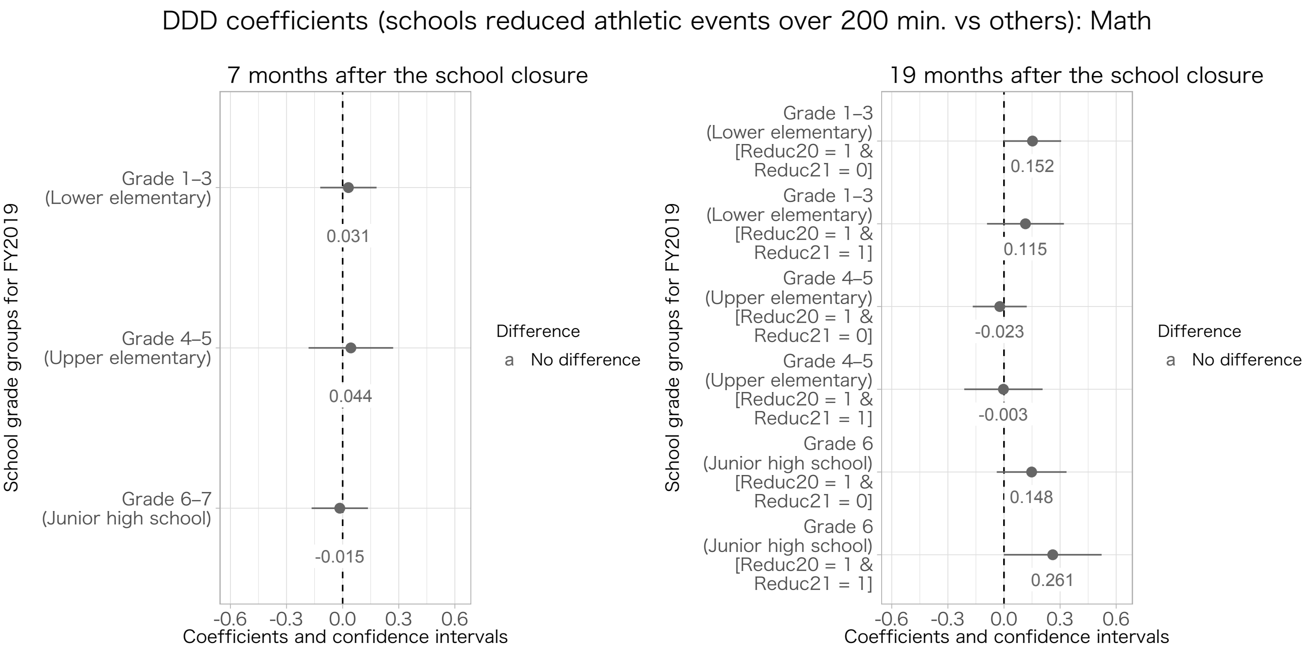




Note: All figures show the coefficients of "COVID19 x After x female", which tests the difference in slope of the treatment effect between female = 1 and female = 0. If the CIs are significantly away from zero, it means that the treatment effect differs between students with female and male in that grade group. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed.

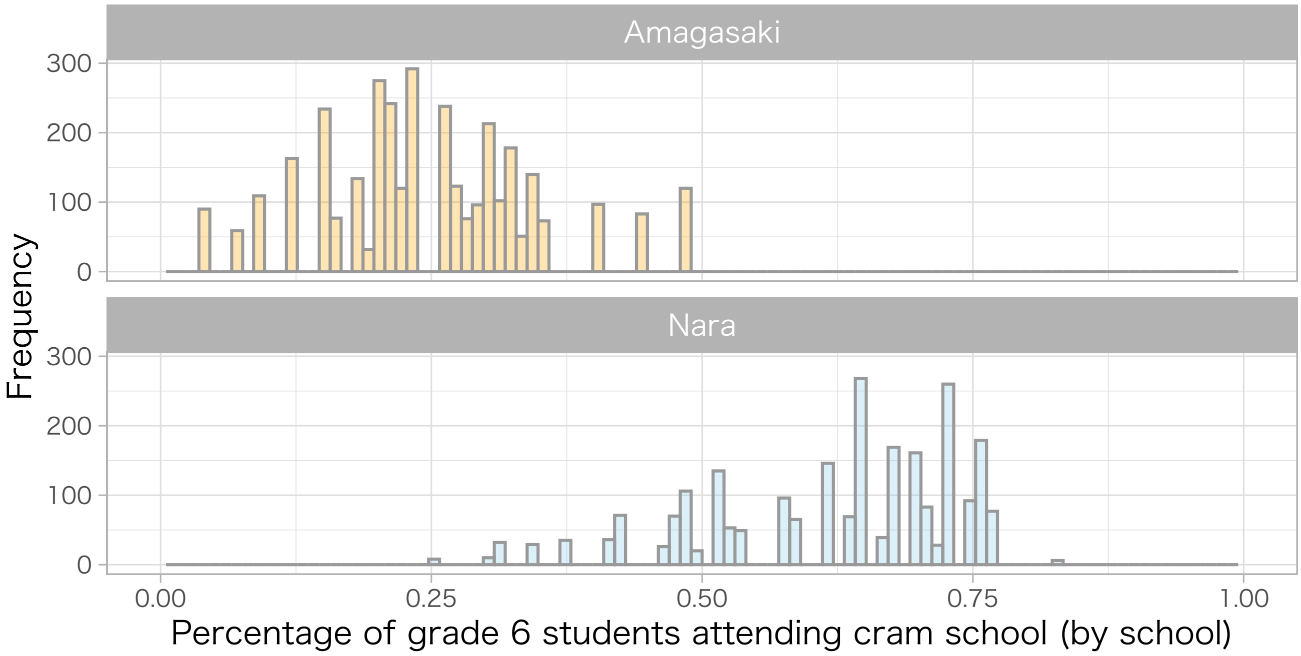
Figure 8 Heterogeneity of effects across groups (female and male students)





Note: "Reduc20" and "Reduc21" are dummy variables that take 1 if the student belongs to a school that reduced athletic events by at least 200 minutes in FY2020 and FY2021, respectively. In the panel of 7 months after school closure, the coefficient of "COVID19 x After x Reduc20" is shown. In the panel of 19 months after school closure, the difference in the effects of COVID-19 pandemic between students in schools with Reduc20 = 1 & Reduc21 = 0 and those with Reduc20 = 0 & Reduc21 = 0 is shown by "COVID19 x After x Reduc20", while the difference in the effects between students in schools with Reduc20 = 1 & Reduc21 = 1 and those with Reduc20 = 0 & Reduc21 = 0 is shown by "COVID19 x After x Reduc20 + COVID19 x After x Reduc21". If the CIs are significantly away from zero, it means that treatment effects differ between students in schools that reduced athletic events by more than 200 minutes and students in other schools. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed.

Figure 9 DDD results (students in the schools reduced athletic events by over 200 minutes and others)



Note: Average cram school attendance rates in Nara and Amagasaki City, weighted by N of 6th graders per school, are 0.621 and 0.247, respectively.
The average cram school attendance rate is calculated for each school in both cities, and the frequency is shown on the y-axis.
For Nara City, we used data from 6th graders in all public elementary schools that took the NAAA in May 2021.
For Amagasaki City, we use data from a randomly selected class of sixth graders in each school in May 2020.
For Amagasaki City, we also assume that the other classes not surveyed have the same rate and calculated frequencies by school using the ratio of the N of students in the class receiving the survey to the N of 6th graders in that school.

Figure 10 The average rate of students using cram schools in Amagasaki and Nara City

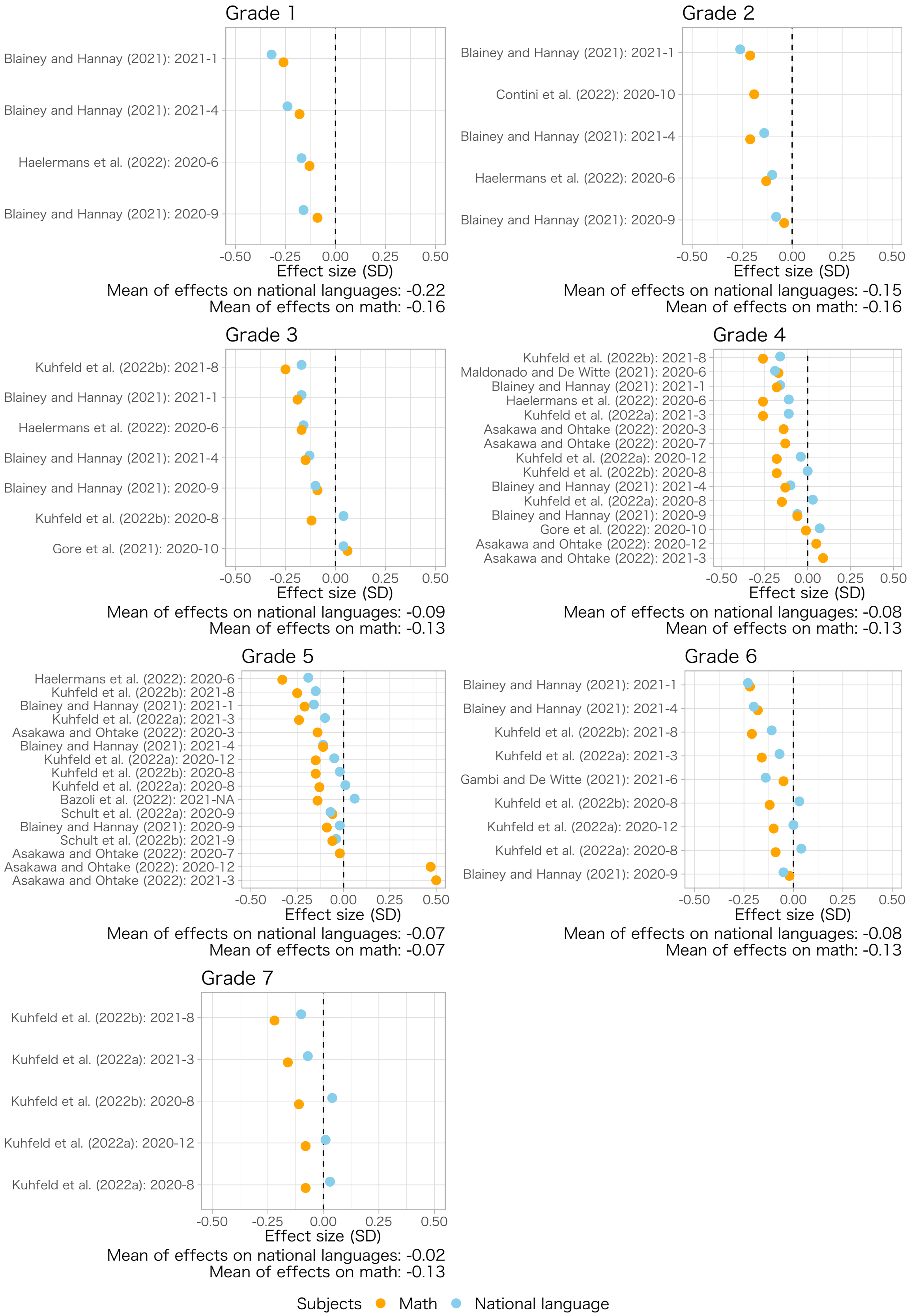


Figure A1 Summary of previous studies (by grade)

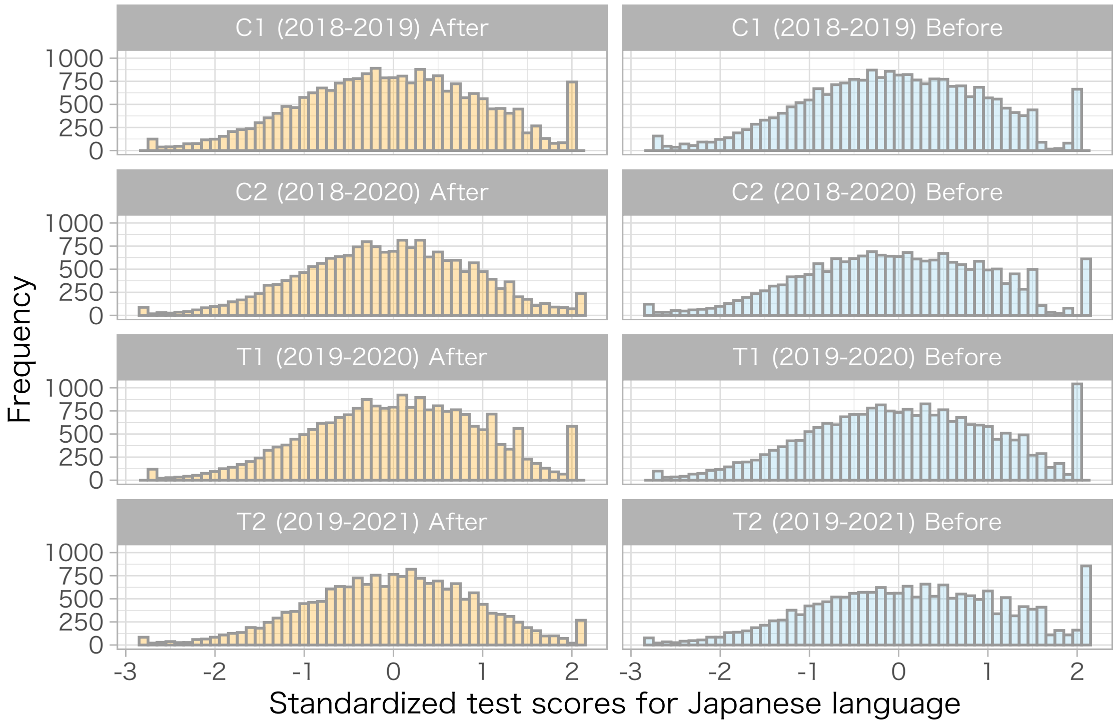


Figure A2 Histogram of standardized test scores in Japanese by cohort and time period

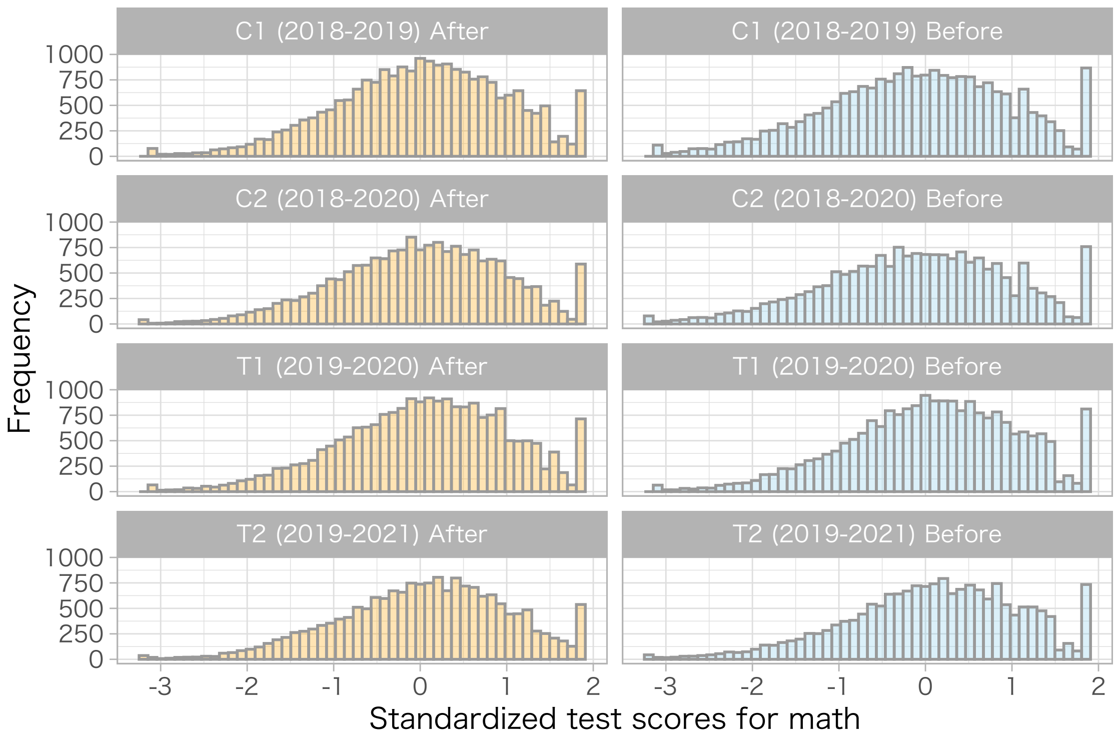


Figure A3 Histogram of standardized test scores in math by cohort and time period

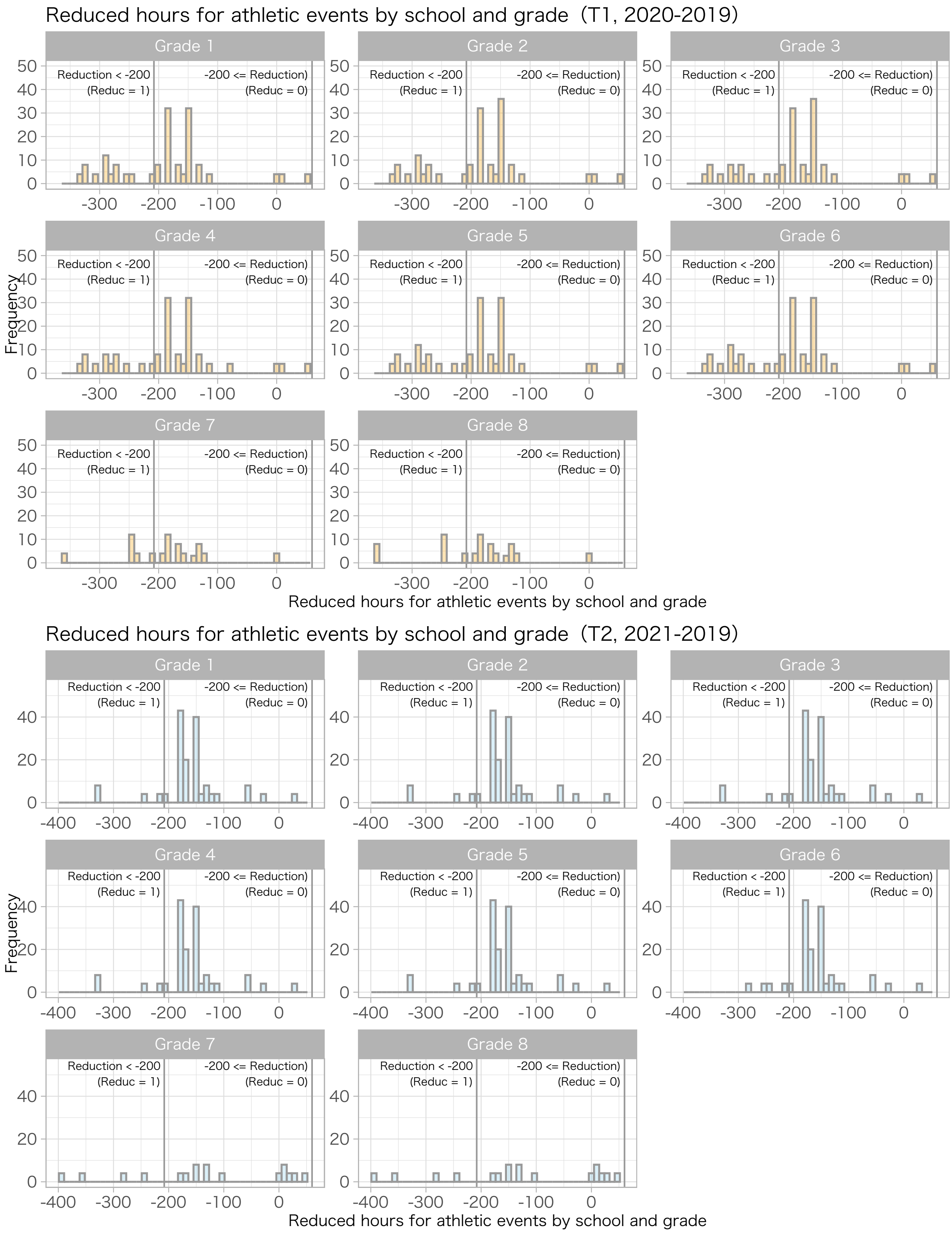
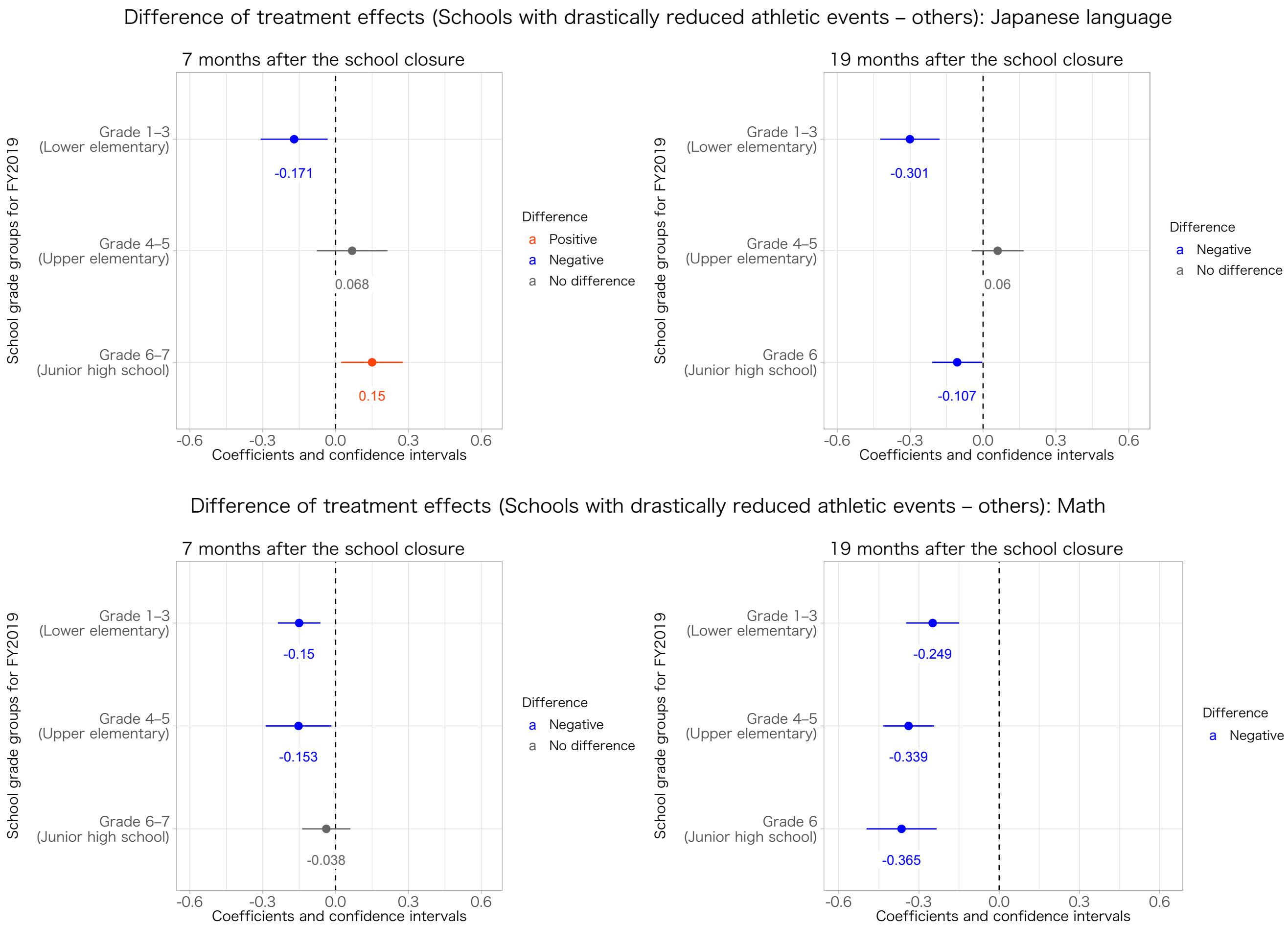


Figure A4 Distribution of reduced hours for athletic events by school and grade

Figure A5 Change in hours of athletic events by school and grade (2018–2021)



Note: "Reduc20" and "Reduc21" are dummy variables that take 1 if the student belongs to a school that reduced athletic events by at least 200 minutes in FY2020 and FY2021, respectively. All panels show the coefficients of "COVID19 x After". In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1–3, 4–5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2–4, 5–6, and 7 in FY2019) and the FY2018 to FY2019 growth of control group C1 (Grade groups 2–4, 5–6, and 7 in FY2018), the impact of COVID-19 on the control group C2 is removed.

Figure A6 DID results (students in the schools reduced athletic events to less than 200 minutes)

Table 1 Summary of previous studies

Country/ reigion	Duration of school closure (weeks)	New school year start month	Test timing after the closure (month)	Months between school closure and the next test	Test year	Effect size (SD) National language	Math	Author(s)
Grade 1								
Netherlands	8	8	6-7	2-3	2020	-0.166	-0.131	Haelermans et al (2022)
UK	10	9	9-12	4-7	2020	-0.155	-0.09	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.315	-0.26	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.245	-0.18	Blainey, Hannay (2021)
Grade 2								
Netherlands	8	8	6-7	2-3	2020	-0.096	-0.129	Haelermans et al (2022)
Italy	15	9	10	4	2020	-	-0.19	Contini et al. (2022)
UK	10	9	9-12	4-7	2020	-0.08	-0.04	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.255	-0.21	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.14	-0.21	Blainey, Hannay (2021)
Grade 3								
Netherlands	8	8	6-7	2-3	2020	-0.158	-0.166	Haelermans et al (2022)
Australia	8	1	10-12	6-8	2020	0.04	0.06	Gore et al. (2021)
US	-	7	8-11	-	2020	0.045	-0.122	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2021	-0.173	-0.248	Kuhfeld, Lewis, Peltier (2022)
UK	10	9	9-12	4-7	2020	-0.095	-0.09	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.175	-0.19	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.13	-0.15	Blainey, Hannay (2021)
Grade 4								
Netherlands	8	8	6 - 7	2-3	2020	-0.107	-0.265	Haelermans et al (2022)
Australia	8	1	10-12	6-8	2020	0.07	-0.01	Gore et al. (2022)
Belgium	9	9	6	1	2020	-0.19	-0.17	Maldonado, De Witte (2021)
US	-	7	8-11	-	2020	-0.004	-0.181	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2021	-0.163	-0.262	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2020	0.03	-0.15	Kuhfeld et al (2022)
US	-	7	12-3	-	2020	-0.04	-0.18	Kuhfeld et al (2022)

US	-	7	3 (late)-6	-	2021	-0.11	-0.26	Kuhfeld et al (2022)
UK	10	9	9-12	4-7	2020	-0.055	-0.06	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.16	-0.18	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.1	-0.13	Blainey, Hannay (2021)
Japan	12	4	3	-3	2020	-	-0.14	Asakawa, Ohtake (2022)
Japan	12	4	7	2	2020	-	-0.13	Asakawa, Ohtake (2022)
Japan	12	4	12	7	2020	-	0.05	Asakawa, Ohtake (2022)
Japan	12	4	3	10	2021	-	0.09	Asakawa, Ohtake (2022)
Grade 5								
Netherlands	8	8	6 - 7	2-3	2020	-0.19	-0.326	Haelermans et al (2022)
Germany	8	9	9	5	2020	-0.07	-0.06	Schult et al. (2022a)
Germany	8	9	9	17	2021	-0.045	-0.063	Schult et al. (2022b)
Italy	15	9	-	-	2021	0.057	-0.142	Bazoli et al. (2022)
US		7	8-11	-	2020	-0.024	-0.151	Kuhfeld, Lewis, Peltier (2022)
US		7	8-11	-	2021	-0.154	-0.247	Kuhfeld, Lewis, Peltier (2022)
US		7	8-11	-	2020	0.01	-0.13	Kuhfeld et al (2022)
US		7	12-3	-	2020	-0.05	-0.15	Kuhfeld et al (2022)
US		7	3 (late)-6	-	2021	-0.1	-0.24	Kuhfeld et al (2022)
UK	10	9	9-12	4-7	2020	-0.02	-0.09	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.16	-0.21	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.105	-0.11	Blainey, Hannay (2021)
Japan	12	4	3	-3	2020	-	-0.14	Asakawa, Ohtake (2022)
Japan	12	4	7	2	2020	-	-0.02	Asakawa, Ohtake (2022)
Grade 6								
Belgium	9	9	6 (2020&21)	1 & 13	2020-21	-0.14	-0.05	Gambi, De Witte (2021)
US	-	7	8-11	-	2020	0.029	-0.123	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2021	-0.11	-0.213	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2020	0.04	-0.09	Kuhfeld et al (2022)
US	-	7	12-3	-	2020	0	-0.1	Kuhfeld et al (2022)
US	-	7	3 (late)-6	-	2021	-0.07	-0.16	Kuhfeld et al (2022)
UK	10	9	9-12	4-7	2020	-0.045	-0.02	Blainey, Hannay (2021)
UK	20-24	9	1-4	8-11 (0-4)	2021	-0.235	-0.22	Blainey, Hannay (2021)
UK	20-24	9	4-7	12-15 (5-8)	2021	-0.195	-0.18	Blainey, Hannay (2021)

Grade 7								
US	-	7	8-11	-	2020	0.045	-0.11	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2021	-0.095	-0.216	Kuhfeld, Lewis, Peltier (2022)
US	-	7	8-11	-	2020	0.03	-0.08	Kuhfeld et al (2022)
US	-	7	12-3	-	2020	0.01	-0.08	Kuhfeld et al (2022)
US	-	7	3 (late)-6	-	2021	-0.07	-0.16	Kuhfeld et al (2022)
Mean: Grades 1-7						-0.088	-0.139	
Mean: Grades 1-3						-0.14	-0.147	
Mean: Grades 4-5						-0.073	-0.14	
Mean: Grades 6-7						-0.058	-0.129	

Note: "-" means that no information was available. "Mean" is the average of the effect size per grade group and subject

UK had a second school closure between December 2020 and January 2021, so the period between the second school closure and the test is shown between parentheses.

In the US, the period of school closure varies from state to state, so "-" is placed in the column for the period of school closure and the interval between school closure and test.

Table 2 Descriptive statistics of Japanese language and math (by cohort, period, and grade groups)

		Analysis for effects 7 months after the school closure (T1 vs C1)																							
		Grade group 1–3 (Lower elementary)								Grade group 4–5 (Upper elementary)								Grade group 6–7 (Junior high)							
		Before (Grade 1-3)				After (Grade 2-4)				Before (Grade 4-5)				After (Grade 5-6)				Before (Grade 6-7)				After (Grade 7-8)			
Variable	Cohort	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value
Test score (J)	C1 (2018–19)	9721	58.57	23.19	< 0.001	9725	57.87	22.79	0.737	6714	51.70	18.05	0.952	6713	57.78	20.32	<0.001	5490	57.93	18.52	0.085	5486	53.69	17.88	< 0.001
	T1 (2019–20)	9438	63.17	23.30		9433	57.76	21.17		6369	51.68	18.93		6369	60.56	19.64		5532	58.55	19.21		5530	55.81	17.86	
Standardized scores (J)	C1 (2018–19)	9721	0.06	1.11	< 0.001	9725	0.02	1.09	0.737	6714	-0.27	0.86	0.952	6713	0.02	0.97	<0.001	5490	0.03	0.89	0.085	5486	-0.18	0.86	< 0.001
	T1 (2019–20)	9438	0.28	1.12		9433	0.02	1.01		6369	-0.27	0.91		6369	0.15	0.94		5532	0.06	0.92		5530	-0.07	0.86	
Test score (M)	C1 (2018–19)	9721	65.27	21.73	< 0.001	9725	66.00	19.51	0.914	6714	54.52	20.64	<0.001	6713	60.06	19.84	0.005	5490	58.26	18.46	<0.001	5486	59.30	17.62	< 0.001
	T1 (2019–20)	9438	67.97	20.35		9433	66.04	19.88		6369	58.82	19.48		6369	61.02	19.53		5532	61.70	16.30		5530	60.98	17.38	
Standardized scores (M)	C1 (2018–19)	9721	0.15	1.09	< 0.001	9725	0.18	0.98	0.914	6714	-0.39	1.04	<0.001	6713	-0.11	0.99	0.005	5490	-0.20	0.93	<0.001	5486	-0.15	0.88	< 0.001
	T1 (2019–20)	9438	0.28	1.02		9433	0.19	1.00		6369	-0.18	0.98		6369	-0.07	0.98		5532	-0.03	0.82		5530	-0.07	0.87	

		Analysis for effects 19 months after the school closure (T2 vs C2)																							
		Grade group 1–3 (Lower elementary)								Grade group 4–5 (Upper elementary)								Grade group 6 (Junior high)							
		Before (Grade 1-3)				After (Grade 3-5)				Before (Grade 4-5)				After (Grade 6-7)				Before (Grade 6)				After (Grade 8)			
Variable	Cohort	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value
Test score (J)	C2 (2018–20)	9487	58.84	23.07	< 0.001	9519	51.11	18.41	<0.001	6186	51.18	17.64	0.652	6202	60.98	19.39	0.469	2703	61.60	17.82	< 0.001	2715	60.31	17.34	0.066
	T2 (2019–21)	9295	63.07	23.37		9288	53.28	18.04		5797	51.33	18.78		5789	60.72	20.22		2679	66.94	17.11		2669	59.44	17.41	
Standardized scores (J)	C2 (2018–20)	9487	0.08	1.13	< 0.001	9519	-0.30	0.90	<0.001	6186	-0.29	0.86	0.652	6202	0.19	0.95	0.471	2703	0.22	0.87	< 0.001	2715	0.15	0.85	0.065
	T2 (2019–21)	9295	0.29	1.15		9288	-0.19	0.88		5797	-0.29	0.92		5789	0.17	0.99		2679	0.48	0.84		2669	0.11	0.85	
Test score (M)	C2 (2018–20)	9487	65.59	21.53	< 0.001	9519	60.88	20.08	0.001	6186	53.94	20.02	<0.001	6202	65.92	16.90	0.274	2703	58.41	16.40	< 0.001	2715	61.17	18.58	<0.001
	T2 (2019–21)	9295	67.82	20.48		9288	61.85	20.01		5797	58.49	19.16		5789	66.27	17.47		2679	65.50	15.29		2669	63.13	18.66	
Standardized scores (M)	C2 (2018–20)	9487	0.14	1.09	< 0.001	9519	-0.09	1.01	0.001	6186	-0.44	1.01	<0.001	6202	0.16	0.85	0.275	2703	-0.22	0.83	< 0.001	2715	-0.08	0.94	<0.001
	T2 (2019–21)	9295	0.26	1.03		9288	-0.04	1.01		5797	-0.21	0.97		5789	0.18	0.88		2679	0.14	0.77		2669	0.02	0.94	

Note: In analysis for effects 7 months after school closure, 'Before' means FY 2018 for the C1 cohort and FY 2019 for the T1 cohort and 'After' means FY 2019 for the C1 cohort and FY 2020 for the T1 cohort. In analysis for effects 19 months after the school closure, 'Before' means FY 2018 for the C2 cohort and FY FY2020 for the C2 cohort and 'After' means FY2020 for the C2 cohort and FY2021 for the T2 cohort. Here, descriptive statistics for each variable at Before and After are listed together by grade groups for each cohort. 'Obs.' means the number of observations. (J) and (M) represent Japanese and math, respectively. 'p.value' means the p-value in the t-test of the difference between the means of the two cohorts.

Table 3 Descriptive statistics of pre-determinant variables and athletic event reduction dummies (by cohort, period, and grade groups)

		Grade group 1–3 (Lower elementary)				Grade group 4–5 (Upper elementary)				Grade group 6–7 (Junior high)			
Variable	Cohort	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value	Obs.	Mean	S.D.	p-value
Analysis for effects 7 months after the school closure (T1 vs C1)													
School attendance assistance (SAA)	C1 (2018–19)	9721	0.19	0.39	<0.001	6714	0.22	0.41	<0.001	5490	0.24	0.43	<0.001
	T1 (2019–20)	9438	0.15	0.36		6369	0.17	0.38		5532	0.20	0.40	
Living with one parent	C1 (2018–19)	9721	0.13	0.34	0.612	6714	0.17	0.37	0.047	5490	0.18	0.38	0.313
	T1 (2019–20)	9437	0.13	0.34		6368	0.15	0.36		5532	0.17	0.38	
Female student	C1 (2018–19)	9721	0.50	0.50	0.579	6714	0.50	0.50	0.345	5490	0.48	0.50	0.111
	T1 (2019–20)	9438	0.49	0.50		6369	0.51	0.50		5532	0.50	0.50	
Reduction over 200 min.	C1 (2018–19)	9416	0.37	0.48	0.446	6503	0.38	0.48	0.797	5403	0.38	0.49	0.583
(between FY2019 and FY2020)	T1 (2019–20)	9126	0.36	0.48		6183	0.37	0.48		5440	0.39	0.49	
Analysis for effects 19 months after the school closure (T2 vs C2)													
School attendance assistance (SAA)	C2 (2018–20)	9487	0.19	0.39	<0.001	6186	0.22	0.41	<0.001	2703	0.23	0.42	<0.001
	T2 (2019–21)	9295	0.15	0.36		5797	0.18	0.38		2679	0.19	0.39	
Living with one parent	C2 (2018–20)	9487	0.13	0.34	0.695	6186	0.16	0.37	0.127	2703	0.17	0.38	0.243
	T2 (2019–21)	9294	0.13	0.33		5796	0.15	0.36		2679	0.16	0.37	
Female student	C2 (2018–20)	9487	0.50	0.50	0.229	6186	0.50	0.50	0.692	2703	0.50	0.50	0.889
	T2 (2019–21)	9295	0.49	0.50		5797	0.50	0.50		2679	0.49	0.50	
Reduction over 200 min.	C2 (2018–20)	9196	0.37	0.48	0.287	5992	0.38	0.49	0.528	2618	0.39	0.49	0.824
(between FY2019 and FY2020)	T2 (2019–21)	8986	0.36	0.48		5628	0.37	0.48		2598	0.39	0.49	
Reduction over 200 min.	C2 (2018–20)	9400	0.14	0.34	0.504	6129	0.14	0.35	0.982	2675	0.12	0.32	0.017
(between FY2018 and FY2021)	T2 (2019–21)	9210	0.13	0.34		5737	0.14	0.34		2648	0.14	0.35	

Note: 'Obs.' means the number of observations. (J) and (M) represent Japanese and math, respectively. 'p.value' means the p-value in the t-test of the difference between the means of the two cohorts.

Table 4 Main full-sample DID results: Effects of COVID-19 pandemic on standardized Japanese and math test scores

	Effects 7 months after school closure					Effects 19 months after school closure				
				Confidence intervals					Confidence intervals	
Grade groups	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Japanese language										
Grade 1-3 (Lower elementary)	**-0.225**	0.060	**<0.001**	-0.343	-0.107	**-0.215**	0.046	**<0.001**	-0.305	-0.125
Grade 4-5 (Upper elementary)	**0.134**	0.059	**0.024**	0.018	0.250	-0.005	0.031	0.88	-0.065	0.056
Grade 6-7 (Junior high)	0.074	0.055	0.178	-0.034	0.183	0.033	0.024	0.166	-0.014	0.080
Math										
Grade 1-3 (Lower elementary)	**-0.134**	0.025	**<0.001**	-0.182	-0.086	**-0.205**	0.030	**<0.001**	-0.265	-0.146
Grade 4-5 (Upper elementary)	**-0.167**	0.069	**0.016**	-0.303	-0.031	**-0.321**	0.022	**<0.001**	-0.365	-0.277
Grade 6 (Junior high)	**-0.086**	0.024	**<0.001**	-0.133	-0.039	**-0.225**	0.034	**<0.001**	-0.292	-0.159

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include individual fixed effects and pre-treatment grade dummies. ‘Est‘ means the estimated coefficients of interaction term ‘COVID19 × After‘. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. ’Est.’ and ’p-value’ are bolded for $p < 0.05$. ‘S.E.‘ presents cluster-robust standard error results at the classroom level.

Table 5 Quantile-DID results: Japanese language

	Effects 7 months after SC					Effects 19 months after SC				
				Confidence interval					Confidence interval	
Quantile	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Grade 1-3 (Lower elementary)										
1st QT	**-0.157**	0.030	**<0.001**	-0.215	-0.099	**-0.126**	0.029	**<0.001**	-0.182	-0.070
2nd QT	**-0.206**	0.029	**<0.001**	-0.263	-0.149	**-0.237**	0.028	**<0.001**	-0.292	-0.182
3rd QT	**-0.248**	0.035	**<0.001**	-0.317	-0.179	**-0.227**	0.033	**<0.001**	-0.292	-0.162
4th QT	**-0.508**	0.040	**<0.001**	-0.586	-0.429	**-0.568**	0.038	**<0.001**	-0.642	-0.494
Grade 4-5 (Upper elementary)										
1st QT	**0.23**	0.030	**<0.001**	0.170	0.289	**0.093**	0.035	**0.008**	0.025	0.162
2nd QT	**0.196**	0.028	**<0.001**	0.142	0.250	0.052	0.030	0.078	-0.006	0.111
3rd QT	**0.11**	0.036	**0.002**	0.039	0.181	-0.011	0.034	0.745	-0.077	0.055
4th QT	**-0.085**	0.040	**0.034**	-0.164	-0.006	**-0.338**	0.033	**<0.001**	-0.403	-0.274
Grade 6-7 / Grade 6 (Junior high)										
1st QT	0.062	0.034	0.07	-0.005	0.130	-0.003	0.042	0.948	-0.086	0.080
2nd QT	0.029	0.032	0.369	-0.034	0.091	**0.087**	0.043	**0.043**	0.003	0.170
3rd QT	**0.105**	0.032	**0.001**	0.043	0.168	0.002	0.043	0.955	-0.081	0.086
4th QT	-0.034	0.032	0.296	-0.096	0.029	-0.053	0.052	0.309	-0.155	0.049

Note: Quartiles were calculated based on 5-95 percentile samples of Japanese language for each grade group and cohort. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1-3, 4-5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2-4, 5-6, and 7 in FY2019) and the FY2018 to FY2020 growth of control group C1 (Grade groups 2-4, 5-6, and 7 in FY2018) for each quartile, the impacts of COVID-19 pandemic on the control group C2 are removed. The standard errors in the estimates are based on nonparametric bootstrap with 300 replications. 'Est' means the estimated coefficients of interaction term "COVID19 $\times$ After". Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. 'Est.' and 'p-value' are bolded for $p < 0.05$.

Table 6 Quantile-DID results: Math

	Effects 7 months after SC					Effects 19 months after SC				
				Confidence interval					Confidence interval	
Quantile	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Grade 1-3 (Lower elementary)										
1st QT	**-0.206**	0.031	**<0.001**	-0.266	-0.145	**-0.332**	0.035	**<0.001**	-0.400	-0.264
2nd QT	**-0.155**	0.028	**<0.001**	-0.210	-0.101	**-0.236**	0.030	**<0.001**	-0.295	-0.177
3rd QT	**-0.098**	0.024	**<0.001**	-0.145	-0.050	**-0.131**	0.026	**<0.001**	-0.182	-0.079
4th QT	**0.251**	0.023	**<0.001**	0.206	0.296	**0.286**	0.048	**<0.001**	0.192	0.380
Grade 4-5 (Upper elementary)										
1st QT	**-0.216**	0.035	**<0.001**	-0.285	-0.147	**-0.345**	0.034	**<0.001**	-0.411	-0.279
2nd QT	**-0.181**	0.033	**<0.001**	-0.244	-0.117	**-0.295**	0.030	**<0.001**	-0.353	-0.236
3rd QT	**-0.09**	0.034	**0.009**	-0.158	-0.023	**-0.217**	0.029	**<0.001**	-0.274	-0.159
4th QT	-0.119	0.061	0.052	-0.239	0.001	**-0.268**	0.039	**<0.001**	-0.344	-0.193
Grade 6-7 / Grade 6 (Junior high)										
1st QT	**-0.112**	0.031	**<0.001**	-0.172	-0.052	**-0.364**	0.047	**<0.001**	-0.455	-0.272
2nd QT	-0.02	0.031	0.521	-0.081	0.041	**-0.248**	0.044	**<0.001**	-0.334	-0.162
3rd QT	0.029	0.032	0.358	-0.033	0.091	-0.007	0.049	0.878	-0.103	0.088
4th QT	0.045	0.048	0.348	-0.049	0.139	0.059	0.060	0.323	-0.058	0.177

Note: Quartiles were calculated based on 5-95 percentile samples of math for each grade group and cohort. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1-3, 4-5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2-4, 5-6, and 7 in FY2019) and the FY2018 to FY2020 growth of control group C1 (Grade groups 2-4, 5-6, and 7 in FY2018) for each quartile, the impacts of COVID-19 pandemic on the control group C2 are removed. The standard errors in the estimates are based on nonparametric bootstrap with 300 replications. 'Est' means the estimated coefficients of interaction term "COVID19 × After". Estimated results for other variables are omitted. All p-values below 0.001 are indicated as p < 0.001. 'Est.' and 'p-value' are bolded for p < 0.05.

Table 7 Heterogeneity of treatment effects across groups (students receiving school attendance assistance and others)

	Effects 7 months after school closure					Effects 19 months after school closure				
				Confidence intervals					Confidence intervals	
Grade groups	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Japanese language										
Grade 1-3 (Lower elementary)	0.001	0.065	0.984	-0.127	0.130	0.090	0.057	0.116	-0.022	0.202
Grade 4-5 (Upper elementary)	0.057	0.061	0.354	-0.064	0.178	0.034	0.058	0.557	-0.080	0.148
Grade 6-7 (Junior high)	-0.042	0.066	0.522	-0.172	0.087	-0.083	0.082	0.312	-0.243	0.078
Math										
Grade 1-3 (Lower elementary)	-0.083	0.067	0.216	-0.213	0.048	-0.025	0.061	0.688	-0.145	0.095
Grade 4-5 (Upper elementary)	0.017	0.070	0.809	-0.120	0.154	-0.073	0.058	0.209	-0.186	0.041
Grade 6 (Junior high)	-0.061	0.065	0.350	-0.188	0.067	-0.142	0.082	0.082	-0.302	0.018

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by status of school attendance assistance, grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include individual fixed effects and pre-treatment grade dummies. ‘Est‘ means the coefficients of ‘COVID19 $\times$ After $\times$ SAA‘, which tests the difference in slope of the treatment effect between SAA = 1 and SAA = 0. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. ’Est.’ and ’p-value’ are bolded for $p < 0.05$. ‘S.E.‘ presents cluster-robust standard error results at the classroom level.

Table 8 Heterogeneity of effects across groups (students living with one parent and others)

	Effects 7 months after school closure					Effects 19 months after school closure				
				Confidence intervals					Confidence intervals	
Grade groups	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Japanese language										
Grade 1-3 (Lower elementary)	0.000	0.072	0.997	-0.141	0.141	0.056	0.063	0.375	-0.068	0.180
Grade 4-5 (Upper elementary)	0.116	0.068	0.089	-0.018	0.249	0.084	0.062	0.177	-0.038	0.207
Grade 6-7 (Junior high)	-0.042	0.071	0.553	-0.181	0.097	-0.060	0.088	0.491	-0.232	0.111
Math										
Grade 1-3 (Lower elementary)	-0.042	0.075	0.580	-0.189	0.106	0.023	0.066	0.727	-0.107	0.153
Grade 4-5 (Upper elementary)	0.065	0.080	0.417	-0.093	0.223	-0.052	0.063	0.414	-0.176	0.072
Grade 6 (Junior high)	-0.070	0.066	0.295	-0.200	0.061	-0.143	0.088	0.103	-0.314	0.029

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by whether the student living with one parent, grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include pre-treatment grade dummies. ‘Est‘ means the coefficients of ‘COVID19 × After × one parent‘, which tests the difference in slope of the treatment effect between one parent = 1 and one parent = 0. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. ’Est.’ and ’p-value’ are bolded for $p < 0.05$. ‘S.E.‘ presents cluster-robust standard error results at the classroom level.

Table 9 Heterogeneity of effects across groups (female and male students)

	Effects 7 months after school closure					Effects 19 months after school closure				
				Confidence intervals					Confidence intervals	
Grade groups	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Japanese language										
Grade 1-3 (Lower elementary)	-0.044	0.042	0.296	-0.128	0.039	-0.049	0.036	0.173	-0.120	0.021
Grade 4-5 (Upper elementary)	-0.031	0.044	0.477	-0.118	0.055	0.056	0.041	0.175	-0.025	0.137
Grade 6-7 (Junior high)	**0.117**	0.047	**0.013**	0.024	0.209	**0.134**	0.062	**0.03**	0.013	0.256
Math										
Grade 1-3 (Lower elementary)	0.004	0.045	0.93	-0.085	0.093	0.008	0.039	0.838	-0.068	0.084
Grade 4-5 (Upper elementary)	0.026	0.047	0.575	-0.066	0.119	0.035	0.042	0.412	-0.048	0.118
Grade 6 (Junior high)	0.032	0.047	0.496	-0.061	0.125	0.001	0.059	0.992	-0.116	0.117

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by gender, grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include pre-treatment grade dummies. ‘Est‘ means the coefficients of ‘COVID19 × After × female‘, which tests the difference in slope of the treatment effect between female = 1 and female = 0. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. ’Est.’ and ’p-value’ are bolded for $p < 0.05$. ‘S.E.‘ presents cluster-robust standard error results at the classroom level.

Table 10 DDD results at 7 months after school closure (Students in the schools reduced athletic events over 200 minutes and others)

Grade groups	Est.	S.E.	p-value	Confidence intervals Lower	Upper
Japanese language					
Grade 1-3 (Lower elementary): Reduc20 = 1	0.012	0.115	0.919	-0.214	0.237
Grade 4-5 (Upper elementary): Reduc20 = 1	0.028	0.124	0.824	-0.215	0.270
Grade 6-7 (Junior high): Reduc20 = 1	0.029	0.102	0.777	-0.171	0.229
Math					
Grade 1-3 (Lower elementary): Reduc20 = 1	0.031	0.077	0.689	-0.120	0.181
Grade 4-5 (Upper elementary): Reduc20 = 1	0.044	0.115	0.703	-0.182	0.270
Grade 6-7 (Junior high): Reduc20 = 1	-0.015	0.077	0.841	-0.166	0.135

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by reduced hours of athletic event, grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include pre-treatment grade dummies. 'Reduc20' is dummy variables that take 1 if the student belongs to a school that reduced athletic events by at least 200 minutes in FY2020. 'Est' means the coefficient of 'COVID19 x After x Reduc20', which tests the difference in slope of the treatment effect between Reduc20 = 1 and Reduc20 = 0. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. 'Est.' and 'p-value' are bolded for $p < 0.05$. 'S.E.' presents cluster-robust standard error results at the classroom level.

Table 11 DDD results at 19 months after school closure (Students in the schools reduced athletic events over 200 minutes and others)

	Reduc20 = 1 & Reduc21 = 0					Reduc20 = Reduc21 = 1				
				Confidence intervals					Confidence intervals	
Grade groups	Est.	S.E.	p-value	Lower	Upper	Est.	S.E.	p-value	Lower	Upper
Japanese language										
Grade 1-3 (Lower elementary)	0.062	0.094	0.510	-0.122	0.246	0.151	0.139	0.279	-0.122	0.424
Grade 4-5 (Upper elementary)	0.029	0.079	0.711	-0.126	0.185	-0.017	0.115	0.879	-0.242	0.207
Grade 6 (Junior high)	-0.014	0.093	0.882	-0.197	0.169	0.005	0.123	0.969	-0.237	0.247
Math										
Grade 1-3 (Lower elementary)	0.152	0.078	0.051	-0.001	0.306	0.115	0.105	0.274	-0.091	0.320
Grade 4-5 (Upper elementary)	-0.023	0.074	0.758	-0.167	0.122	-0.003	0.107	0.977	-0.213	0.206
Grade 6 (Junior high)	0.148	0.095	0.120	-0.039	0.334	0.261	0.133	0.051	-0.001	0.522

Note: This table shows the heterogeneity of the effect of the COVID-19 pandemic on standardized test scores by reduced hours of athletic event, grade group and subject. The outcome variable is the standardized Japanese and math test scores. Control variables include pre-treatment grade dummies. ‘Reduc20‘ and ‘Reduc21‘ are dummy variables that take 1 if the student belongs to a school that reduced athletic events by at least 200 minutes in FY2020 and FY2021, respectively. ‘Est‘ in column 2 and 7 mean the coefficient of ‘COVID19 x After x Reduc20‘ and ‘COVID19 x After x Reduc20 + COVID19 x After x Reduc21‘, representing the difference in the effects of COVID-19 pandemic between Reduc20 = 1 & Reduc21 = 0 and Reduc20 = Reduc21 = 0 and between Reduc20 = Reduc21 = 1 and Reduc20 = Reduc21 = 0, respectively. In estimating the impacts at 19 months after the school closure, the effects of COVID-19 pandemic are included between FY2019 and FY2020 for control group C2 (Grade groups 1-3, 4-5, and 6 in FY2018). Therefore, by adding back the differences between the FY2019 to FY2020 growth of treatment group T1 (Grade groups 2-4, 5-6, and 7 in FY2019) and the FY2018 to FY2020 growth of control group C1 (Grade groups 2-4, 5-6, and 7 in FY2018) for each quartile, the impacts of COVID-19 pandemic on the control group C2 are removed. Estimated results for other variables are omitted. All p-values below 0.001 are indicated as $p < 0.001$. ’Est.’ and ’p-value’ are bolded for $p < 0.05$. ‘S.E.‘ presents cluster-robust standard error results at the classroom level.

Table A1 Balance test of outcome and control variables by athletic events min. (Before)

	T1 (2019–20)					T2 (2019–21)				
Variable	Athletic events reduction (min.)	Obs.	Mean	S.D.	p-value	Athletic events reduction (min.)	Obs.	Mean	S.D.	p-value
Test score (J)	< 200 min.	13053	59.11	21.57	<0.001	< 200 min.	15197	59.92	21.90	0.010
	≥ 200 min.	7696	57.50	21.59		≥ 200 min.	2399	58.67	22.09	
Standardized scores (J)	< 200 min.	13053	0.08	1.03	<0.001	< 200 min.	15197	0.13	1.07	0.010
	≥ 200 min.	7696	0.01	1.03		≥ 200 min.	2399	0.07	1.08	
Test score (M)	< 200 min.	13053	64.51	19.31	<0.001	< 200 min.	15197	64.56	19.79	0.001
	≥ 200 min.	7696	62.26	19.71		≥ 200 min.	2399	63.17	19.81	
Standardized scores (M)	< 200 min.	13053	0.11	0.97	<0.001	< 200 min.	15197	0.09	1.00	0.001
	≥ 200 min.	7696	0.00	0.99		≥ 200 min.	2399	0.02	1.00	
School attendance assistance (SAA)	< 200 min.	13053	0.16	0.37	<0.001	< 200 min.	15197	0.17	0.37	0.096
	≥ 200 min.	7696	0.18	0.39		≥ 200 min.	2399	0.16	0.36	
Living with one parent	< 200 min.	13052	0.14	0.35	0.043	< 200 min.	15195	0.14	0.35	0.210
	≥ 200 min.	7695	0.16	0.36		≥ 200 min.	2399	0.15	0.36	
Female student	< 200 min.	13053	0.50	0.50	0.980	< 200 min.	15197	0.50	0.50	0.094
	≥ 200 min.	7696	0.50	0.50		≥ 200 min.	2399	0.48	0.50	