Online Appendix

Impact of COVID-19 pandemic on the cognitive and non-cognitive skills of elementary school students

Shinsuke ASAKAWA
Saga University and The University of Osaka

Fumio OHTAKE
The University of Osaka

Sections 3 and 4 demonstrated that students residing in disadvantaged learning conditions (DLCs) during and after the COVID-19 school closures exhibited slower recovery in cognitive and non-cognitive skills. However, living in a DLC is unlikely to be the sole determinant of learning recovery. To more fully understand the heterogeneity in the effects of school closures, we further investigate two additional factors that vary across schools: class reorganization and the use of cram schools or private tutors.

Class reorganization may influence students' educational outcomes by altering peer group dynamics. Prior research has shown that peer effects can significantly shape student performance (Burke & Sass, 2013; Sacerdote, 2011). In Nara City, class compositions are typically restructured each April, yet the COVID-19 closures from March to May coincided with this period. Thus, students in schools with only a single class per grade (i.e., without class reorganization) may have experienced different impacts from the closures compared to those in multi-class schools. To assess this, we re-estimate Equations (1)–(6) using data from single-class schools only. As shown in Tables A7–A10 and Figures A6–A12, the estimated effects for these students are largely consistent with the main findings for the full sample (Figures 2–8). This suggests that class reorganization is unlikely to be a major confounding factor.

In addition to classroom structure, the use of cram schools and private tutors may play a compensatory role in learning recovery. Nara Prefecture reports a higher rate of supplementary education usage among elementary students than other regions in Japan.[1] Prior studies, such as Liu (2012) in the Taiwanese context, have found that cram school attendance can improve student academic performance, particularly in subjects like mathematics. Given these findings, we explore whether higher rates of supplementary education use buffered the negative effects of school closures.

To examine this, we calculate the average proportion of students using cram schools or tutors at each school using data from the 2021 National Survey of Academic Achievement. Based on this metric, schools are divided into three groups by usage rate, and Equations (1)–(6) are estimated separately for each subgroup. The results, presented in Figures A13–A16, show that schools with higher rates of cram school/tutor use experienced faster recovery in both cognitive and non-cognitive domains. However, the magnitude of the difference between the high-use group (above 67.74%) and the low-use group (below 51.81%) is smaller than the gap observed between students from DLC and non-DLC environments. Therefore, we conclude that while access to supplementary education may aid recovery, its influence appears weaker than that of living in a DLC.

However, these findings should be interpreted with caution due to several limitations. First, the measure of cram school and tutor attendance is based on school-level aggregates, which may introduce measurement error and obscure individual-level heterogeneity. Second, the lack of longitudinal data prevents us from examining dynamic changes in supplementary education participation over time. Specifically, the 2021 National Survey

[1] According to the results of national survey of school performance and learning (MEXT, 2021), the percentage of students who do not use cram schools or tutors was 42.8% in Nara Prefecture, compared to the national average of 52.6% in Japan. The data can be downloaded from the following URL (https://www.nier.go.jp/21chousakekkahoukoku/factsheet/prefecture-City.html).

of Academic Achievement provides the only available cross-sectional information on sixth-grade students' involvement in such activities. Future research would benefit from panel data that allow for the identification of causal effects of changes in supplementary education on the evolution of cognitive and non-cognitive skill formation in the aftermath of school disruptions.

Appendix references

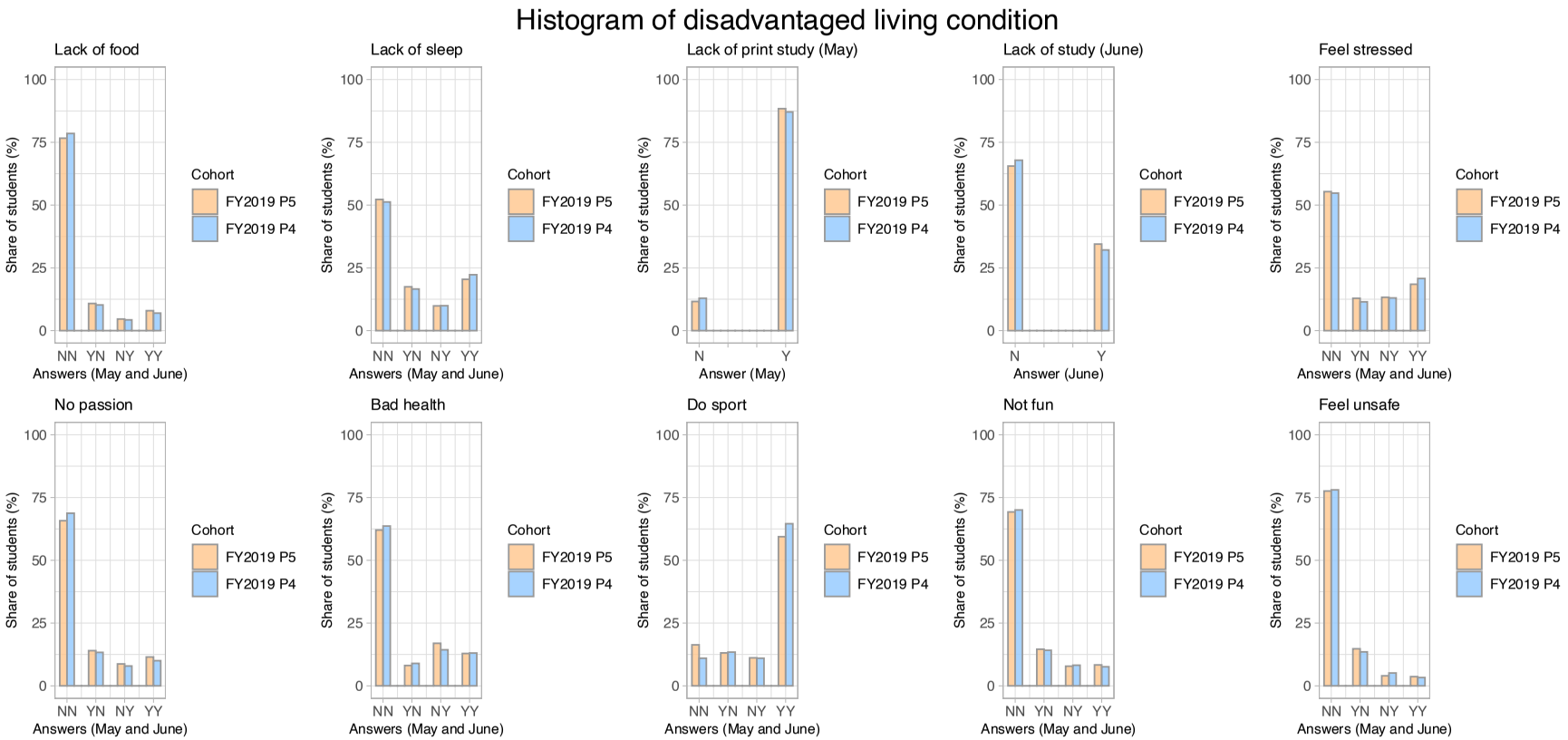
Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics, 31*(1), 51-82. https://doi.org/10.1086/666653

Liu, J. (2012). Does cram schooling matter? Who goes to cram schools? Evidence from Taiwan. *International Journal of Educational Development 32*(1), 46-52. https://doi.org/10.1016/j.ijedudev.2011.01.014

Ministry of Education, Culture, Sports, Science, and Technology (MEXT). (2021). *Results of the 2021 National Survey of Academic Achievement and Learning by prefecture*. https://www.nier.go.jp/21chousakekkahoukoku/factsheet/prefecture-City.html. Retrieved July 24, 2022.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education* (Vol. 3, pp. 249-277). Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00004-1

List of Appendix Figures



Note: 'N' means 'No / Partly No' and 'Y' means 'Yes / Partly Yes' for each disadvantaged living condition. We show changes over the two periods for variables that both the May and June surveys included. (e.g., 'YN' means 'Yes / Partly Yes' in May and 'No / Partly No' in June). 'Lack print study (May)' and 'Lack study (June)' show the percentages as of May and June, respectively.

Fig. A1 Histogram of the disadvantaged living conditions

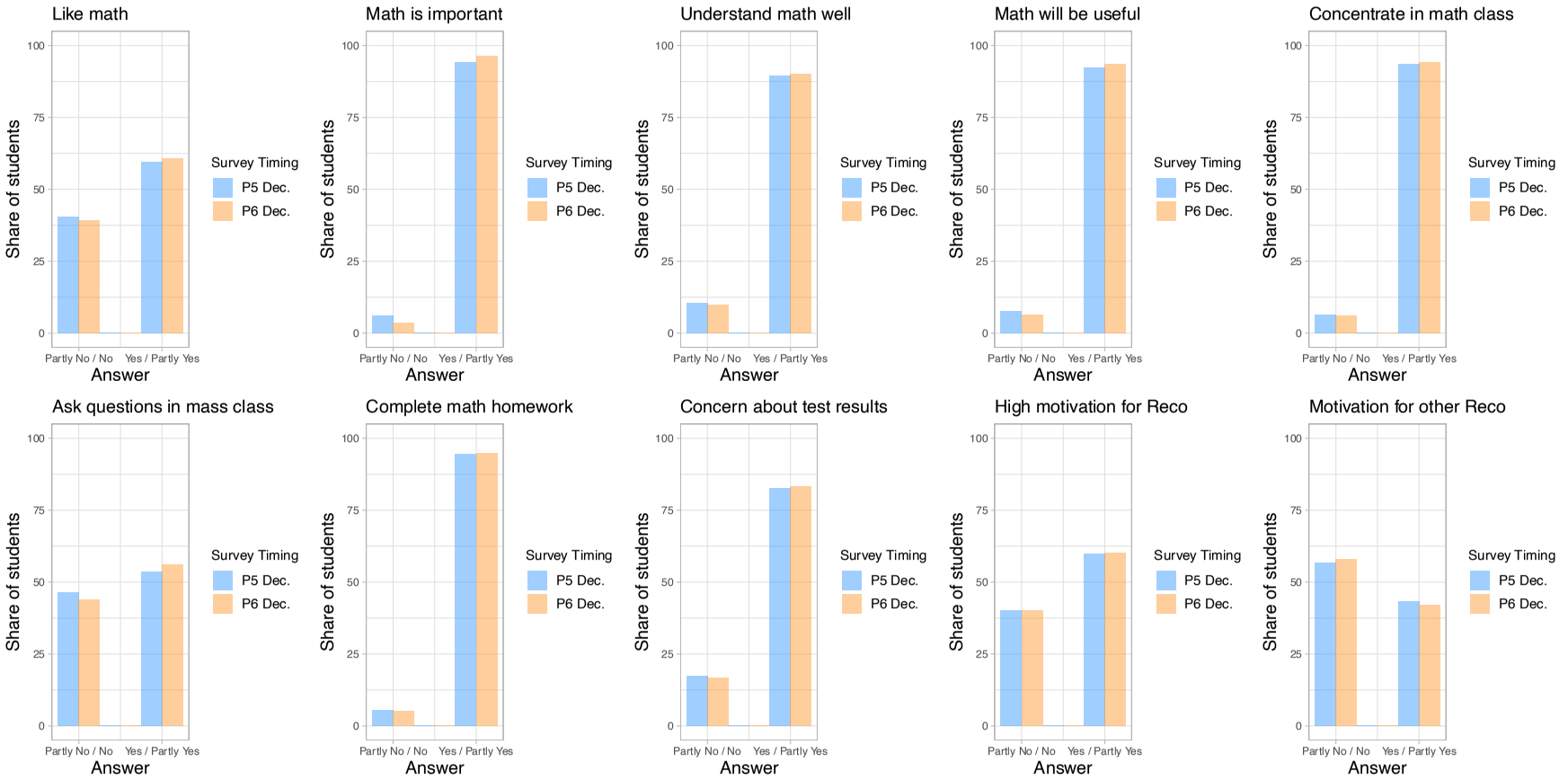


Fig. A2 Histogram of the attitudes toward proactive learning in math (FY2019 P6 Cohort)

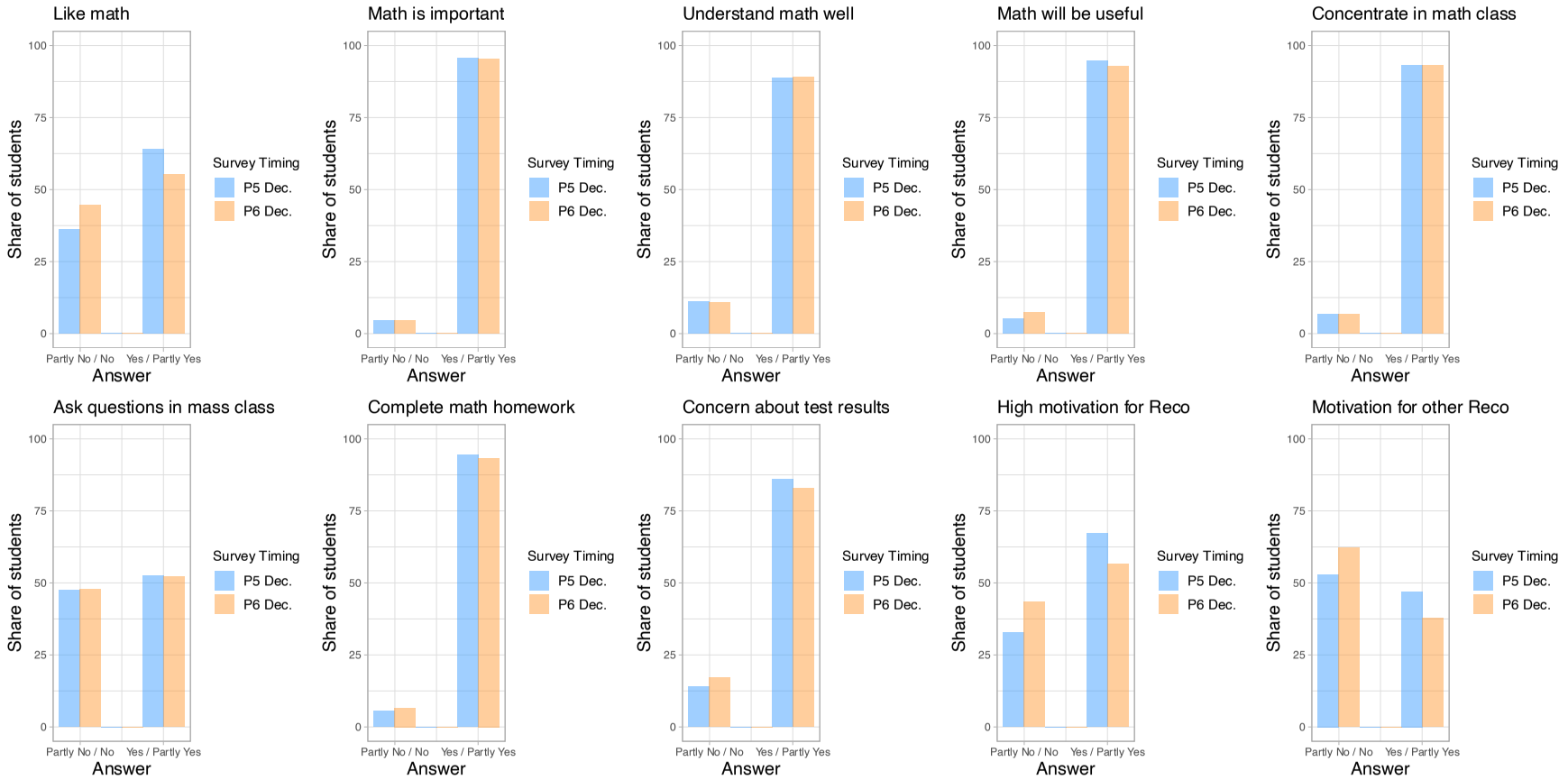
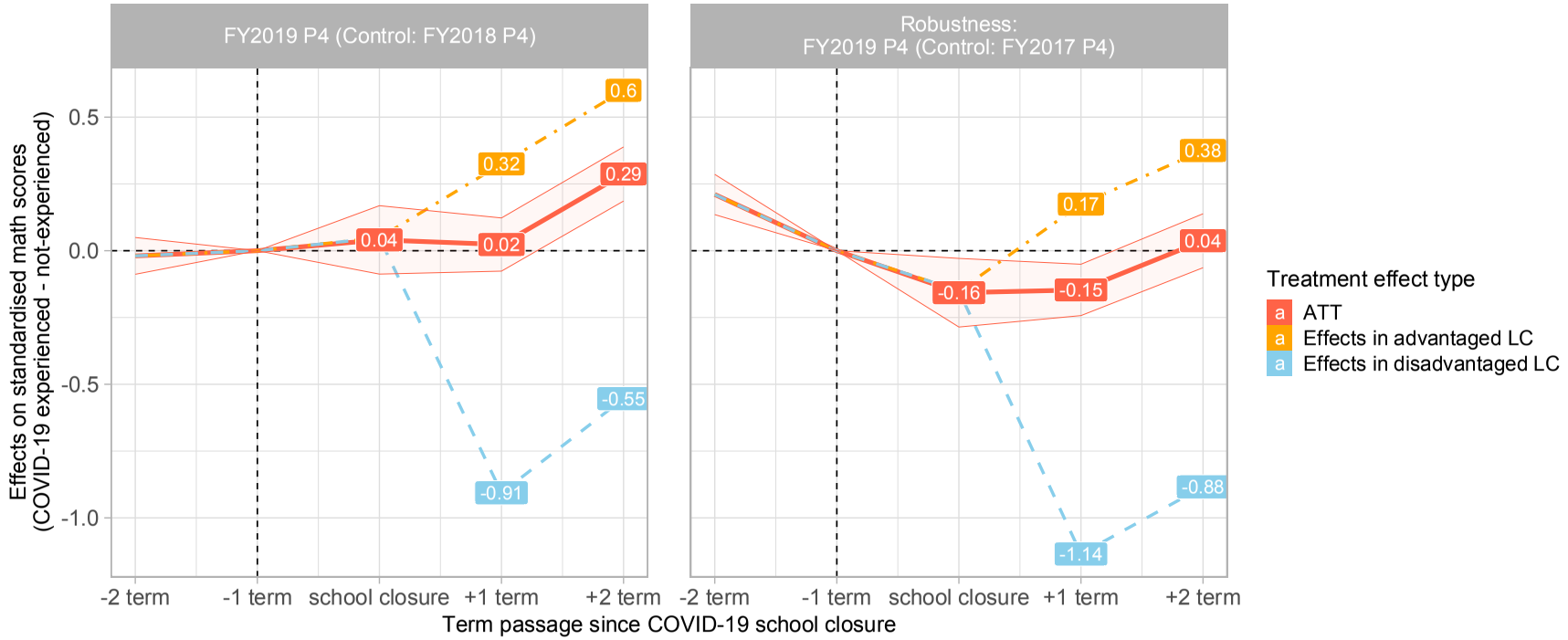
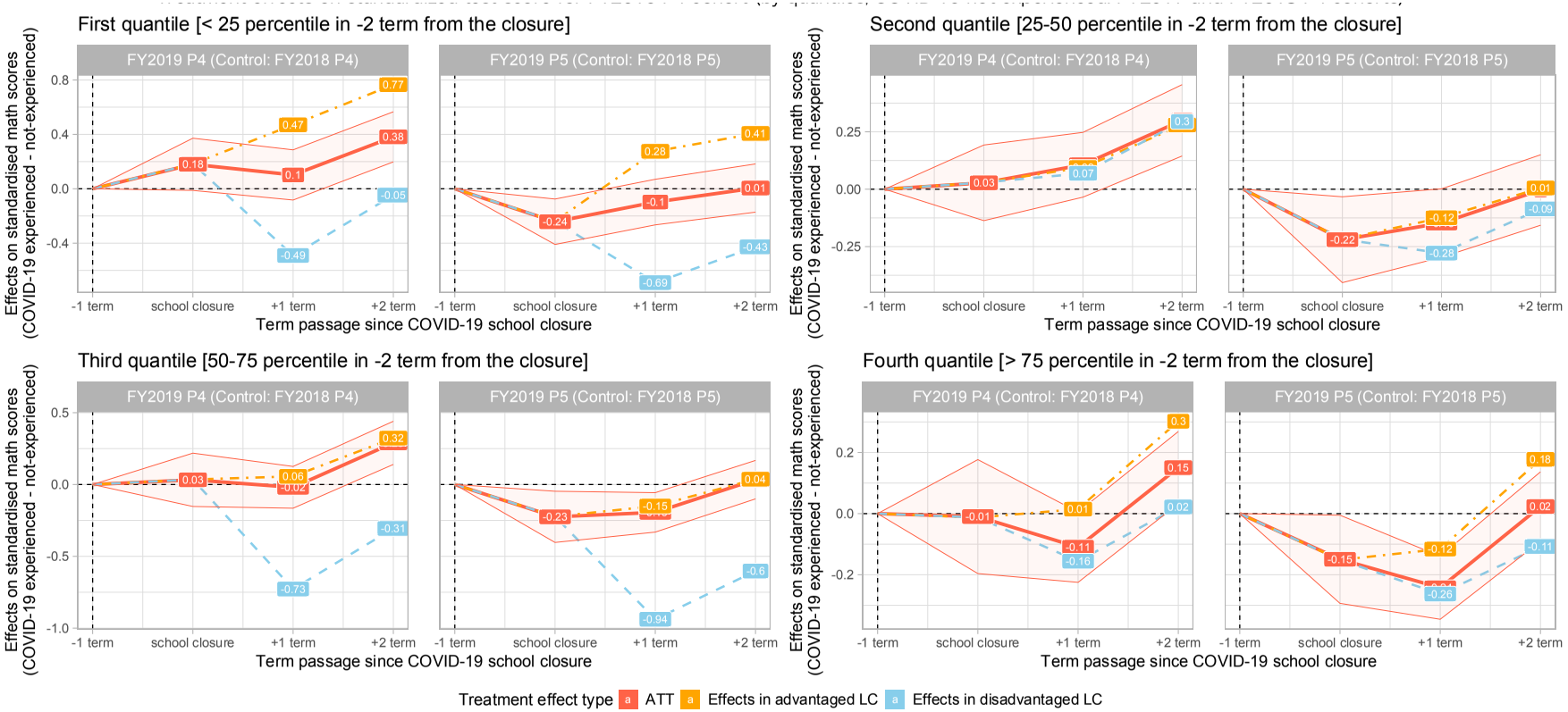


Fig. A3 Histogram of the attitudes toward proactive learning in math (FY2019 P5 Cohort)



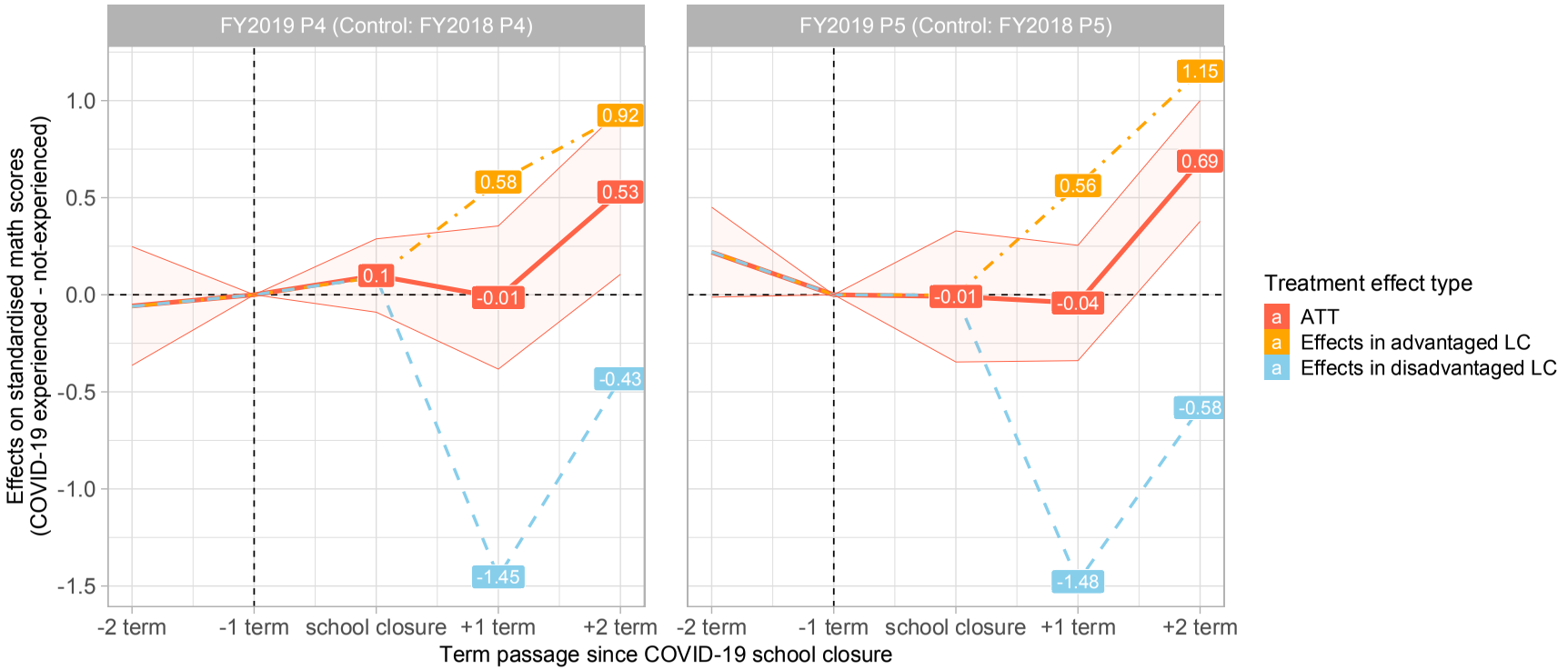
Note: For each cohort, the solid (red), dotdashed (orange), dashed (blue) lines indicate the ATT and the coefficient of the effect of COVID-19 for the students with the most ALCs or DLCs, respectively. The effects for the students with the most ALCs or DLCs are estimated by equations (2.1) and (2.2), respectively. Only for the ATT, the 95% confidence intervals are indicated by solid lines and filled ranges. For all identification strategies, the reference time period is t = -1.

Fig. A4 Treatment effects on standardized test score for FY2019 P4 cohort (COVID-19 not-experienced: FY2017 and FY2018 P4 cohorts)



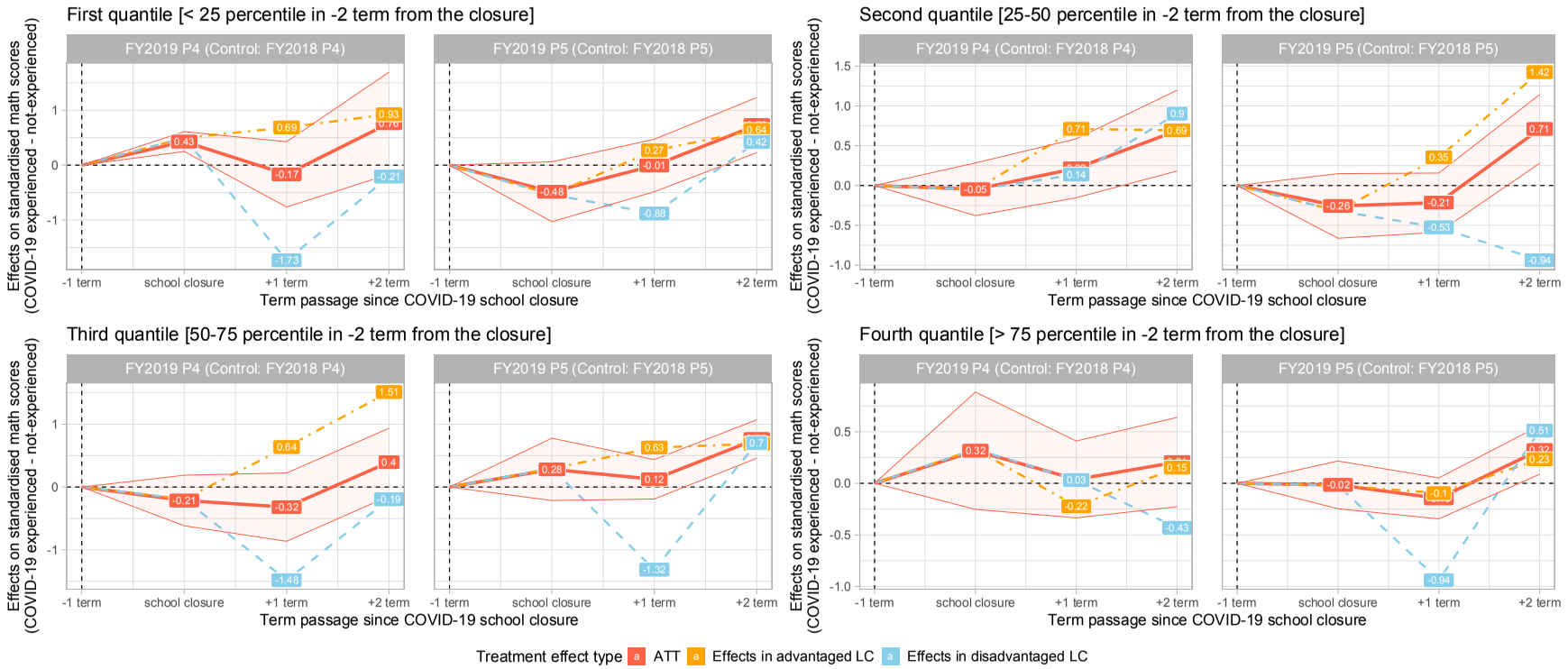
Note: The quantile is calculated based on the score of -2 term. For each cohort, the solid (red), dotdashed (orange), dashed (blue) lines indicate the ATT and the coefficient of the effect of COVID-19 for the students with the most ALCs or DLCs, respectively. The effects for the students with the most ALCs or DLCs are estimated by equations (2.1) and (2.2), respectively. Only for the ATT, the 95% confidence intervals are further indicated by solid lines and filled ranges. For all identification strategies, the reference time period is t = -1.

Fig. A5 Treatment effects on standardized test score for FY2019 P4 cohort (by quantile, COVID-19 not-experienced: FY2017 and FY2018 P4 cohorts)



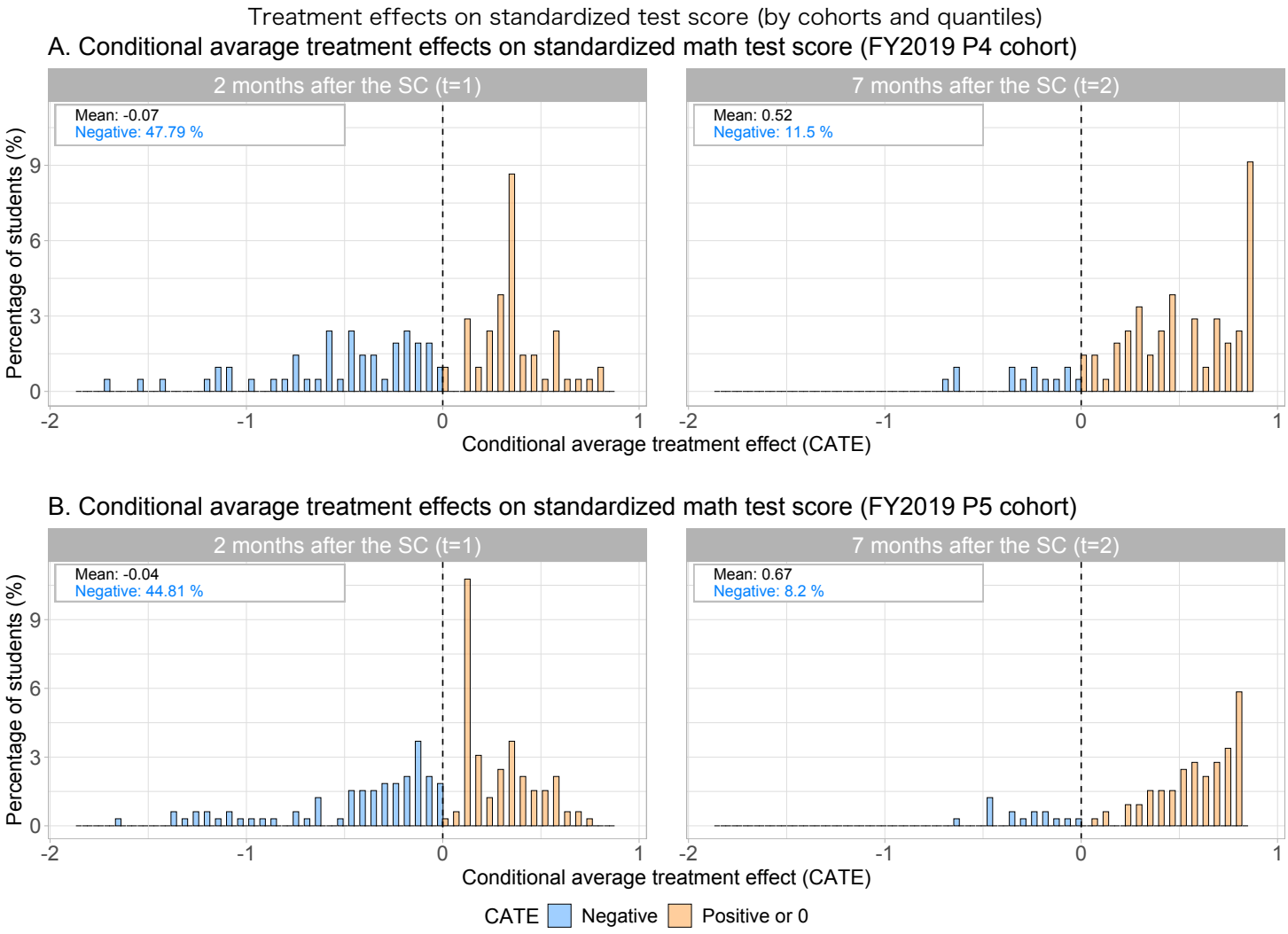
Note: We use only single class schools and grades for the sample. For each cohort, the solid (red), dotdashed (orange), dashed (blue) lines indicate the ATT and the coefficient of the effect of COVID-19 for the students with the most ALCs or DLCs, respectively. The effects for the students with the most ALCs or DLCs are estimated by equations (2.1) and (2.2), respectively.
Only for the ATT, the 95% confidence intervals are further indicated by solid lines and filled ranges. For all identification strategies, the reference time period is t = -1.

Fig. A6 Treatment effects on standardized math test score (DID, Full sample of schools without class changes)



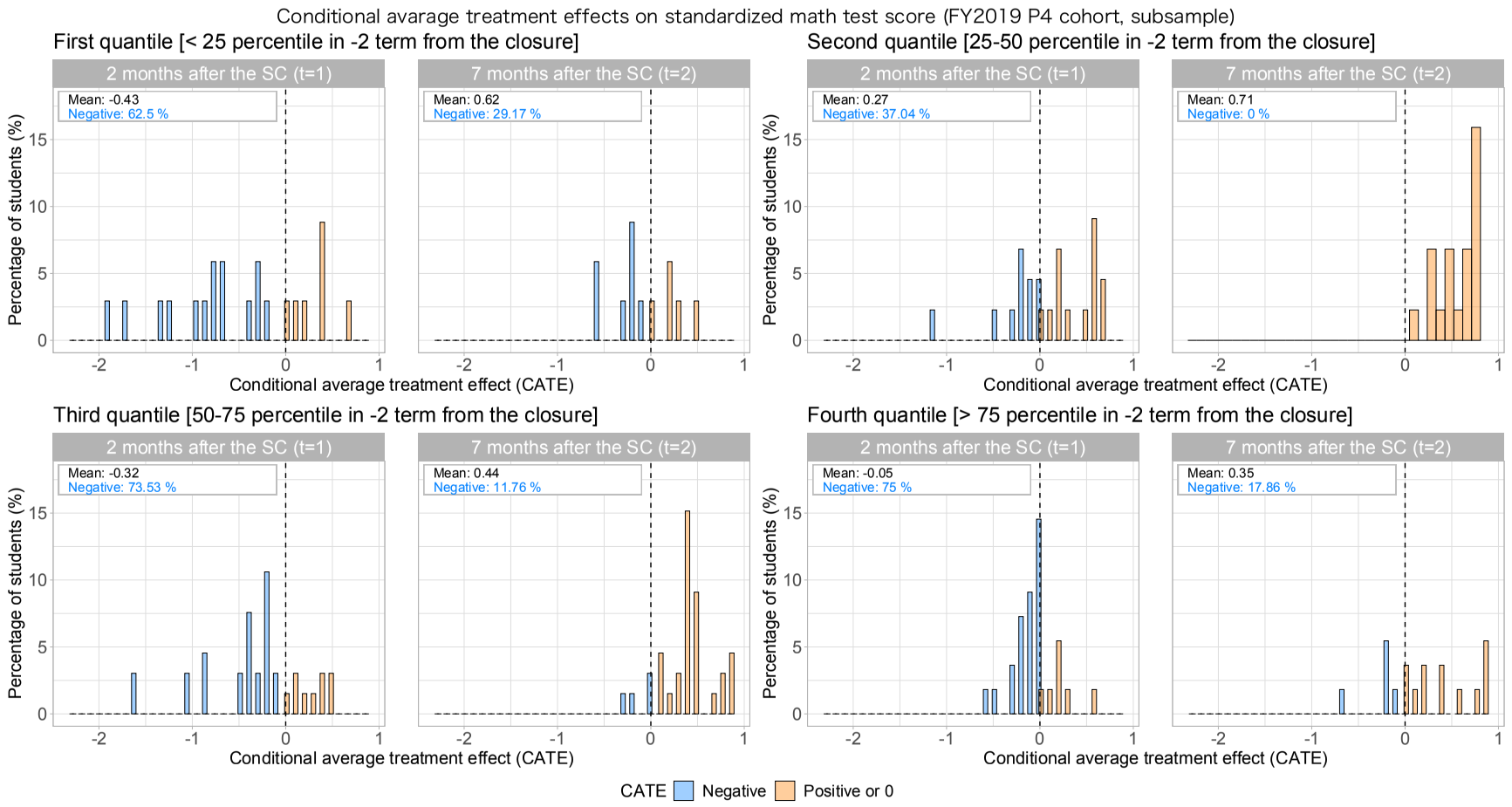
Note: Use only single class schools and grades for the sample. The quantile is calculated based on the score of -2 term. For each cohort, the solid (red), dotdashed (orange), dashed (blue) lines indicate the ATT and the coefficient of the effect of COVID-19 for the students with the most ALCs or DLCs, respectively. The effects for the students with the most ALCs or DLCs are estimated by equations (2.1) and (2.2), respectively. Only for the ATT, the 95% confidence intervals are further indicated by solid lines and filled ranges. For all identification strategies, the reference time period is t = -1.

Fig. A7 ATE and CATE on standardized math test score (Full sample of schools without class changes)



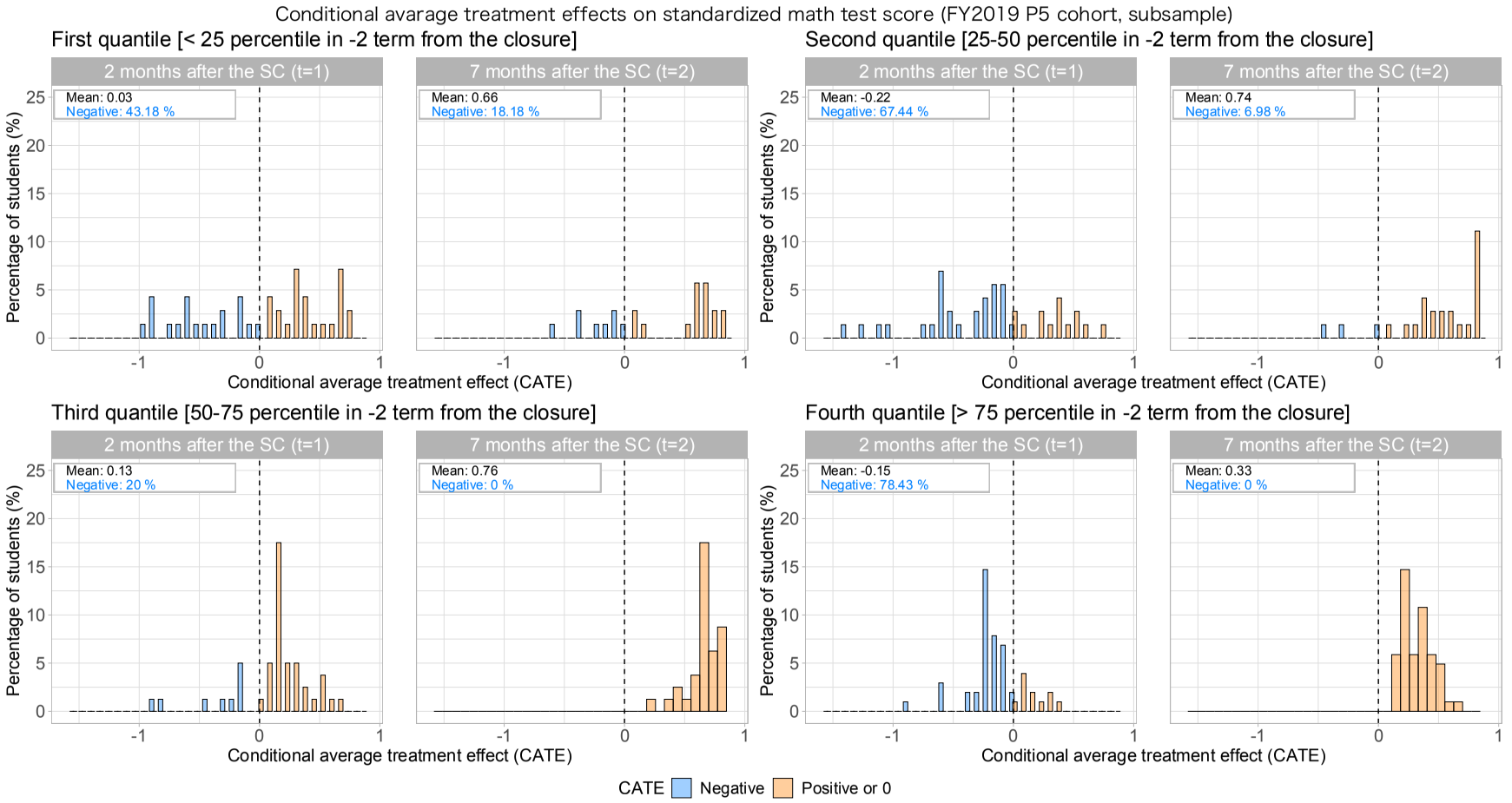
Note: Use only single class schools and grades for the sample. `Percentage of students` is defined as the number of students per bin divided by the total number of students taking the test in the treatment cohort for each term. `Mean` and `Negative` refer to the mean of the CATE and the percentage of students with a negative CATE. The number of bins in the histogram is set to 35.

Appendix Figure A8 Treatment effects on standardized math test score (DID, Subsample of schools without class changes)



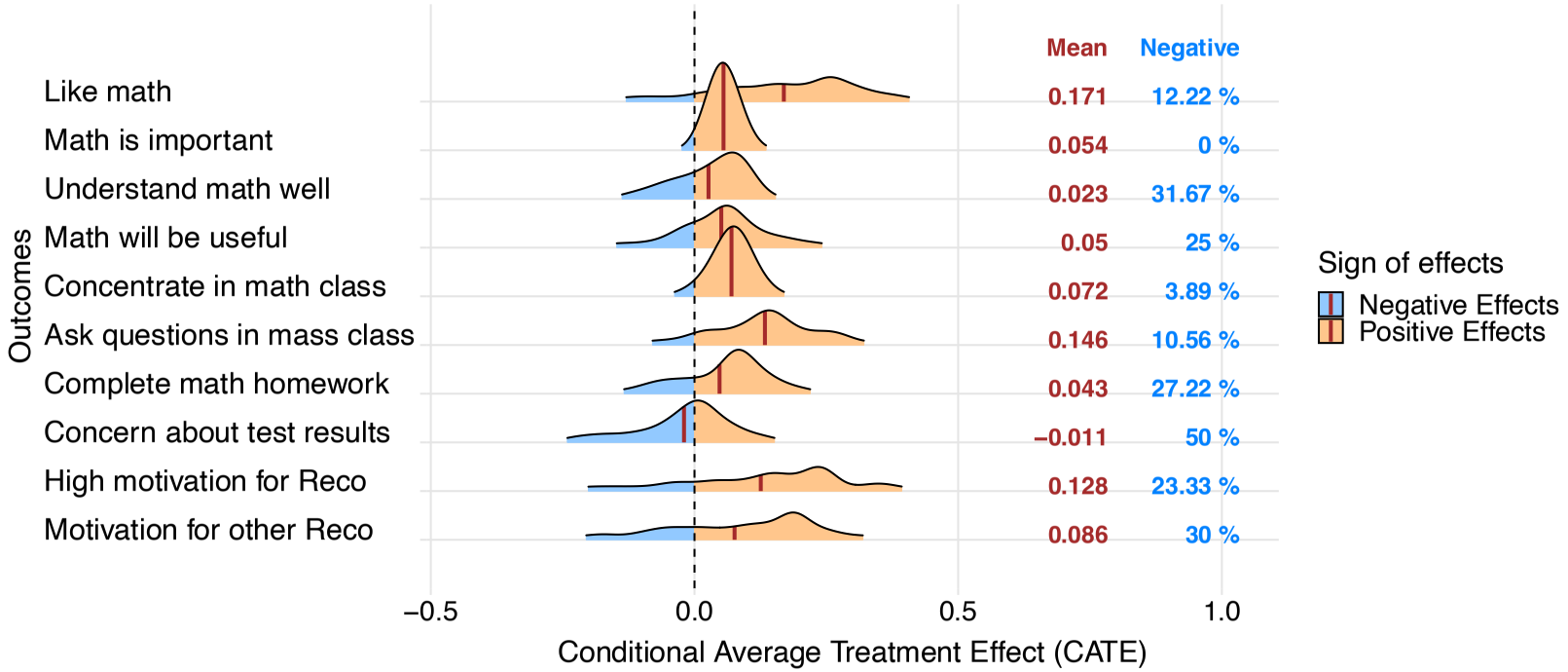
Note: Use only single class schools and grades for the sample. The quantile is calculated based on the score of -2 term. `Percentage of students` is defined as the number of students per bin divided by the total number of students taking the test in the treatment cohort for each term. `Mean` and `Negative` refer to the mean of the CATE and the percentage of students with a negative CATE. The number of bins in the histogram is set to 35.

Fig. A9 Conditional average treatment effects on standardized math test score (FY2019 P4 cohort, Subsample of schools without class changes)



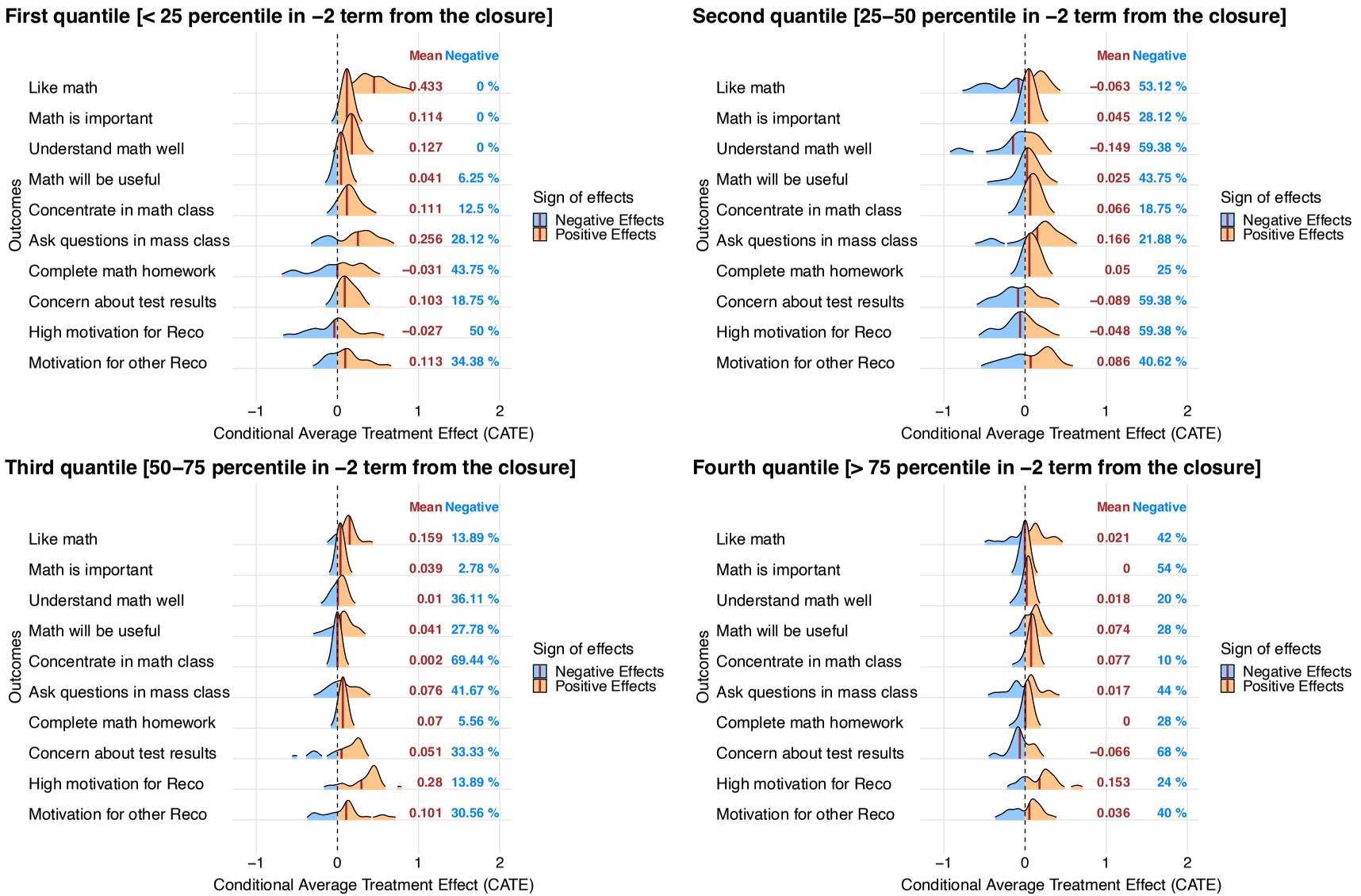
Note: Use only single class schools and grades for the sample. The quantile is calculated based on the score of -2 term. `Percentage of students` is defined as the number of students per bin divided by the total number of students taking the test in the treatment cohort for each term. `Mean` and `Negative` refer to the mean of the CATE and the percentage of students with a negative CATE. The number of bins in the histogram is set to 35.

Fig. A10 Conditional average treatment effects on standardized math test score (FY2019 P5 cohort, Subsample of schools without class changes)



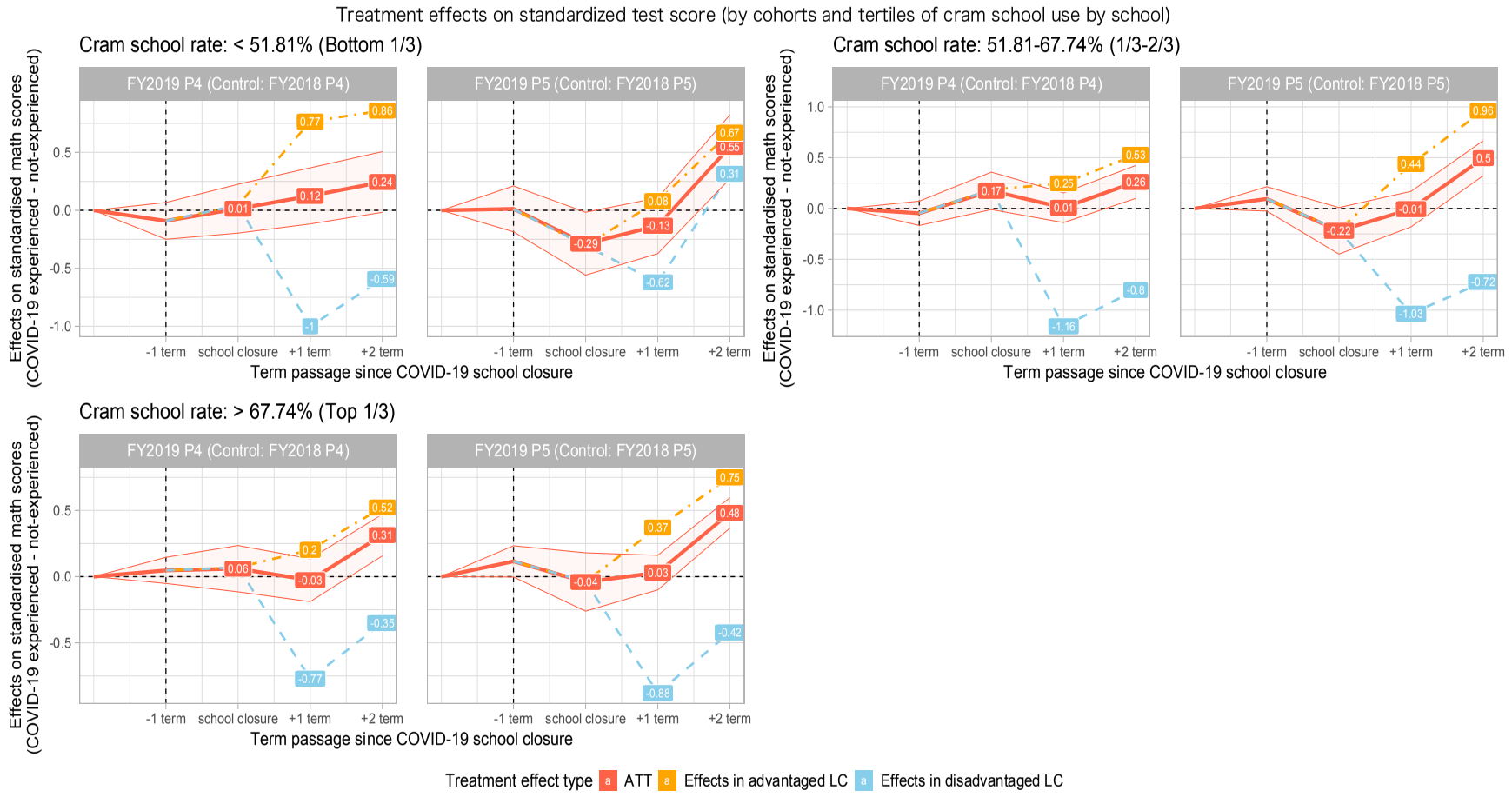
Note: To compare the effect sizes, we standardized all outcome variables separately (one outcome on math comprehension and nine on math non–cognitive skills) 'Mean' and 'Negative' refer to the mean and the percentage of students with a negative CATE.

Fig. A11 Treatment effects on non-cognitive skills (Full sample of schools without class changes)



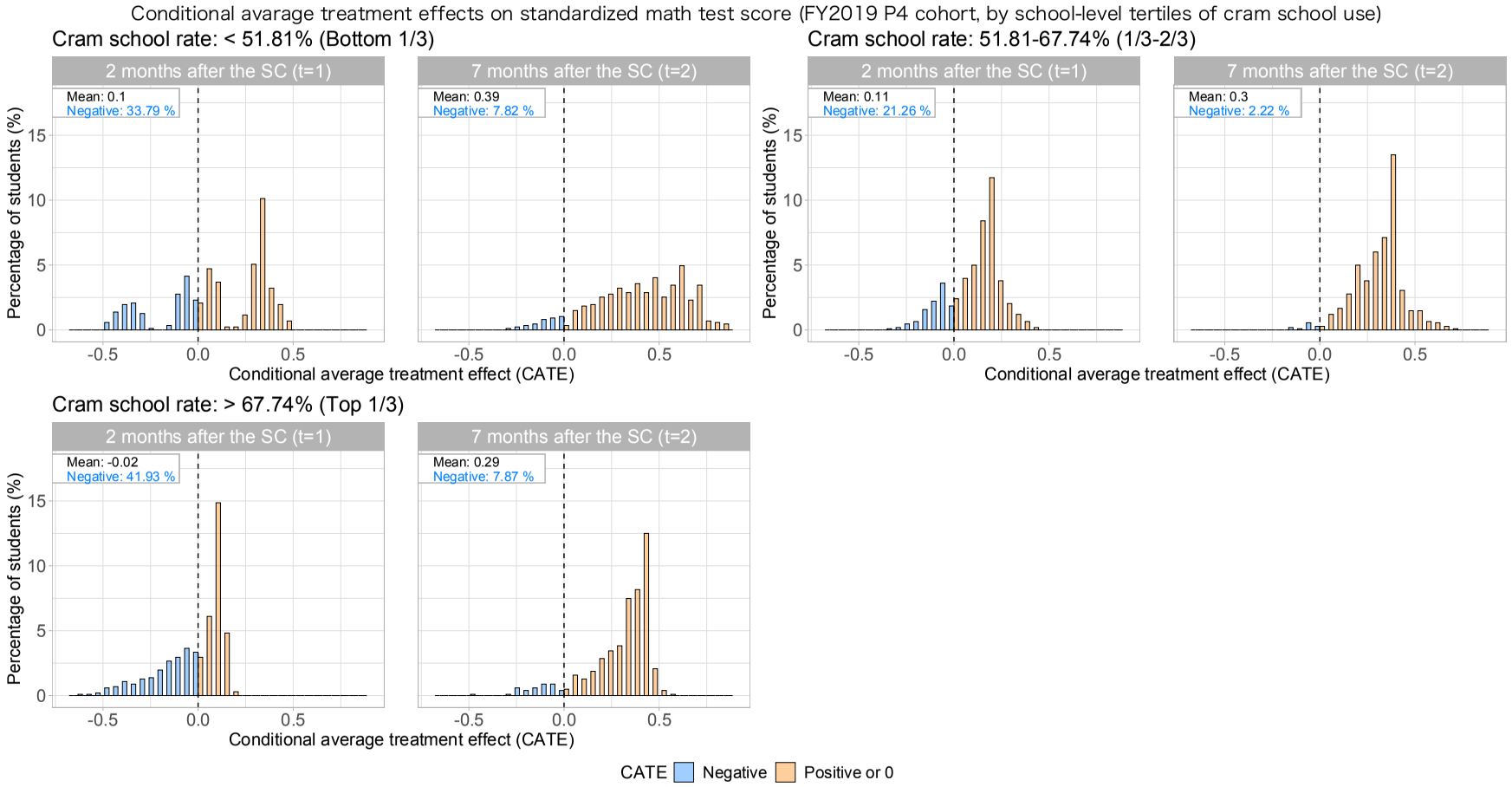
Note: The quantile is calculated based on the standardized math test score of −2 term. We standardized all outcome variables separately (one outcome on math comprehension and nine on math non−cognitive skills). 'Mean' and 'Negative' refer to the mean and the percentage of students with a negative CATE.

Fig. A12 Treatment effects on non-cognitive skills (Subsample of schools without class changes)



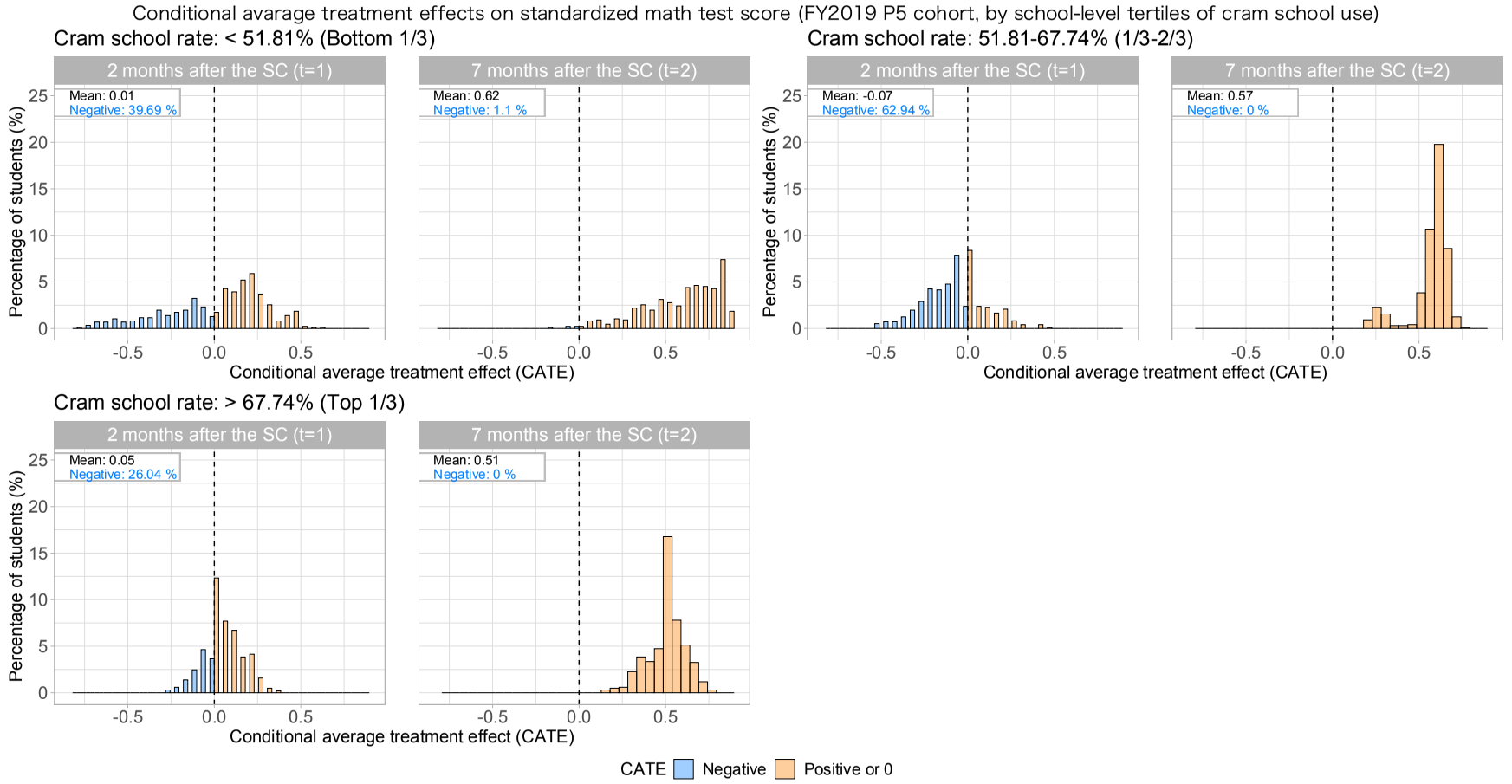
Note: The tertile is calculated based on a school-by-school basis based on school averages of using a cram school/a tutor from the 2021 National Survey of Academic Achievement. For each cohort, the solid (red), dotdashed (orange), dashed (blue) lines indicate the ATT and the coefficient of the effect of COVID-19 for the students with the most ALCs or DLCs, respectively. The effects for the students with the most ALCs or DLCs are estimated by equations (2.1) and (2.2), respectively. Only for the ATT, the 95% confidence intervals are further indicated by solid lines and filled ranges.For all identification strategies, the reference time period is t = -1.

Fig. A13 Treatment effects on standardized math test score (DID, Subsample by the average rate of attending a cram school /using a tutor)



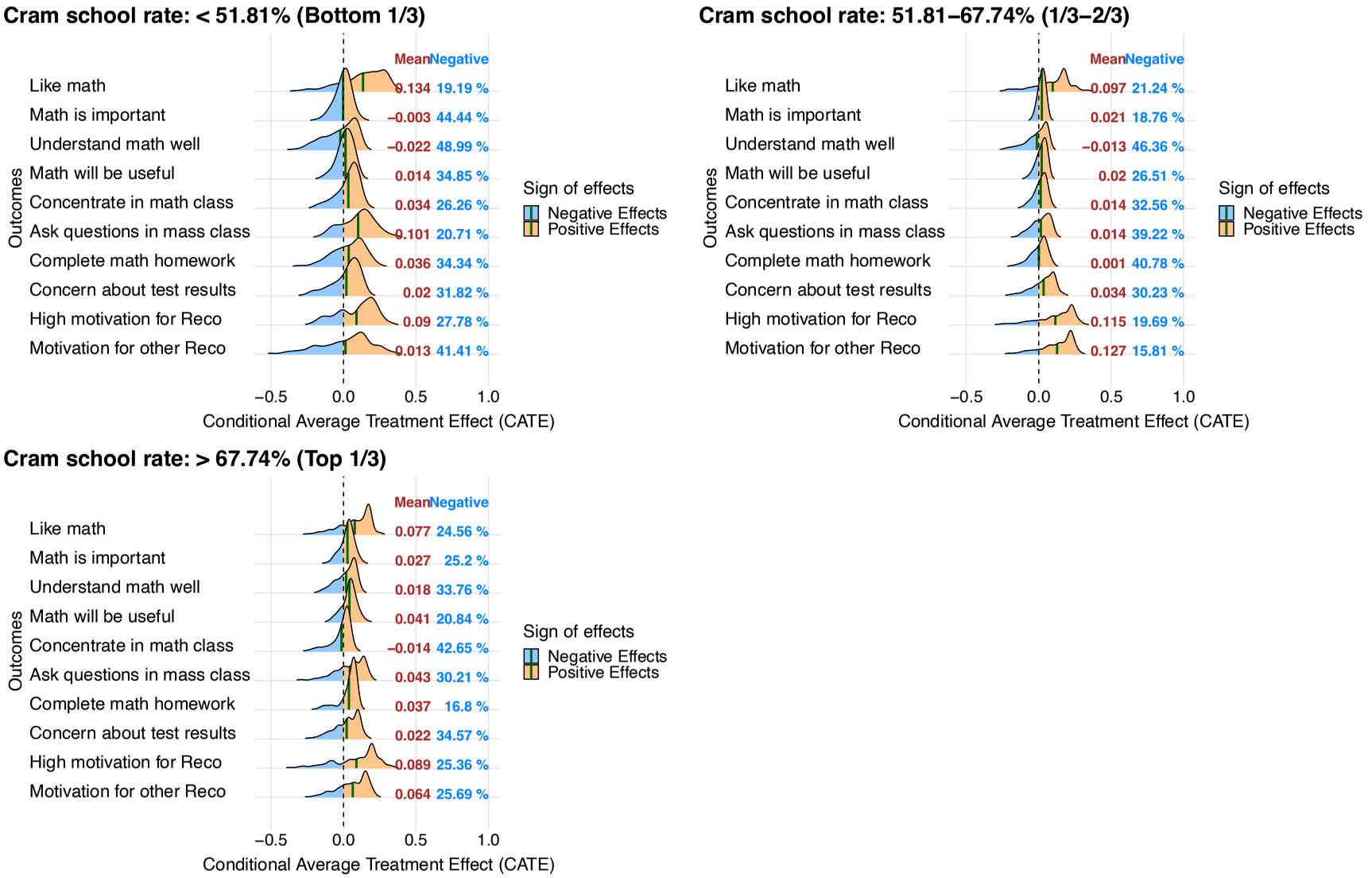
Note: The tertile is calculated based on a school-by-school basis based on school averages of using a cram school/a tutor from the 2021 National Survey of Academic Achievement. `Percentage of students` is defined as the number of students per bin divided by the total number of students taking the test in the treatment cohort for each term. `Mean` and `Negative` refer to the mean of the CATE and the percentage of students with a negative CATE. The number of bins in the histogram is set to 35.

Fig. A14 CATE on standardized math test score (FY2019 P4 cohort, Subsample by the average rate of attending a cram school /using a tutor)



Note: The tertile is calculated based on a school-by-school basis based on school averages of using a cram school/a tutor from the 2021 National Survey of Academic Achievement. `Percentage of students` is defined as the number of students per bin divided by the total number of students taking the test in the treatment cohort for each term. `Mean` and `Negative` refer to the mean of the CATE and the percentage of students with a negative CATE. The number of bins in the histogram is set to 35.

Fig. A15 CATE on standardized math test score (FY2019 P5 cohort, Subsample by the average rate of attending a cram school /using a tutor)



Note: The quantile is calculated based on a school–by–school basis based on school averages of using a cram school/a tutor from the 2021 National Survey of Academic Achievement. We standardized all outcome variables separately (one outcome on math comprehension and nine on math 'Mean' and 'Negative' refer to the mean and the percentage of students with a negative CATE.

Fig. A16 Treatment effects on non-cognitive skills (Subsample by the average rate of attending a cram school /using a tutor)

List of Tables

Table A1: Balance test of school-level math scores between schools tested in FY2019 term 3 and those in other terms

Comparison:	FY2019 term 3 vs. Other terms							
Cohorts:	FY2019 P4				FY2019 P5			
Tests:	T-test		F-test		T-test		F-test	
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Test score	3.77	0.104	0.720	0.654	−8.77	0.0776*	1.55	0.352

Note: *p<0.1; **p<0.05; ***p<0.01

Table A2: Definitions for disadvantaged living condition dummy

Living Condition	Definition
Lack food	= 1 if the student answered, "not applicable" or "not really applicable" to the question "I eat breakfast and lunch every day," or 0 otherwise
Lack sleep	= 1 if the student answered, "applicable" or "mostly applicable" to the question "I sometimes have difficulty sleeping," or 0 otherwise
Lack print study (May)	= 1 if the student answered, "not applicable" or "not really applicable" to the question "I studied using handouts from school during the temporary primary school closure," or 0 otherwise
Lack study (June)	= 1 If the student answered, "applicable" or "mostly applicable" to the question "I sometimes have difficulty concentrating on studies," or 0 otherwise
Feel stressed	= 1 if the student answered, "applicable" or "mostly applicable" to the question "I get upset, frustrated, or angry," or 0 otherwise
No passion	= 1 if the student answered, "applicable" or "mostly applicable" to the question "I have no motivation to do anything," or 0 otherwise
Bad health	= 1 if the student answered, "applicable" or "mostly applicable" to the question "I sometimes have physical problems such as a headache or stomachache," or 0 otherwise
No sport	= 1 if the student answered, "not applicable" or "not really applicable" to the question "I exercise a lot," or 0 otherwise
Not fun	= 1 if the student answered, "not applicable" or "not really applicable" to the question "I enjoy every day," or 0 otherwise
Feel unsafe	= 1 if the student answered, "applicable" or "mostly applicable" to the question "I have felt anxious about something," or 0 otherwise

Table A3: Robustness check of full-sample DID (Treatment: FY2019 P4 cohort, Control: FY2017 and FY2018 P4 cohort)

	Dependent variables: Standardized math test score					
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort) [Main estimation]			Treatment: FY2019 P4 cohort (Control: FY2017 P4 cohort) [Robustness check]		
	ATT	(1)+DLC [Most ALC]	(1)+ALC [Most DLC]	ATT	(4)+DLC [Most ALC]	(4)+ALC [Most DLC]
	(1)	(2)	(3)	(4)	(5)	(6)
COVID19 x 8 months before SC ($t = -2$)	-0.019	-0.019	-0.019	0.210***	0.210***	0.210***
	(0.035)	(0.035)	(0.035)	(0.038)	(0.038)	(0.038)
COVID19 x SC beginning term ($t = 0$)	0.041	0.049	0.049	-0.157*	-0.151*	-0.151*
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
COVID19 x 2 months after SC ($t = 1$)	0.023	0.325***	-0.906***	-0.147**	0.174	-1.14***
	(0.051)	(0.087)	(0.143)	(0.049)	(0.091)	(0.145)
COVID19 x 7 months after SC ($t = 2$)	0.287***	0.600***	-0.554***	0.037	0.375***	-0.884***
	(0.051)	(0.088)	(0.158)	(0.051)	(0.096)	(0.158)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	19,668	19,668	19,668	21,485	21,485	21,485

Note: For all estimations, reference period is $t = -1$. All test scores are standardized per treatment and control group pair. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A4: Robustness check of subsample DID (Treatment: FY2019 P4 cohort, Control: FY2017 and FY2018 P4 cohort): ATT

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort) [Main estimation]				Treatment: FY2019 P4 cohort (Control: FY2017 P4 cohort) [Robustness check]			
	1st QT	2nd QT	3rd QT	4th QT	1st QT	2nd QT	3rd QT	4th QT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
COVID19 x SC beginning term ($t = 0$)	0.180	0.028	0.033	-0.010	-0.242**	-0.220*	-0.225*	-0.149*
	(0.097)	(0.084)	(0.094)	(0.095)	(0.085)	(0.095)	(0.091)	(0.073)
COVID19 x 2 months after SC ($t = 1$)	0.102	0.106	-0.020	-0.110	-0.098	-0.147	-0.194**	-0.243***
	(0.093)	(0.071)	(0.074)	(0.058)	(0.085)	(0.076)	(0.070)	(0.052)
COVID19 x 7 months after SC ($t = 2$)	0.382***	0.300***	0.290***	0.150*	0.005	-0.003	0.034	0.023
	(0.093)	(0.079)	(0.076)	(0.060)	(0.090)	(0.078)	(0.068)	(0.058)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,536	3,947	3,948	3,516	3,955	4,369	4,222	3,935

Note: 'Test score quantile (QT)' is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A5: Robustness check of subsample DID: Treatment effects on students with the most advantaged living conditions (FY2019 P4 cohort vs. FY2017–2018)

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort) [Main estimation]				Treatment: FY2019 P4 cohort (Control: FY2017 P4 cohort) [Robustness check]			
	1st QT (1)	2nd QT (2)	3rd QT (3)	4th QT (4)	1st QT (5)	2nd QT (6)	3rd QT (7)	4th QT (8)
COVID19 x SC beginning term ($t = 0$)	0.183	0.029	0.035	-0.013	-0.239**	-0.218*	-0.223*	-0.151*
	(0.097)	(0.084)	(0.094)	(0.095)	(0.084)	(0.095)	(0.090)	(0.073)
COVID19 x 2 months after SC ($t = 1$)	0.470**	0.095	0.057	0.015	0.275	-0.124	-0.150	-0.116
	(0.174)	(0.119)	(0.123)	(0.089)	(0.168)	(0.125)	(0.125)	(0.086)
COVID19 x 7 months after SC ($t = 2$)	0.766***	0.280*	0.323*	0.302***	0.407*	0.005	0.037	0.178*
	(0.174)	(0.123)	(0.132)	(0.086)	(0.165)	(0.127)	(0.137)	(0.085)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,536	3,947	3,948	3,516	3,955	4,369	4,222	3,935

Note: ‘Test score quantile (QT)‘ is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. ‘COVID19‘ is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. ‘DLC‘ and ‘ALC‘ are disadvantaged and advantaged living conditions, respectively. ‘SC‘ is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A6: Robustness check of subsample DID: Treatment effects on students with the most disadvantaged living conditions (FY2019 P4 cohort vs. FY2017–2018)

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort) [Main estimation]				Treatment: FY2019 P4 cohort (Control: FY2017 P4 cohort) [Robustness check]			
	1st QT (1)	2nd QT (2)	3rd QT (3)	4th QT (4)	1st QT (5)	2nd QT (6)	3rd QT (7)	4th QT (8)
COVID19 x SC beginning term ($t = 0$)	0.183	0.029	0.035	-0.013	-0.239**	-0.218*	-0.223*	-0.151*
	(0.097)	(0.084)	(0.094)	(0.095)	(0.084)	(0.095)	(0.090)	(0.073)
COVID19 x 2 months after SC ($t = 1$)	-0.490	0.069	-0.728**	-0.156	-0.690*	-0.278	-0.938***	-0.263
	(0.271)	(0.187)	(0.220)	(0.194)	(0.272)	(0.187)	(0.222)	(0.188)
COVID19 x 7 months after SC ($t = 2$)	-0.046	0.295	-0.306	0.022	-0.427	-0.086	-0.603**	-0.107
	(0.268)	(0.209)	(0.230)	(0.192)	(0.295)	(0.218)	(0.224)	(0.177)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,536	3,947	3,948	3,516	3,955	4,369	4,222	3,935

Note: ‘Test score quantile (QT)‘ is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. ‘COVID19‘ is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. ‘DLC‘ and ‘ALC‘ are disadvantaged and advantaged living conditions, respectively. ‘SC‘ is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A7: DID results for single-class schools: Average treatment effects on the treated and subgroup effects for students with the most advantaged and disadvantaged living conditions

	Dependent variables: Standardized math test score					
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort)			Treatment: FY2019 P5 cohort (Control: FY2018 P5 cohort)		
	ATT	(1)+DLC [Most ALC]	(1)+ALC [Most DLC]	ATT	(4)+DLC [Most ALC]	(4)+ALC [Most DLC]
	(1)	(2)	(3)	(4)	(5)	(6)
COVID19 x 8 months before SC ($t = -2$)	-0.058	-0.060	-0.060	0.220	0.220	0.220
	(0.149)	(0.150)	(0.150)	(0.115)	(0.116)	(0.116)
COVID19 x SC beginning term ($t = 0$)	0.099	0.090	0.090	-0.009	-0.002	-0.002
	(0.092)	(0.099)	(0.099)	(0.168)	(0.166)	(0.166)
COVID19 x 2 months after SC ($t = 1$)	-0.014	0.581*	-1.45***	-0.042	0.562*	-1.48***
	(0.180)	(0.279)	(0.337)	(0.148)	(0.253)	(0.399)
COVID19 x 7 months after SC ($t = 2$)	0.529*	0.925***	-0.433	0.689***	1.15***	-0.581
	(0.207)	(0.240)	(0.429)	(0.154)	(0.219)	(0.363)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,642	1,642	1,642	1,949	1,949	1,949

Note: Sample is restricted to schools and grades with only one class for two years from $t = -2$ to $t = 2$. For all estimations, reference period is $t = -1$. All test scores are standardized per treatment and control group pair. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A8: Subsample DID results for single-class schools: Average treatment effects on the treated

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort)				Treatment: FY2019 P5 cohort (Control: FY2018 P5 cohort)			
	1st QT (1)	2nd QT (2)	3rd QT (3)	4th QT (4)	1st QT (5)	2nd QT (6)	3rd QT (7)	4th QT (8)
COVID19 x SC beginning term ($t = 0$)	0.429***	-0.047	-0.214	0.316	-0.485	-0.257	0.281	-0.015
	(0.087)	(0.161)	(0.196)	(0.276)	(0.270)	(0.202)	(0.246)	(0.115)
COVID19 x 2 months after SC ($t = 1$)	-0.168	0.216	-0.320	0.038	-0.006	-0.215	0.123	-0.146
	(0.288)	(0.180)	(0.265)	(0.181)	(0.236)	(0.185)	(0.156)	(0.099)
COVID19 x 7 months after SC ($t = 2$)	0.763	0.690**	0.395	0.206	0.730**	0.711**	0.764***	0.323**
	(0.450)	(0.247)	(0.263)	(0.210)	(0.249)	(0.215)	(0.152)	(0.117)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	317	315	324	290	350	358	372	391

Note: Sample is restricted to schools and grades with only one class for two years from $t = -2$ to $t = 2$. 'Test score quantile (QT)' is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A9: Subsample DID results for single-class schools: Subgroup effects for students with the most advantaged living conditions

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort)				Treatment: FY2019 P5 cohort (Control: FY2018 P5 cohort)			
	1st QT (1)	2nd QT (2)	3rd QT (3)	4th QT (4)	1st QT (5)	2nd QT (6)	3rd QT (7)	4th QT (8)
COVID19 x SC beginning term ($t = 0$)	0.496***	-0.053	-0.192	0.331	-0.536	-0.321	0.293	-0.020
	(0.096)	(0.171)	(0.201)	(0.281)	(0.304)	(0.189)	(0.259)	(0.113)
COVID19 x 2 months after SC ($t = 1$)	0.687	0.712*	0.636	-0.223	0.269	0.355	0.628*	-0.096
	(0.430)	(0.281)	(0.564)	(0.296)	(0.437)	(0.242)	(0.247)	(0.171)
COVID19 x 7 months after SC ($t = 2$)	0.926	0.693*	1.51**	0.152	0.642	1.42***	0.689**	0.233
	(0.609)	(0.300)	(0.479)	(0.271)	(0.386)	(0.396)	(0.202)	(0.160)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	317	315	324	290	350	358	372	391

Note: Sample is restricted to schools and grades with only one class for two years from $t = -2$ to $t = 2$. 'Test score quantile (QT)' is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A10: Subsample DID results for single-class schools: Subgroup effects for students with the most disadvantaged living conditions

	Dependent variables: Standardized math test score							
	Treatment: FY2019 P4 cohort (Control: FY2018 P4 cohort)				Treatment: FY2019 P5 cohort (Control: FY2018 P5 cohort)			
	1st QT (1)	2nd QT (2)	3rd QT (3)	4th QT (4)	1st QT (5)	2nd QT (6)	3rd QT (7)	4th QT (8)
COVID19 x SC beginning term ($t = 0$)	0.496***	-0.053	-0.192	0.331	-0.536	-0.321	0.293	-0.020
	(0.096)	(0.171)	(0.201)	(0.281)	(0.304)	(0.189)	(0.259)	(0.113)
COVID19 x 2 months after SC ($t = 1$)	-1.73	0.138	-1.48	0.032	-0.880	-0.525	-1.32*	-0.939***
	(0.862)	(0.849)	(1.19)	(0.388)	(0.555)	(0.289)	(0.628)	(0.265)
COVID19 x 7 months after SC ($t = 2$)	-0.208	0.903	-0.195	-0.432	0.421	-0.939*	0.700**	0.512
	(1.03)	(0.679)	(0.503)	(0.416)	(0.820)	(0.374)	(0.245)	(0.259)
Time dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	317	315	324	290	350	358	372	391

Note: Sample is restricted to schools and grades with only one class for two years from $t = -2$ to $t = 2$. 'Test score quantile (QT)' is calculated by the test score in 8 months before the COVID-19 school closure ($t = -2$). For all estimations, reference period is $t = -1$. All test scores are standardized. Standard errors in parentheses are clustered at the classroom level. All coefficients for living conditions and FEs are omitted for simplicity. 'COVID19' is a dummy variable taking one if the student belongs to the FY2019 P4 cohort or FY2019 P5 cohort. 'DLC' and 'ALC' are disadvantaged and advantaged living conditions, respectively. 'SC' is the COVID-19 school closure. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A11: Definition of outcome variables for attitude toward proactive learning of math

Outcome variables	Definition
Like math (Like math)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you like to study math?," or 0 otherwise
Math important (Math is important)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you think it is important to study math?," or 0 otherwise
Understand math well (Understand math)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you understand the content of the math class well?," or 0 otherwise
Math will be useful (Math useful)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you think that what you learned in math class will be useful in the future when you start working?," or 0 otherwise
Concentrate in math Classroom FE (Math concentration)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you listen carefully to the teacher in math class?," or 0 otherwise
Ask questions in math Classroom FE (Math question)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you ask questions to your teacher in math class if you don't understand something?," or 0 otherwise
Complete math homework (Math homework)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you complete your math homework regularly?," or 0 otherwise
Concern about test results (Math results)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you care about the results of the test?," or 0 otherwise
High motivation for Reco (Reco motivation)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "Do you have a high motivation to study using the Reco sheets (reflection study sheets)?," or 0 otherwise
Motivation for other Reco (Other Reco)	= 1 if the student answered, "applicable" or "mostly applicable" to the question "You receive 3 Reco-sheets each time. Do you want to try to work on the other numbered Reco-sheets besides your own?," or 0 otherwise

Note: Words in parentheses are abbreviations for the outcome variables used in the manuscript.