Learning inequality during the COVID-19 pandemic

Per Engzell*[a,b,c], Arun Frey[d], and Mark Verhagen[a,b]

[a]Leverhulme Centre for Demographic Science, 42 Park End St, Oxford OX1 1JD, UK

[b]Nuffield College, University of Oxford, 1 New Rd, Oxford OX1 1NF, UK

[c]Swedish Institute for Social Research, Stockholm University, 106 91 Stockholm, Sweden

[d]Department of Sociology, University of Oxford, 42 Park End St, Oxford OX1 1JD, UK

October 2020

Abstract

Suspension of face-to-face instruction in schools during the COVID-19 pandemic has led to concerns about consequences for student learning. So far, data to study this question have been limited. In this paper, we evaluate the effect of school closures on primary school students in the Netherlands, where schools remained closed for 8 weeks. We examine nationally standardized test scores before and after lockdown, and compare progress during this period to the same period in previous years. Our results reveal a significant loss of learning progress of about 3 percentile points or 0.08 standard deviations. This loss is 35–40% larger among students from low-educated homes, confirming worries about the uneven toll of the pandemic on children and families. Our results are on the same order of magnitude as best-case projections by the European Commission and the World Bank, and suggest losses many times larger in countries less prepared for remote learning.

*To whom correspondence may be addressed: per.engzell@nuffield.ox.ac.uk

1 Introduction

The COVID-19 pandemic is transforming society in profound ways, often exacerbating social and economic inequalities in its wake. In an effort to curb its spread, governments around the world have moved to suspend face-to-face teaching in schools, affecting some 95% of the world's student population—the largest disruption to education in history (*1*). The UN Convention on the Rights of the Child states that governments should provide primary education for all on the basis of equal opportunity (*2*). To weigh the costs of school closures against public health benefits (*3*, *4*), it is crucial to know whether students are learning less in lockdown, and whether disadvantaged students do so disproportionately.

Whereas previous research examined the impact of summer recess on learning, or disruptions from events such as extreme weather or teacher strikes (*5*, *6*), COVID-19 presents a unique challenge that makes it unclear how to apply past lessons. Concurrent effects on the economy make parents less equipped to provide support, as they struggle with economic uncertainty or demands of working from home (*1*). The health and mortality risks of the pandemic incur further psychological costs, as does the toll of social isolation (*7*, *8*). Family violence is projected to rise, putting already vulnerable students at increased risk (*1*). At the same time, the scope of the pandemic may compel governments and schools to respond more forcefully than during other disruptive events.

Data on learning loss during lockdown have been limited. Unlike societal sectors like the economy or the healthcare system, school systems do usually not post data at high-frequency intervals. Schools and teachers have been struggling to adopt online-based solutions for instruction, let alone for assessment and accountability (*5*, *6*). Early data from online learning platforms suggest a drop in coursework completed (*9*) and an increased dispersion of test scores (*10*). More recently, evidence has emerged from classrooms among students returning to school (*11*). Existing studies on learning in lockdown suffer from limitations, however: they typically do not contain high-stakes tests, raising doubts about external validity. Moreover, they lack individual covariates that are needed to adjust for potential biases, e.g., from time trends or selective attrition.

2 This study and its context

Here we present evidence on the pandemic's effect on student progress in the Netherlands, using the natural experiment that occurred as national exams took place before and after school closures.

The Netherlands has several features that makes it attractive as a testing ground for learning loss during the pandemic. While close to the OECD average in terms of school spending and educational performance (*12*), the country leads the world in broadband penetration (*13*, *14*). National and local governments also took swift action to ensure that students had access to appropriate technology (*15*). School closures were short in comparative perspective, and the first wave of the pandemic was relatively mild (*16*, *17*). The Netherlands therefore presents a best-case scenario, providing a lower bound on learning loss elsewhere in Europe and the world. Despite these favorable conditions, survey evidence from lockdown indicates socioeconomic differences in students' access to learning resources and help with schoolwork (*18*).

Our main interest is whether learning stalled during lockdown, and whether students from less-educated homes were disproportionately affected. In addition, we examine differences by sex, school grade, subject, and prior performance. Hypotheses and analysis protocols for this study were pre-registered (*19*). We obtained anonymized data through partnership with a digital platform that supplies teachers and principals with tools to track student performance. The sample, which covers 15% of all Dutch primary schools (n≈350,000), is broadly representative of the national student body (fig. S2). We assess standardized tests in Maths & Arithmetics, Spelling, and Reading Comprehension for students aged 7–11 (grades 4–7). Results are transformed into a percentile score using a common norm for all years, which allows them to be interpreted on an absolute scale.

Key to our study design is the fact that national assessments take place twice a year in the Netherlands (*20*): halfway into the school year in January–February and at the end of the school year in May–June. In 2020, these testing dates occurred just before and after the nationwide school closures that lasted 8 weeks starting March 16 (Fig. 1). Access to

data from 3 years prior to the pandemic allows us to create a natural benchmark against which to assess learning loss. We do so using a difference-in-differences design (*21*), and address loss to follow-up using various techniques: regression adjustment, re-balancing on the propensity score and maximum entropy weights, and comparison within schools and families.

3 Results

Fig. 2 shows our difference-in-differences estimate of learning loss in 2020 compared to the three previous years, using a composite score of students' performance in Maths & Arithmetics, Spelling, and Reading Comprehension. Students lost on average 3 percentile points in the national distribution, equivalent to 0.08 standard deviations (*SD*). Assuming a regular learning progress of 0.30–0.40 *SD* per year (*22*), this implies that 20–27% of total yearly progress was lost. Losses are not distributed equally but concentrated among students from less-educated homes. Those in the two lowest categories of parental education—together accounting for 8% of the population—suffered losses 35–40% larger than the average student. In contrast, we find little evidence that the effect differs by sex, school grade, subject, or prior performance.

We examine the robustness of these results in several ways. As Fig. 1 shows, the end-of-year tests in 2020 were delayed relative to earlier years and a smaller proportion of students were tested. In our analysis in Fig 2, we include terms for the time elapsed between testing dates and a linear trend in year. To confirm that this specification delivers unbiased estimates, we perform a placebo analysis assigning treatment status to each of the three comparison years (fig. S4). In each case, the 95% confidence interval of our main effect spans zero. We also re-estimate our main specification dropping comparison years one at a time (fig. S5). These results align with those of our main analysis.

Fig. 2 adjusts for secular time trends and differences in test timing, but not the fact that only a subset of students returned after lockdown (see Fig. 1). Our design discards with those students who did not do so, which might lead to bias if their performance differs from

those we observe. To confirm that this is not the case, we take several measures. First, we control for individual covariates (table S5). We also restrict analysis to schools where at least 75% of students returned (table S7). Moreover, we balance treatment and control groups on a wider set of covariates using the estimated propensity of treatment (*23*) and maximum entropy weights (*24*) (fig. S6–S7). Finally, we adjust for unobserved heterogeneity at the school and family level using within-school (table S7–S8) and within-family comparison (fig. S8). Results remain robust across these analyses.

One challenge in interpreting our results is whether they actually reflect the cumulative impact of knowledge learned, or more transient "day of exam" effects. Social distancing measures may alter factors such as seating arrangements or indoor climate that in turn can influence student performance (*25*, *26*). Following school re-openings, tests were taken in-person under normal conditions and with minimal social distancing. Still, students may have been under stress or simply unaccustomed to the classroom after several weeks at home. Similarly, if remote teaching covered the requisite material but put less emphasis on test-taking skills, results may have declined while knowledge remained stable. We assess this by inspecting how students perform on tasks not designed to test curricular content (fig. S9). Doing so, our treatment effect shrinks by 62%, implying that knowledge learned is the main channel.

4 Discussion

During the pandemic-induced lockdown in 2020, schools in many countries were forced to close for extended periods. It is of great policy interest to know whether students are able to have their educational needs met in these circumstances, and identify groups at special risk. In this study, we offered micro-level evidence of learning loss during the pandemic using anonymized test scores from Dutch primary school students. There is clear evidence that students are learning less during lockdown than in a typical year. These losses are evident in all grades 4 through 7 and across three subject areas: Maths, Spelling, and Reading. The size of these effects is on the order of 3 percentile points or 0.08 *SD*,

but students from disadvantaged homes are disproportionately affected. In the most low-resourced households, the size of the learning slide is 35–40% larger than in the general population.

Are these losses large or small? To answer this, we turn to projections made early in the pandemic (*6*, *14*, *27*–*30*). The most credible estimates are from internationally recognized bodies like the World Bank (*27*) or the Joint Research Centre of the European Commission (*14*). Helpfully, these projections span a range of scenarios that let us position the Netherlands as a "best case" based on its resilience in the first wave of the pandemic. The World Bank's "optimistic" scenario—schools closed for 3 months and remote learning operating at 60% efficiency—projects a 0.06 *SD* loss in standardized test scores (*27*). This is less than our 0.08 *SD*, despite the fact that Dutch schools only stayed closed for 8 weeks. The European Commission estimates a lower bound of learning loss of 0.008 *SD* per week (*14*). Multiplied by 8 weeks this translates to 0.064 *SD*, on the same order of magnitude as our findings but again marginally smaller.

Taken together, our estimates suggest that existing projections of learning loss are, if anything, too conservative. This is alarming in light of the much larger losses projected in countries less prepared for the pandemic. At the same time, our results may underestimate costs even in the context that we study. Schools remained at reduced capacity following re-openings. Dynamic models demonstrate how small initial losses can accumulate into large disadvantages with time (*29*). Test scores do not consider children's psycho-social development (*7*), neither productivity decline or heightened pressure among parents (*8*). Overall, our results highlight the importance of social investment strategies to "build back better" and enhance resilience and equity. Further research is needed to assess the success of such initiatives, and address the long-term fallout of the pandemic for student learning and well-being.

References

1. United Nations, *Education during COVID-19 and beyond* (UN Policy Briefs, 2020).

2. United Nations, *United Nations, Treaty Series* **1577** (1989).

3. S. K. Brooks *et al.*, *Eurosurveillance* **25** (2020).

4. R. M. Viner *et al.*, *The Lancet Child & Adolescent Health* (2020).

5. G. Defeyter *et al.*, *Back to school post COVID-19: Rebuilding a better future for all children* (Education Committee, UK Parliament, 2020).

6. M. Kuhfeld *et al.*, *Projecting the potential impacts of COVID-19 school closures on academic achievement*, EdWorkingPaper, Annenberg Institute at Brown University, 2020, (https://doi.org/10.26300/cdrv-yw05).

7. E. Golberstein, H. Wen, B. F. Miller, *JAMA Pediatrics* (2020).

8. H. van Ballegooijen, L. Goossens, R. H. Bruin, R. Michels, M. Krol, *Concerns, quality of life, access to care and productivity of the general population during the first 8 weeks of the coronavirus lockdown in Belgium and the Netherlands*, medRxiv, 2020, (https://www.medrxiv.org/content/10.1101/2020.07.24.20161554v1).

9. R. Chetty, J. N. Friedman, N. Hendren, M. Stepner, *How did covid-19 and stabilization policies affect spending and employment?*, (https://www.nber.org/papers/w27431).

10. DELVE Initiative, *Balancing the risks of pupils returning to schools*, DELVE Report No. 4. Published 24 July, 2020, (http://rs-delve.github.io/reports/2020/07/24/balancing-the-risk-of-pupils-returning-to-schools.html).

11. J. E. Maldonado, K. De Witte, *The effect of school closures on standardised student test outcomes*, KU Leuven Department of Economics Discussion Paper DPS20.17, 2020, (https://feb.kuleuven.be/research/economics/ces/documents/DPS/2020/dps2017.pdf).

12. A. Schleicher, *PISA 2018: Insights and interpretations* (OECD Publishing, 2018).

13. OECD, *Students, computers and learning: Making the connection* (OECD Publishing, 2015).

14. G. Di Pietro, F. Biagi, P. Costa, Z. Karpinski, J. Mazza, *The likely impact of COVID-19 on education: Reflections based on the existing literature and recent international datasets* (Publications Office of the European Union, 2020).

15. F. M. Reimers, A. Schleicher, *A framework to guide an education response to the COVID-19 Pandemic of 2020* (OECD Publishing).

16. M. de Haas, R. Faber, M. Hamersma, *Transportation Research Interdisciplinary Perspectives*, 100150 (2020).

17. M. E. Kuiper *et al.*, *The intelligent lockdown: Compliance with COVID-19 mitigation measures in the Netherlands*, PsyArXiv, 2020, (https://psyarxiv.com/5wdb3).

18. T. Bol, *Inequality in homeschooling during the Corona crisis in the Netherlands: First results from the LISS Panel* (SocArXiv, 2020), (https://osf.io/preprints/socarxiv/hf32q/).

19. P. Engzell, A. Frey, M. D. Verhagen, *Pre-analysis plan for: Learning inequality during the COVID-19 pandemic*, Open Science Framework, 2020, (https://osf.io/qtndg/).

20. K. F. Vlug, *Education and Information Technologies* **2**, 287–306 (1997).

21. J. D. Angrist, J. Pischke, *Mostly harmless econometrics: An empiricist's companion* (Princeton University Press, 2008).

22. H. S. Bloom, C. J. Hill, A. R. Black, M. W. Lipsey, *Journal of Research on Educational Effectiveness* **1**, 289–328 (2008).

23. G. W. Imbens, J. M. Wooldridge, *Journal of Economic Literature* **47**, 5–86 (2009).

24. J. Hainmueller, *Political Analysis*, 25–46 (2012).

25. P. D. Marshall, M. Losonczy-Marshall, *Psychological Reports* **107**, 567–577 (2010).

26. R. J. Park, J. Goodman, A. P. Behrer, *Nature Human Behaviour* (2020).

27. J. P. Azevedo, A. Hasan, D. Goldemberg, S. A. Iqbal, K. Geven, *Simulating the potential impacts of covid-19 school closures on schooling and learning outcomes: A set of global estimates* (The World Bank, 2020).

28. E. Dorn, B. Hancock, J. Sarakatsannis, E. Viruleg, *COVID-19 and student learning in the United States: The hurt could last a lifetime* (McKinsey & Company, 2020).

29. M. Kaffenberger, *Modeling the long-run learning impact of the COVID-19 learning shock: actions to (more than) mitigate loss*, RISE Insight Series, 17, 2020.

30. G. Psacharopoulos, V. Collis, H. A. Patrinos, E. Vegas, *Lost wages: The COVID-19 cost of school closures* (The World Bank, 2020).

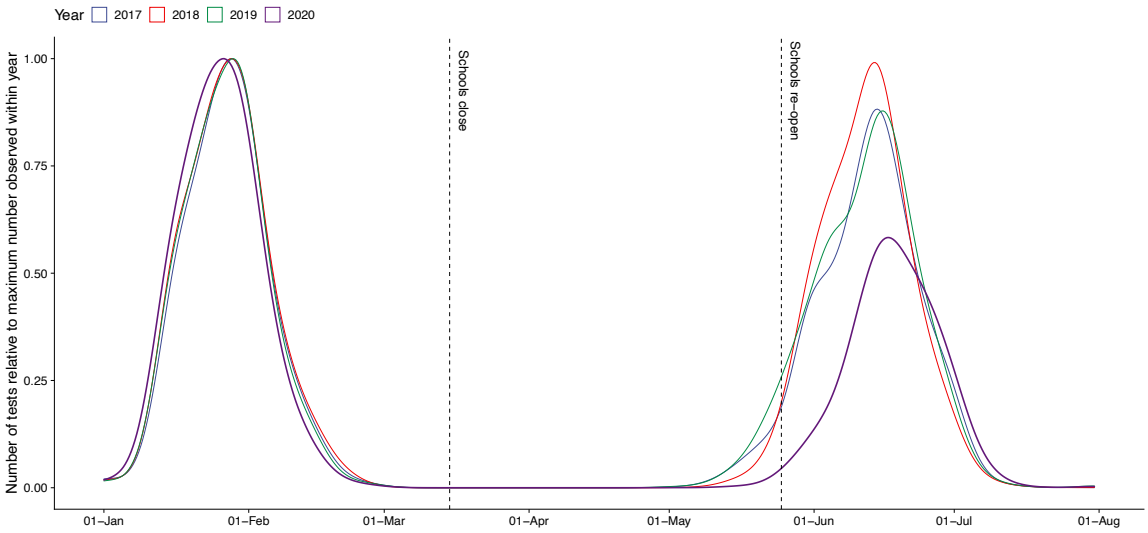


Figure 1. Distribution of testing dates 2017–2020 and timeline of 2020 school closures. Density curves show the distribution of testing dates for national standardized assessments in 2020 and the three comparison years 2017–2019. Vertical lines show the beginning and end of nationwide school closures in 2020. Schools closed nationally on March 16 and re-opened on May 11, after 8 weeks of remote learning. Our difference-in-differences design compares learning progress between the two testing dates in 2020 to that in the three previous years. End-of-year tests in 2020 were delayed relative to earlier years and a smaller proportion of students were tested. Our baseline model adjusts for time elapsed between testing dates and a linear trend in year, and only includes students who took a test on both occasions in a given year. In supplementary analyses, we address loss to follow-up by regression adjustment, re-balancing on the propensity score and maximum entropy weights, and fixed-effects comparison within schools and families.

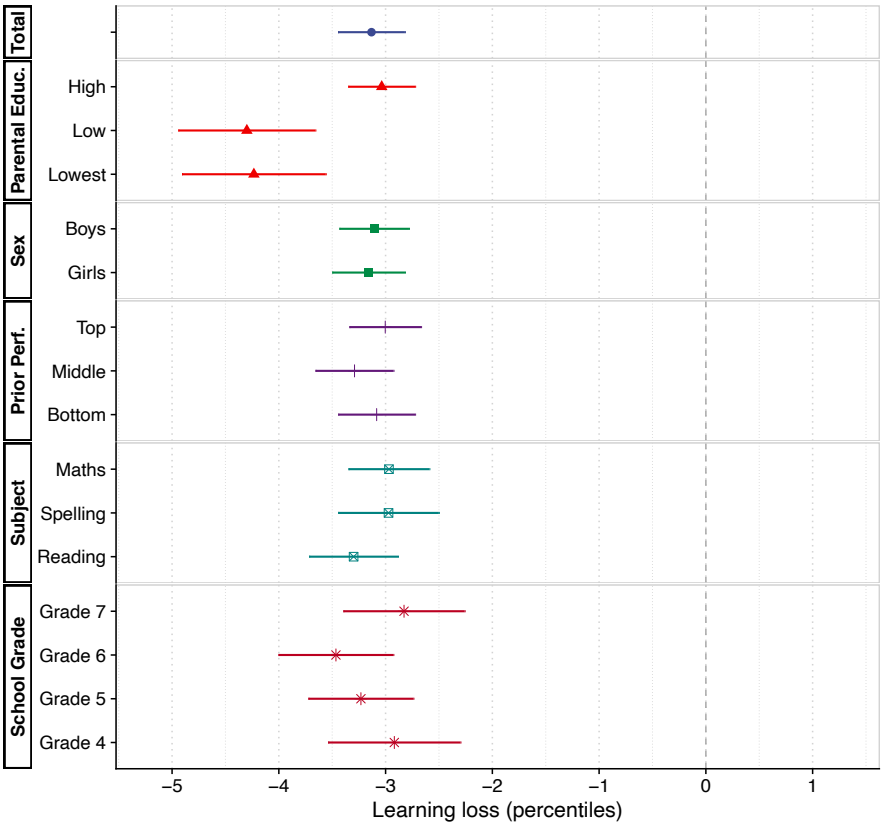


Figure 2. Estimates of learning loss for the whole sample and by subgroup and test. The graph shows estimates of learning loss from a difference-in-differences specification that compares learning progress between the two testing dates in 2020 to that in the three previous years. Statistical controls include time elapsed between testing dates and a linear trend in year; additional analyses adjusting for observed and unobserved heterogeneity are presented in the Supplementary Materials. Point estimates with 95% confidence intervals, robust standard errors accounting for clustering at the school level. Tests are graded on a percentile scale, but using the same grading norm across years, so the score can be interpreted on an absolute scale. One percentile point corresponds to approximately 2.5% of a standard deviation. Where not otherwise noted, effects refer to a composite score of Maths & Arithmetics, Spelling, and Reading Comprehension.

Supplementary Materials for

Learning inequality during the COVID-19 pandemic

Per Engzell[a,b,c], Arun Frey[d], and Mark Verhagen[a,b]

[a]Leverhulme Centre for Demographic Science, 42 Park End St, Oxford OX1 1JD, UK

[b]Nuffield College, University of Oxford, 1 New Rd, Oxford OX1 1NF, UK

[c]Swedish Institute for Social Research, Stockholm University, 106 91 Stockholm, Sweden

[d]Department of Sociology, University of Oxford, 42 Park End St, Oxford OX1 1JD, UK

This PDF file includes:

Contents

List of Figures

List of Tables

1 Study context

Figure S1 provides a timeline of school closures in the Netherlands, compared with 32 OECD countries for which data could be located. These data are sourced from the Oxford COVID-19 Government Response Tracker located at the Blavatnik School of Government, University of Oxford. In comparative perspective, the length of school closures in the Netherlands was brief: schools throughout the country closed on March 16, and reopened on May 11, after eight weeks of remote learning. While students initially attended classes every other day, since June 8 in-person education at primary schools is once again open for all students and has returned back to normal activity.[1]

The Netherlands is located close to the OECD average in terms of school spending and reading performance in PISA, but places nearer the top in mathematics performance (*1*). Several reasons lead us to see the country as a best-case scenario that likely presents a lower bound on consequences of school closures elsewhere. Non-pharmaceutical interventions have been mild compared to its European neighbors (*2*), as has the pandemic's toll on human lives and livelihoods. Prior to the pandemic, the Netherlands was a global leader in technology adoption (*3*), and in 2019, more than 90% of households enjoyed broadband access even among the poorest quartile (*4*). Adding to this advantage, the policy response has been swift: already in March 2020, the Ministry of Education devoted 2.5 million euros to online learning devices for students in need (*5*), and this scheme has subsequently been extended with another 3.8 million in June 2020 (*6*).

Despite these measures, initial survey evidence suggests that the pandemic-induced school closure has taken a toll on parents and children, particularly in disadvantaged households. A representative survey on learning in lockdown (*7*) found that only 2 in 3 primary school children had access to their own workspace and computer or tablet for schoolwork. A similar proportion of parents reported regularly helping their child with schoolwork, a figure which was lower among low-educated parents.

[1] See a description of the Dutch measures against COVID-19: https://www.government.nl/topics/coronavirus-covid-19/tackling-new-coronavirus-in-the-netherlands/public-life

2 Data

The main question we ask is whether learning stalled during lockdown, and whether students from less-educated homes were disproportionately affected. In addition, we examine differences by student sex, school grade, subject domain, and prior performance. To answer our research questions, we leverage a dataset containing individual-level test scores and demographic characteristics of a large number of primary school students in the Netherlands. The dataset has been made available to us in a fully anonymized way (not traceable to any individual student or school) by an educational service provider, and contains data on the educational performance of students at 15% of all primary schools in The Netherlands.

Although covering a subset of schools, the sample is large (n≈350,000) and broadly representative of the student population. Figure S2 compares the distribution of key school characteristics in the sample and the population as a whole. As this figure shows, there is some over-representation of mid-sized schools (101–200 students). Crucially, however, the relative representation of public as opposed to private schools is identical to that in the student population, as is the socioeconomic composition of the sample as represented by the "student weight" assigned by the Central Bureau of Statistics on the basis of parental education. We describe this measure in greater detail in Section 2.2 below.

2.1 Dependent variables

To explore the impact of the COVID-19-induced lockdown on educational outcomes, we examine how students' progress on nationally-standardized tests compares to progress in previous cohorts. These tests are conducted across three main didactic areas: Maths & Arithmetics, Spelling, and Reading Comprehension, each of them lasting about 45–60 minutes. All students across schools in the Netherlands take the same exam within a given year. Test results are transformed to percentile scores, but the norm for transformation is the same across years so absolute changes in performance over time are preserved. We rely on translation keys provided by the test producer Cito (https://www.cito.nl/) to assign each student a percentile score in the national distribution. However, as these keys

are actually based on smaller samples than that at our disposal, we further re-norm the distribution to ensure that it is uniform within our sample, standardizing the distribution within each subject and school grade (but not year).

Our main outcome is a composite score that takes the average of all non-missing values in the three areas Maths & Arithmetics, Spelling, and Reading Comprehension. In sensitivity analyses in Section 4.1 we require a student to have a valid score on all three subjects. We also display separate results for the three sub-tests in main manuscript Figure 2, and throughout. The test in Maths & Arithmetics is based on a set of abstract problems, as well as contextual problems where a student is asked to solve a concrete task. The test in Reading Comprehension assesses the student's ability to understand written texts, in particular business texts and fictional, narrative and literary texts. The test in Spelling asks students to write down a series of words. The spelling rules themselves are not explicitly asked for; instead the student indirectly shows that he or she has mastered them by writing down the requested words correctly.[2]

As an alternative outcome we also assess students' performance on short, 3-minute tests of "Information Processing". The purpose of these tests is somewhat different and they were therefore not included among our preregistered analyses (*8*). Nevertheless, we include them in this Supplementary Material for completeness, and to shed additional light on the mechanisms underlying learning loss. These tests aim to establish how well students can understand technical material and how this skill develops over the years. These assessments are not designed to test actual new didactic material, but instead students are presented with an infographic (e.g., a map) and asked to answer a set of questions regarding their comprehension. As this part of the assessment does not test for the retention of curricular content, we would expect it to be less affected by school closures, which is indeed what we find. We display these results in Section 4.3 below.

[2]See documentation at https://www.expertgroeptoetsenpo.nl/c/cito-bv

2.2 Independent variables

Parental education Our main measure of socioeconomic disadvantage is parental education as reflected in the so-called "student weight" used by the Dutch Ministry of Education to allocate resources to schools. The measure is produced by CBS, the Dutch Central Bureau for Statistics. This is a proxy similar to the eligibility for free school meals (FSM) used in some English speaking countries. Unlike FSM which usually draws on parental income, the Dutch student weight aggregates information about parental education and takes on three values:

- 'High education': At least one parent has a degree above lower secondary education.
- 'Low education': Both parents have a degree above primary education but neither has one above lower secondary education.
- 'Lowest education': At least one parent has no degree above primary education and neither has a degree above lower secondary education.

Student sex We rely on information on student's sex to differentiate between the effect of learning loss among female and male students.

Prior performance We use information on test results in the previous year for a given student to operationalize past performance, and create a categorical variable that distinguishes between students whose performance places them in the top, middle, and bottom tertile.

School grade We include students in grades 4–7, aged approximately 7–11 (students in the Netherlands start school at age 4). The didactic material in the first three grades (the first two of which are practically kindergarten) is less intense than the later grades and grade 8 does not feature much additional didactic material. The final grade is dedicated to transitioning to secondary education and is shorter than the other grades.

2.3 School-level variables

In Section 4.4, we plot school-level treatment effects by two measures of school composition: socioeconomic disadvantage and the proportion of non-Western inhabitants in the school neighborhood. We also enter these variables into our matched analyses in Section 4.2.4, where we balance observations on the propensity of treatment and using maximum entropy weights. The school-level variables are defined as follows.

Socioeconomic disadvantage This score ranges from 20 to 40, and is a composite based on the parental education variable used in our main analyses, with weights assigned according to the norm used by the Central Bureau of Statistics: 0.3 for "Low education" (both parents have a degree above primary education but neither has one above lower secondary education) and 1.2 for "Lowest education" (at least one parent has no degree above primary education and neither has a degree above lower secondary).

Proportion immigrant background The proportion of non-Western inhabitants is based on the neighborhood in which a school is located, which closely correlates with a school's student composition as most students attend a school close to their home. A person is defined as having a non-Western background if they or at least one of their parents were born in Turkey or countries in Africa, Latin America and Asia, except former Dutch colonies and Japan. This variable is not available at an individual level in our data.

School denomination In our matched analysis in Section 4.2.4, we include school denomination as an additional control variable. Here we distinguish between three categories: public schools, Christian schools (including Protestant, Catholic, and Reformist denominations), and other denominations (including *inter alia* Islamic and Waldorf schools).

3 Analytical strategy

3.1 Difference-in-difference analysis

The standardized tests that we use for our analysis are taken biannually, halfway into and at the end of each academic year. Since these tests happened to have been conducted immediately prior to and following the COVID-19 lockdown, it is possible to examine how the lockdown impacted student learning by comparing the change in performance pre- and post-lockdown to similar changes in previous years.

To identify the overall level of learning loss due to the primary school closure, we compare educational achievement pre-lockdown (measured using the middle-of-year test) to achievement post-lockdown (measured using the end-of-year test): $\Delta y_i^{2020} = y_i^{\text{2020-post}} - y_i^{\text{2020-pre}}$, where y_i is some achievement measure for student i and the superscript 2020 denotes the treatment year when school closures took place. In particular, we are interested in how Δy_i^{2020} compares to the counterfactual situation without the COVID-19 pandemic. The latter can be approximated through $\Delta y_i^{2017-2019}$, which is the same difference in the three years prior to the pandemic, amounting to a difference-in-differences setup. The overall effect of the pandemic can then be analyzed as:

$$\Delta y_i = \beta_0 + \mathbf{X}_i'\gamma + \beta_1 T_i + \epsilon_i \tag{1}$$

where Δy reflects the difference between end-of-year achievement and middle-of-year achievement, $\mathbf{X}_i'\gamma$ is a set of individual control variables, and T_i is a dummy reflecting the pandemic year 2020. The coefficient $\hat{\beta}_1$ reflects overall learning loss due to the pandemic. From here, we estimate a model including one of our various interest variables and its interaction with the treatment variable to reflect possible heterogeneity in pandemic-related learning loss:

$$\Delta y_i = \beta_0 + \mathbf{X}_i'\gamma + \beta_1 T_i + \beta_2 A_i + \beta_3 T_i A_i + \epsilon_i \tag{2}$$

where A_i is a categorical variable reflecting some student characteristic. We estimate

Equation (2) in turn including parental education, student sex, and prior performance in A_i. In addition, we estimate Equation (1) separately for the various available grades (grade 4–7) as well as by subject domain (Maths, Spelling, or Reading Comprehension) to examine hetereogeneity by grade and subject. Throughout all our analyses, we adjust confidence intervals for clustering on schools using robust standard errors.

We include two sets of control variables in the vector $\mathbf{X}_i$. In our main specification (used in Figure 2 of the main manuscript), we include controls for the time elapsed between testing dates as well as a linear term for year to adjust for secular time trends. In robustness analyses in Section 4.1, we additionally include controls for parental education, student sex, and prior performance. In Section 4.2, we perform additional analyses using propensity score and entropy balancing techniques to match control and treatment group on a wider range of individual- and school- level characteristics, and estimate results using school and sibling fixed effects.

3.2 Propensity score and entropy weighting

To adjust for potential selection into the test-taking sample in 2020 and thereby ensure the comparability of treatment and control groups, we match on a wider range of individual- and school-level characteristics in Section 4.2.4. We do so using two methods: reweighting on the propensity of treatment (*9*–*11*) and using an entropy balancing procedure (*12*, *13*). In both cases, we constrain the treatment and comparison group to be balanced on the following covariates: parental education, student sex, prior performance, school-level socioeconomic disadvantage, proportion immigrant background, and school denomination. Propensity of treatment weights (*11*) involve first estimating the probability of treatment using a binary response (logit) model and then reweighting observations so that they are balanced on this propensity across comparison and treatment groups. The entropy balancing procedure (*12*) instead uses maximum entropy weights that are calibrated to directly balance comparison and treatment groups on the observed covariates.

3.3 School fixed effects

As a further test for selection into the sample we introduce a within-school analysis using a fixed-effects specification in Section 4.2.5. The fixed-effects design discards all variation between schools by introducing a separate intercept for each school (*14*). By doing so, it eliminates all unobserved heterogeneity across schools which might have biased our results if, for example, schools that perform worse in all years are over-represented during the treatment year.

3.4 Family fixed effects

By the same logic, we introduce family-level fixed effects in Section 4.2.6. Similarly to the within-school analyses, this within-family analysis discards all variation between families by introducing a separate intercept for each group of siblings identified in our data (*14*). This step reduces the size of our sample by approximately 60%, as not every student has a sibling attending a sampled school within the years that we are able to observe. The benefit is that it allows us to remove any unobserved confounding at the family level.

3.5 Mixed model

In Section 4.4 we explore heterogeneity in the treatment effect across schools by fitting a mixed model with random slopes (*15*, *16*). We estimate two version of this model: one without any covariates except controls for testing date and year trend included in our baseline specification, and one that further includes individual-level controls: parental education, prior performance, and sex. In both cases, we plot the predicted school-level treatment effects against school-level socioeconomic disadvantage and the share of non-Western immigrants in the school neighborhood (see Section 2.3 above for definitions).

4 Additional results

This section contains table output underlying the results reported in Figure 2 of the main manuscript; placebo analyses for non-treatment years; robustness analyses leaving out individual years; strategies for dealing with attrition, including covariate balancing and school and family fixed effects; results using the alternative outcome of Information Processing; and finally school-level treatment effects.

4.1 Regression tables

Table S1 reports regression output for the results reported in main text Figure 2, panels reporting pooled learning loss across all three main subjects and separately by subject domain. The treatment effect amounts to –3.13 percentile points for the composite achievement score, and ranges between –2.97 and –3.30 for the three separate subjects. Table S2 shows results by parental education for the composite score as reported in main text Figure 2, as well as results for separate subjects. Table S3 does the same for student sex and Table S4 does so for prior performance. In Table S5 we report additional regression results simultaneoulsy controlling for individual-level covariates: parental education, student sex, prior performance. Doing so does not appreciably alter the treatment effect which remains estimated at –3.12 and significant at the 0.1% level. In Table S6 we restrict the sample to only those students with a valid score in all three subjects, with highly similar results.

4.2 Robustness analyses

In this subsection we perform a series of placebo analyses and robustness checks. First, we run placebo analyses assigning treatment status to each of the three comparison years to confirm that estimated learning loss in these years is zero. Next, we re-estimate our main specification dropping comparison years one at a time to see that our results are not driven by individual years. Thereafter we introduce various ways to address loss to follow-up: by restricting analysis to schools with near-complete return after lockdown, balancing treatments and controls using maximum entropy and propensity score weights, and adjusting

for unobserved heterogeneity at the school and family levels using fixed effects.

4.2.1 Placebo comparisons

To confirm that our main specification delivers unbiased estimates, in Figure S4 we perform a placebo analysis on non-treated years. We do so by keeping the specification identical to our main analysis but excluding the actual treatment year and, in turn, assigning treatment status to each of the three comparison years. Doing so reveals few significant effects, and those that are so by chance are mostly in the opposite direction of the results reported in the main manuscript. Our identification strategy thus appears robust to false positives, and if anything, is likely to underestimate the treatment effect somewhat given the small bias towards a positive treatment effect in two of three control years. Crucially, however, the pooled effect across all groups is not significantly different from zero in any year.

4.2.2 Dropping comparison years

Next we reestimate our main specification dropping comparison years one at a time to confirm that our results are not driven by any one comparison year. Figure S5 reports the results from these analyses. Although the estimates are less precisely estimated, especially in the last analysis dropping the year immediately preceding the treatment, the qualitative results remain unchanged. Specifically, the difference in effect size between students from high- and low-educated homes remains similar and is significant at the 0.1% level throughout these analyses. The robustness of these results thus further corroborates our preferred specification reported in the main text.

4.2.3 Near-complete schools

As a first way to address loss to follow-up we restrict the sample to only schools where at least 75% of students were tested in the treatment year. Table S7 reports the main treatment effect using this restriction, which remains significant at the 0.1% level and near identical in magnitude to that of our main analysis (–3.18 in Table S7 vs –3.13 in Table S1 and main manuscript Figure 2). Table S8 further displays differential treatment effects by parental

education. The estimated treatment at low levels of education is, if anything, slightly larger than estimates reported in our main analysis (low parental education: –1.33 in Table S7 vs –1.26 in Table S2 and main manuscript Figure 2; lowest parental education: –1.42 in Table S7 vs –1.20 in Table S2 and main manuscript Figure 2). These tables report interaction effects, so the full disadvantage among the the groups with less-educated parents is found by adding the main effect estimate to the interaction term for each group.

4.2.4 Maximum entropy and propensity score weights

To further address loss to follow-up we implement re-weighting on the propensity of treatment and maximum entropy balancing as described in Section 4.2.4 above. Figure S6 shows that both methods achieve a sample that is balanced on the desired characteristics. The weighted regressions use either of these sets of unit weights to rebalance treatment and control groups and achieve comparability. Figure S7 displays our main results using both sets of weights. The results correspond closely across both weighting schemes and are also not appreciably different from our main specification as reported in Figure 2 of the main manuscript. These results further highlight that our results are not vulnerable to selective attrition of the sample in the treatment year.

4.2.5 School fixed effects

Next we estimate a model including separate intercepts for each school, to ensure that the relative representation of individual schools across years does not bias our treatment effect. This specification nets out stable differences across schools and thus adjusts for the fact that schools with higher or lower achievement might have tested a greater proportion of students in the treatment year. Table S9 shows the replication of the main results adding school fixed effects whereas Table S10 does so for the interaction by parental education. Again, the estimated treatment effect is significant at the 0.1% level and remains similar in magnitude to our estimates reported in the main text: –3.17 for the whole student body pooled (Table S9) with an added penalty of –1.30 and –1.33 for the groups from less-educated homes, over the baseline of –3.07 for those with at least one higher educated parent (Table S10).

4.2.6 Family fixed effects

Our last and most stringent robustness test introduces separate intercepts for each family. Thus we evaluate our main specification on students who had at least one or more sibling during the period under study (and at least one in the treatment year) to allow for the estimation of family fixed effects. We sacrifice roughly 60% of our original sample size as not every student has a sibling attending a sampled school within the years that we are able to observe. This makes the results somewhat less stable, but nevertheless they remain qualitatively similar. As Figure S8 shows, results remain similar in magnitude and highly significant. The main difference is that inequality by parental education increases slightly in our within-family design. Tables S11–S12 show the regression output underlying these results. The added penalty for children with parents in the 'low' and 'lowest' categories of education is respectively –1.34 and –1.74, which is larger than our baseline estimates of –1.26 and –1.20 in Table S2, although not statistically significantly so.

4.3 Information processing test

To adjudicate between mechanisms underlying our results, we further assess students' performance on short, 3-minute tests of "Information Processing" not designed to test curricular content. If our main estimates of learning loss reflect the actual cumulative impact of knowledge learned, we would expect these effects to be small or zero. In contrast, if our estimates of learning loss mainly reflect "day of exam" effects due to stress exposure, testing conditions, or familiarity with the school setting, we would expect similarly large losses on both kinds of test. Figure S9, top panel, reveals that the treatment effect on this outcome is on average 62% smaller than for our main outcome. Figure S9, bottom panel, shows that this is not the case in a non-treatment year, where estimated null effects on the both tests are instead near identical. Taken together, these results suggest that the main channel behind learning loss is not classroom conditions at the day of testing, but rather that actual exposure to and/or retention of curricular content is lower with remote instruction.

4.4 School-level treatment effects

Schools and student bodies may differ considerably in their ability to cope with learning during lockdown. Our results demonstrate that the impact of school closures is not equal, and that students from disadvantaged backgrounds fell further behind their more advantaged peers. A separate question is whether certain schools or areas are disproportionately affected—or put differently: are there schools that successfully manage to mitigate the effects of the pandemic? Figures S10–S11 report estimates from a mixed-effects model that lets the learning loss differ between schools. The results reveal considerable variation, with some schools seeing a learning slide of 10 percentile points or more, and others recording no losses or even small gains. Losses are larger in schools with a high proportion of students from less-educated homes and of immigrant background (Figure S10), and this holds further when adjusting for individual-level covariates (Figure S11).

References

1. A. Schleicher, *PISA 2018: Insights and interpretations* (OECD Publishing, 2018).
2. M. E. Kuiper *et al.*, *The intelligent lockdown: Compliance with COVID-19 mitigation measures in the Netherlands*, PsyArXiv, 2020, (https://psyarxiv.com/5wdb3).
3. OECD, *Students, computers and learning: Making the connection* (OECD Publishing, 2015).
4. G. Di Pietro, F. Biagi, P. Costa, Z. Karpinski, J. Mazza, *The likely impact of COVID-19 on education: Reflections based on the existing literature and recent international datasets* (Publications Office of the European Union, 2020).
5. F. M. Reimers, A. Schleicher, *A framework to guide an education response to the COVID-19 Pandemic of 2020* (OECD Publishing).
6. SIVON, *Opnieuw extra geld voor laptops en tablets voor onderwijs op afstand*, (https://www.sivon.nl/actueel/opnieuw-extra-geld-voor-laptops-en-tablets-voor-onderwijs-op-afstand/).

7. T. Bol, *Inequality in homeschooling during the Corona crisis in the Netherlands: First results from the LISS Panel* (SocArXiv, 2020), (https://osf.io/preprints/socarxiv/hf32q/).

8. P. Engzell, A. Frey, M. D. Verhagen, *Pre-analysis plan for: Learning inequality during the COVID-19 pandemic*, Open Science Framework, 2020, (https://osf.io/qtndg/).

9. P. R. Rosenbaum, D. B. Rubin, *Biometrika* **70**, 41–55 (1983).

10. K. Hirano, G. W. Imbens, *Health Services and Outcomes Research Methodology* **2**, 259–278 (2001).

11. G. W. Imbens, J. M. Wooldridge, *Journal of Economic Literature* **47**, 5–86 (2009).

12. J. Hainmueller, *Political Analysis*, 25–46 (2012).

13. Q. Zhao, D. Percival, *Journal of Causal Inference* **5** (2016).

14. J. M. Wooldridge, *Econometric analysis of cross section and panel data* (MIT press, 2010).

15. T. A. Snijders, R. J. Bosker, *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (Sage, 2011).

16. A. Gelman, J. Hill, *Data analysis using regression and multilevel/hierarchical models* (Cambridge University Press, 2006).

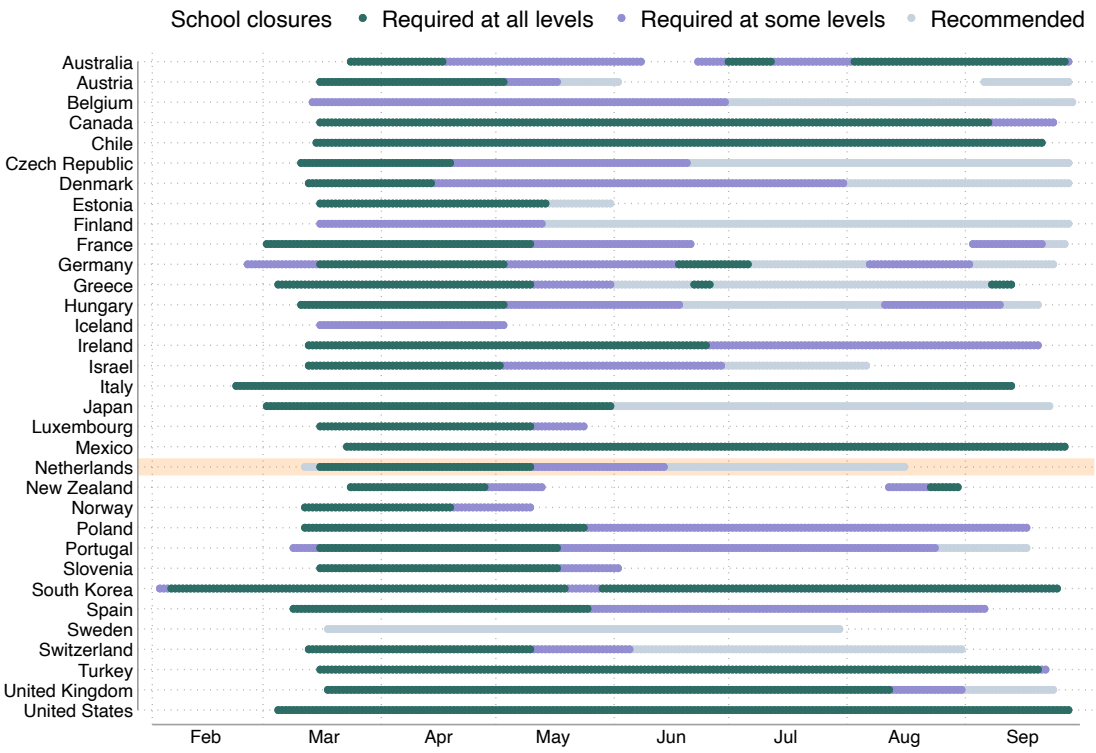


Figure S1. School closures in the OECD. The graph shows the onset and duration of school closures in 33 OECD countries, with the Netherlands marked in orange. Includes all OECD countries for which data could be located. Source: Oxford COVID-19 Government Response Tracker (`https://covidtracker.bsg.ox.ac.uk/`).

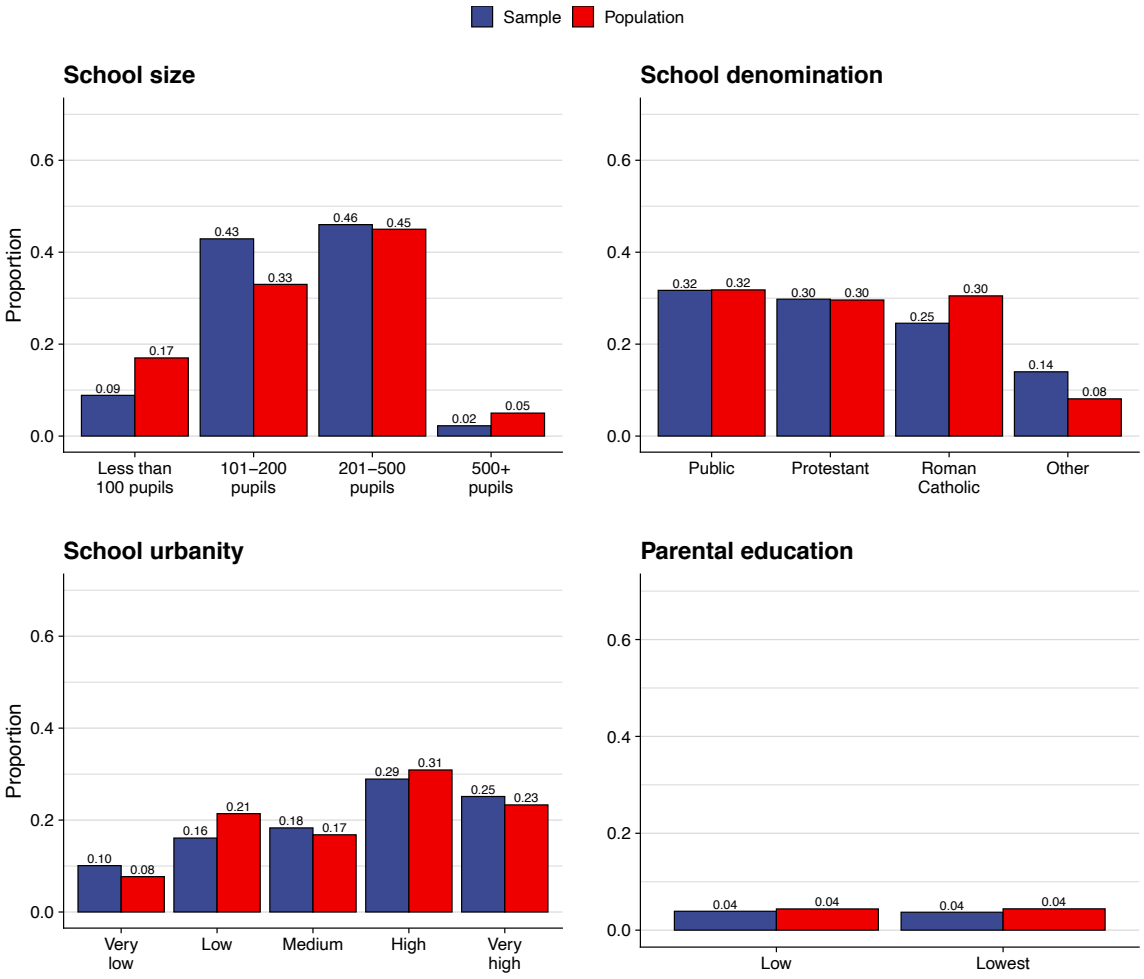


Figure S2. Representativity of the sample. The graph compares the distribution of school characteristics in our sample, shown in blue, with that of the universe of primary schools in the Netherlands, shown in red.

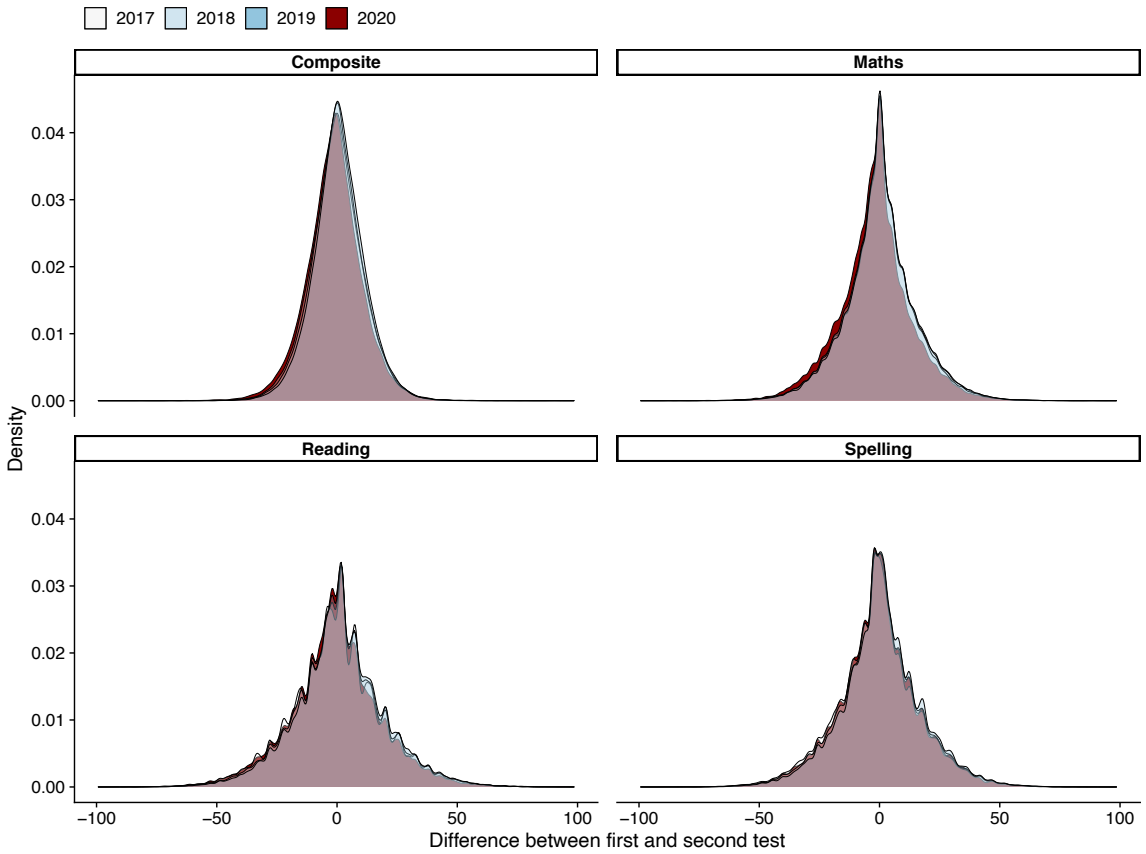
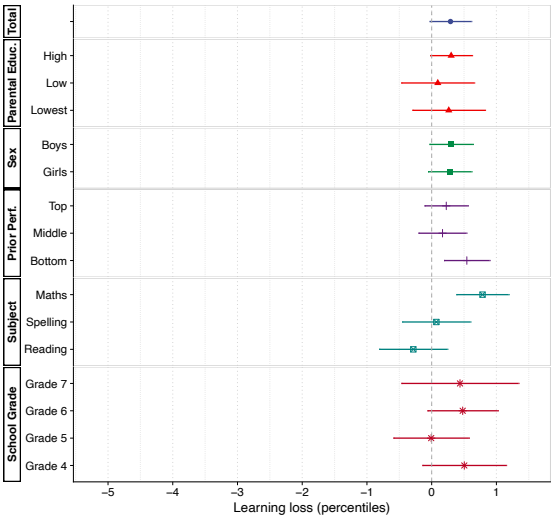
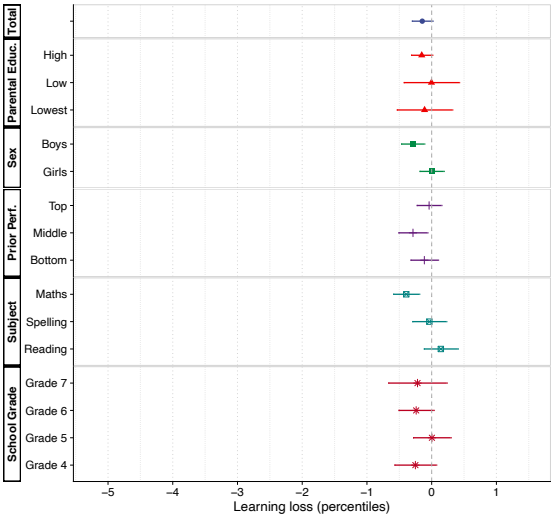


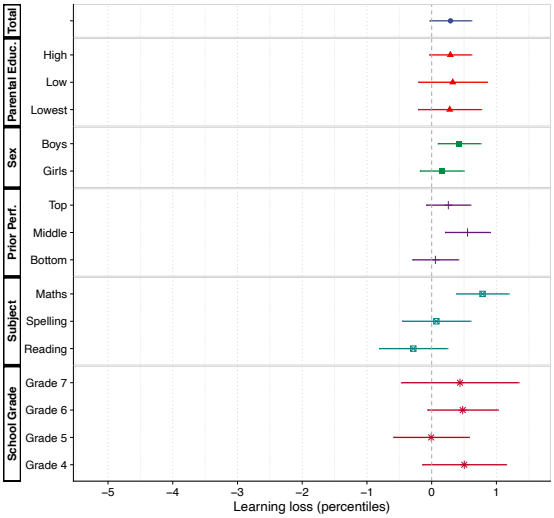
Figure S3. Difference in test scores 2017–2020. The graph shows the difference score that forms the main outcome of our analysis separately by year, with 2020 marked in red. The dark red area reflects learning loss in the treatment year. Note that this graph does not adjust for trends, testing date, or individual covariates.



(a) 2017 as treated

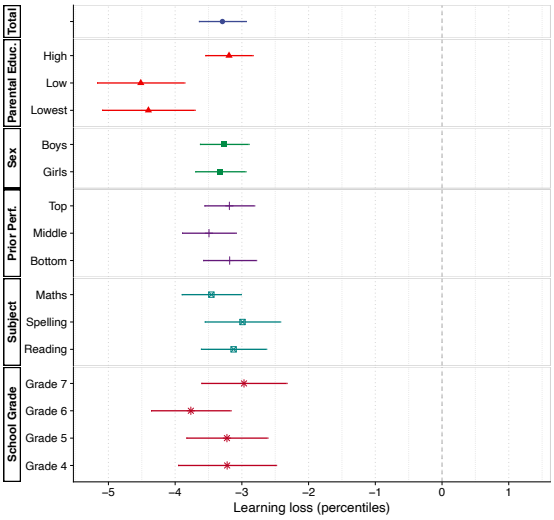


(b) 2018 as treated

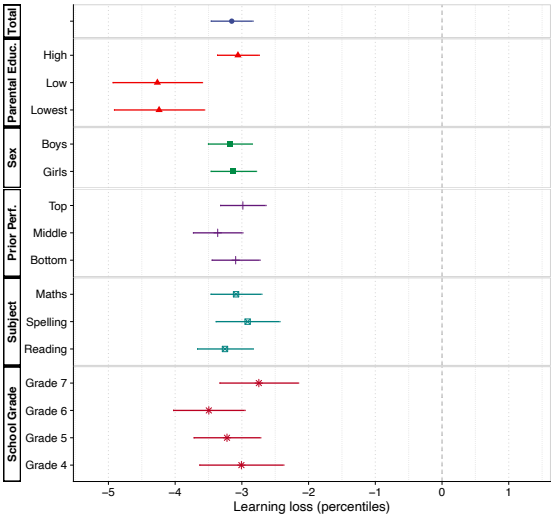


(c) 2019 as treated

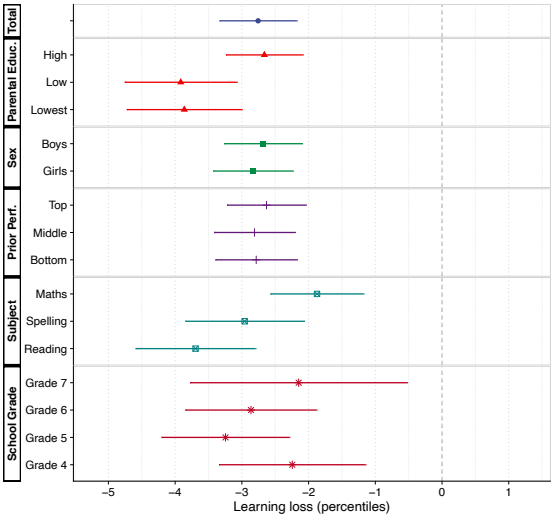
Figure S4. Placebo effects for non-treated years. The graphs show results using our main specification but excluding the actual treatment year and instead assigning treatment status to each comparison year.



(a) 2017 excluded



(b) 2018 excluded



(c) 2019 excluded

Figure S5. Robustness dropping comparison years. The graphs show results using our main specification but in turn excluding each comparison year from the analysis.

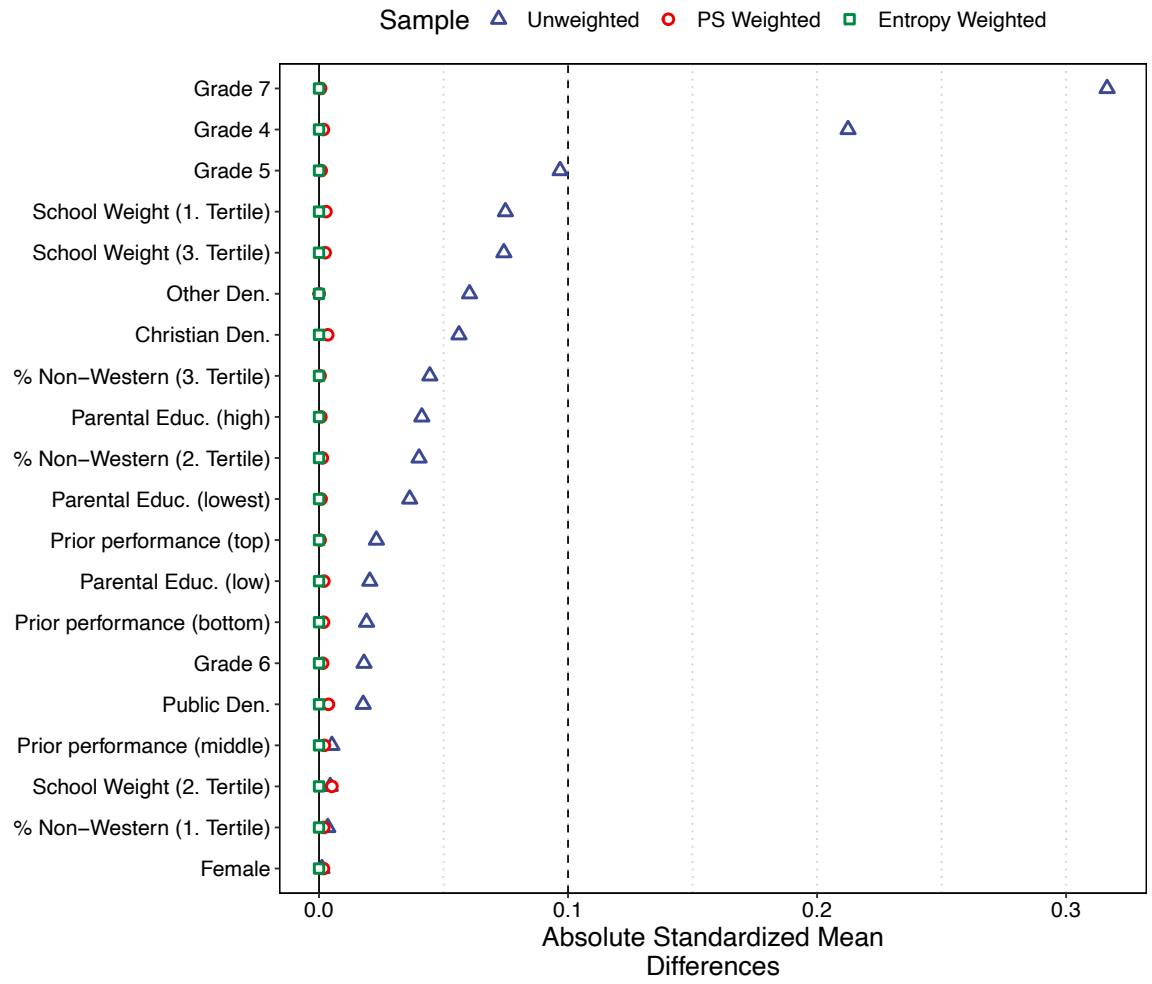


Figure S6. Balancing plot for weighted comparisons. The graph shows absolute standardized mean differences on balancing covariates between treatment and comparison years before adjustment and after reweighting on maximum entropy weights and the estimated propensity of treatment.

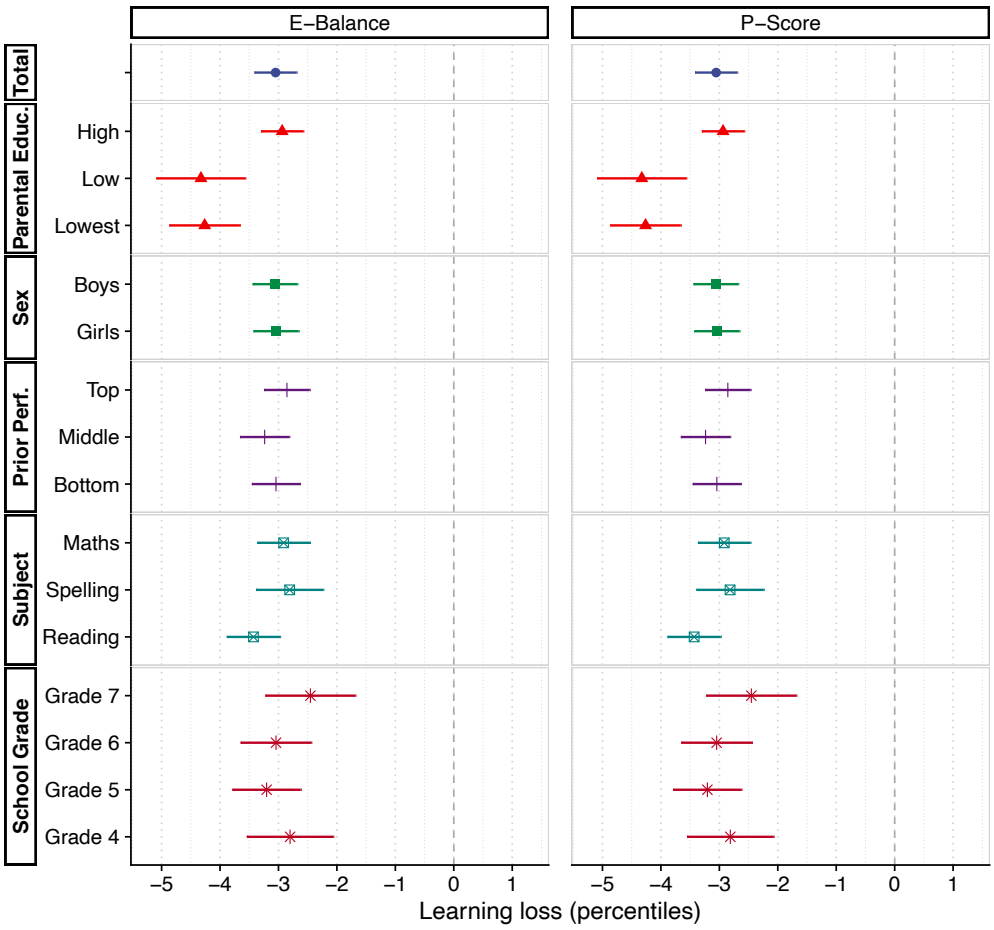


Figure S7. Entropy balanced and propensity-score weighted results. The graph shows results using our main specification while balancing treatment and control years on maximum entropy weights ("E-Balance," left) and the estimated propensity of treatment ("P-Score," right).

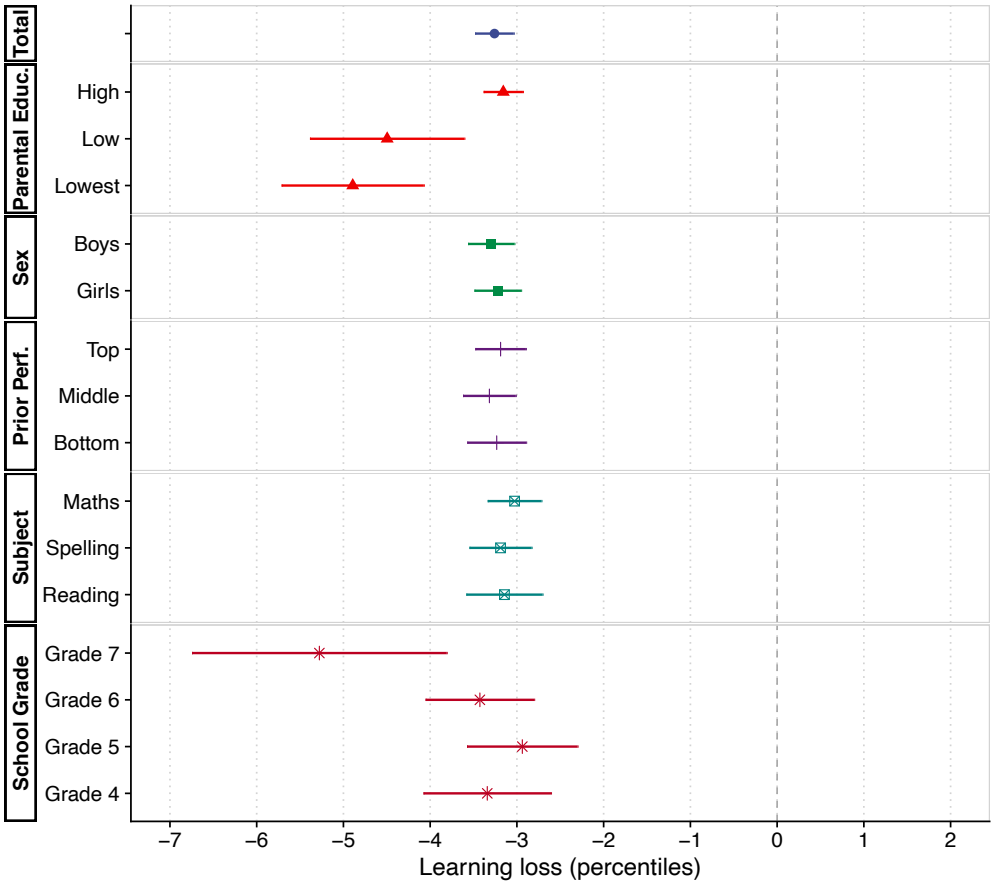
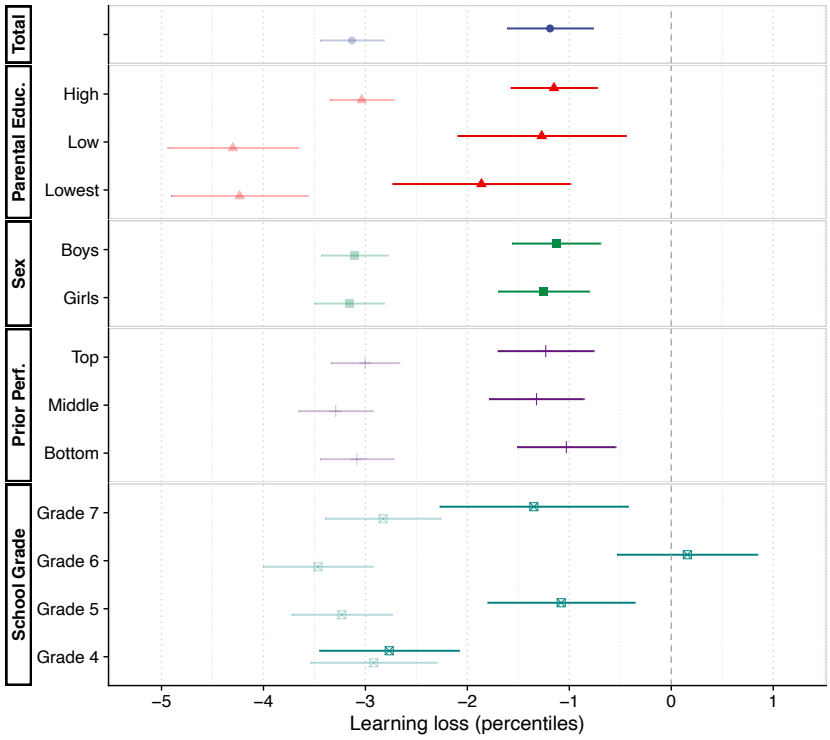
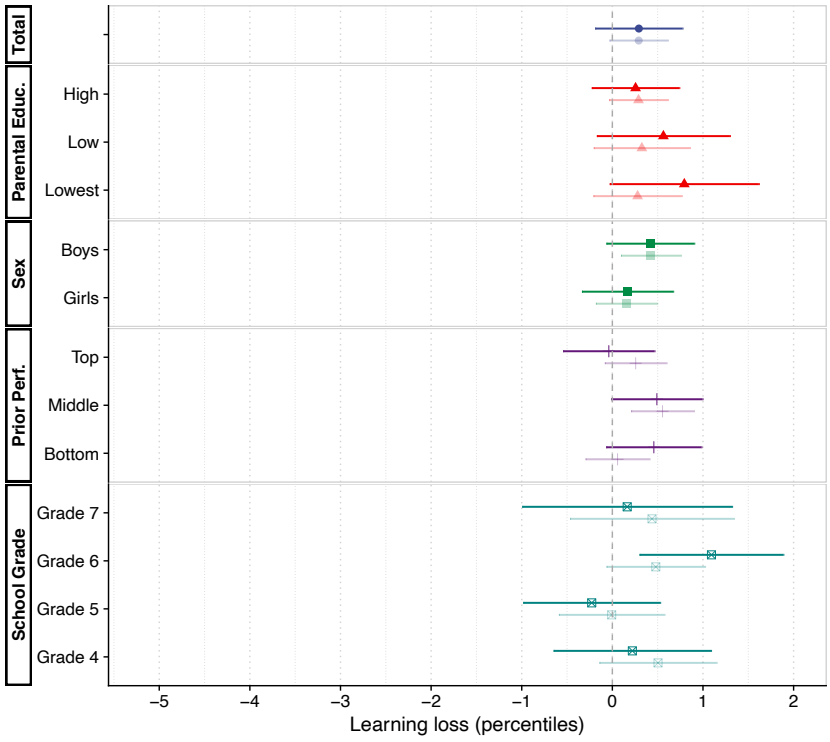


Figure S8. Family fixed effects. The graph shows results combining our difference-in-differences with family fixed effects. This analysis discards all variation between families by introducing a separate intercept for each sibling group, thus adjusting for any heterogeneity across families.

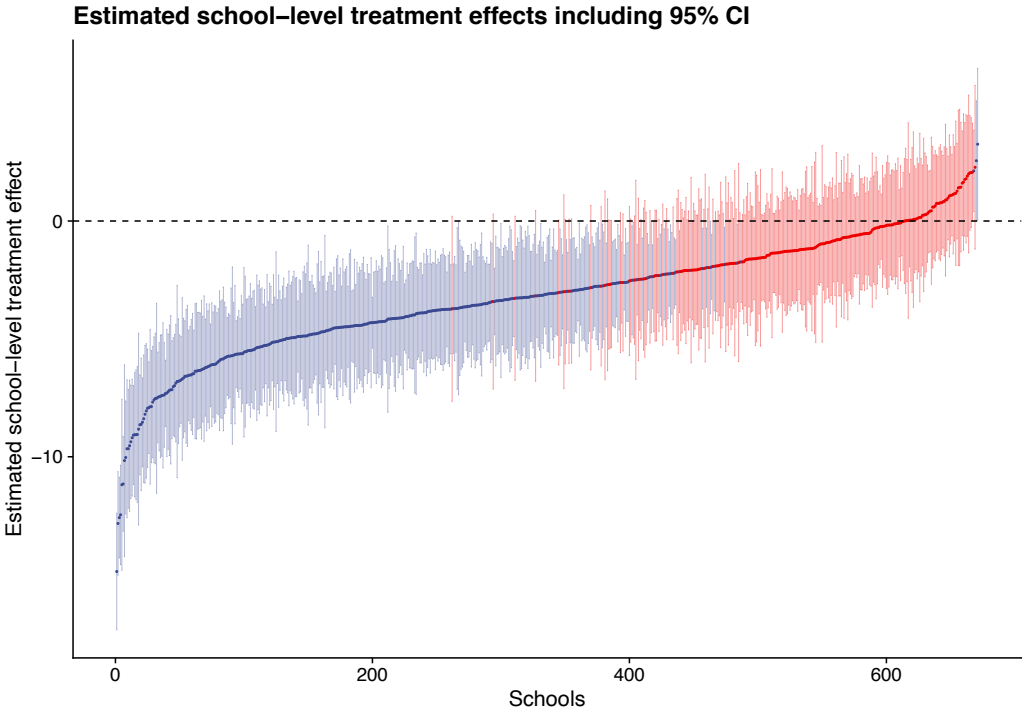


(a) Difference in treatment year

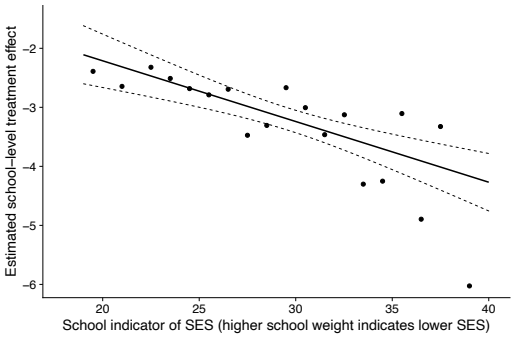


(b) Difference in placebo (2019)

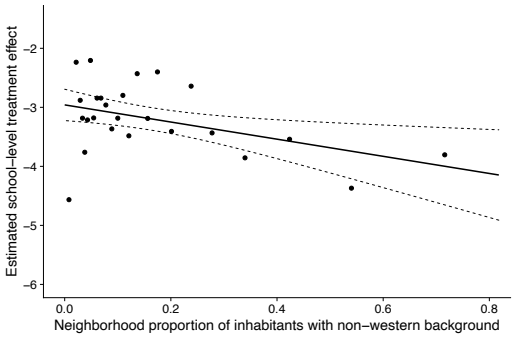
Figure S9. Results for information processing. The graph contrasts results on "Information Processing", in solid colors, with our composite achievement score, in transparent colors. The top panel shows estimated treatment effects for 2020, the bottom panel placebo results for 2019. The pooled treatment effect in 2020 is 62% smaller than that of our main analysis, arguably due to the fact that these tests do not assess curricular content.



(a) School-level treatment effects

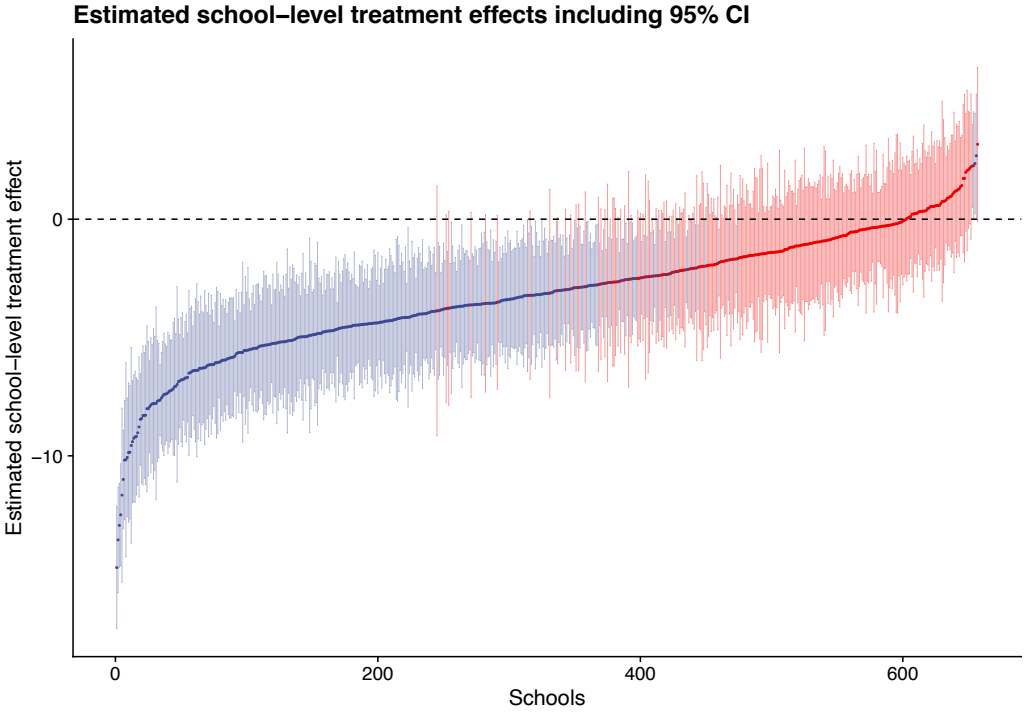


(b) School effect by SES

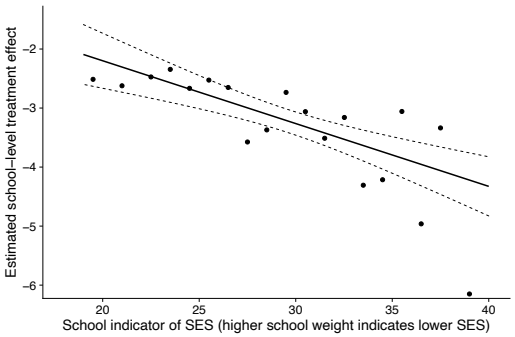


(c) School effect by proportion immigrants

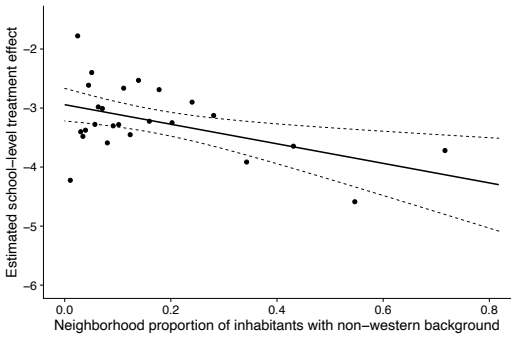
Figure S10. School-level effects. The top panel shows estimates of learning loss by school from a linear mixed model allowing learning loss to differ across schools. The bottom panels plot the predicted effects against school-level covariates: socioeconomic disadvantage and proportion non-Western immigrant background.



(a) School-level treatment effects



(b) School effect by SES



(c) School effect by proportion immigrants

Figure S11. School-level effects with individual controls. The top panel shows estimates of learning loss by school, the bottom panels plot predicted effects against school-level covariates. These results are identical to those in Figure S10 except school-level effects are adjusted for individual-level covariates: parental education, student sex, and prior performance.

Table S1. Main effects by subject.

	Composite	Maths	Reading	Spelling
Treatment	-3.13^{***}	-2.97^{***}	-3.30^{***}	-2.97^{***}
	(0.16)	(0.19)	(0.21)	(0.24)
Year (std.)	0.70^{***}	0.21^{**}	0.94^{***}	0.98^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.48^{***}	0.54^{***}	0.20^{**}	0.64^{***}
	(0.05)	(0.05)	(0.07)	(0.09)
(Intercept)	0.69^{***}	0.56^{***}	1.11^{***}	0.60^{***}
	(0.06)	(0.07)	(0.08)	(0.11)
R^2	0.01	0.00	0.00	0.00
Adj. R^2	0.01	0.00	0.00	0.00
Num. obs.	358407	352656	272382	343151
RMSE	11.04	14.86	18.78	17.00
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S2. Results by parental education and subject.

	Composite	Maths	Reading	Spelling
Treatment	-3.04^{***}	-2.85^{***}	-3.24^{***}	-2.86^{***}
	(0.16)	(0.19)	(0.21)	(0.24)
Treat x Par. Educ. (low)	-1.26^{***}	-1.11^{**}	-1.15^{*}	-1.72^{***}
	(0.29)	(0.39)	(0.48)	(0.44)
Treat x Par. Educ. (lowest)	-1.20^{***}	-1.81^{***}	-0.41	-1.29^{*}
	(0.33)	(0.42)	(0.43)	(0.51)
Parental Educ. (low)	-0.26^{*}	-0.29	-0.58^{**}	0.11
	(0.12)	(0.16)	(0.21)	(0.20)
Parental Educ. (lowest)	0.21	0.41^{*}	-1.00^{***}	0.89^{**}
	(0.16)	(0.17)	(0.20)	(0.29)
Year (std.)	0.70^{***}	0.21^{**}	0.94^{***}	0.98^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.48^{***}	0.54^{***}	0.20^{**}	0.64^{***}
	(0.05)	(0.05)	(0.07)	(0.09)
(Intercept)	0.70^{***}	0.55^{***}	1.18^{***}	0.56^{***}
	(0.06)	(0.07)	(0.08)	(0.12)
R^2	0.01	0.00	0.00	0.00
Adj. R^2	0.01	0.00	0.00	0.00
Num. obs.	358407	352656	272382	343151
RMSE	11.03	14.86	18.78	17.00
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S3. Results by student sex and subject.

	Composite	Maths	Reading	Spelling
Treatment	-3.11^{***}	-2.97^{***}	-3.28^{***}	-3.00^{***}
	(0.17)	(0.20)	(0.23)	(0.26)
Treat x Female	-0.05	-0.01	-0.04	0.05
	(0.12)	(0.15)	(0.19)	(0.18)
Female	0.26^{***}	0.66^{***}	-0.91^{***}	0.69^{***}
	(0.04)	(0.06)	(0.08)	(0.07)
Year (std.)	0.70^{***}	0.21^{**}	0.94^{***}	0.98^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.48^{***}	0.54^{***}	0.20^{**}	0.64^{***}
	(0.05)	(0.05)	(0.07)	(0.09)
(Intercept)	0.56^{***}	0.23^{**}	1.56^{***}	0.26^{*}
	(0.06)	(0.07)	(0.09)	(0.12)
R^2	0.01	0.00	0.00	0.00
Adj. R^2	0.01	0.00	0.00	0.00
Num. obs.	358407	352656	272382	343151
RMSE	11.04	14.85	18.78	17.00
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S4. Results by prior performance and subject.

	Composite	Maths	Reading	Spelling
Treatment	-3.00^{***}	-3.11^{***}	-3.15^{***}	-2.46^{***}
	(0.17)	(0.21)	(0.24)	(0.27)
Treat x Prior Perf. (middle)	-0.29^{*}	-0.10	-0.22	-0.67^{***}
	(0.13)	(0.18)	(0.23)	(0.20)
Treat x Prior Perf. (bottom)	-0.08	0.62^{**}	-0.23	-0.90^{***}
	(0.16)	(0.21)	(0.25)	(0.24)
Prior Perf. (middle)	0.59^{***}	0.60^{***}	0.36^{***}	0.75^{***}
	(0.06)	(0.08)	(0.10)	(0.09)
Prior Perf. (bottom)	1.02^{***}	0.88^{***}	0.76^{***}	1.34^{***}
	(0.07)	(0.09)	(0.11)	(0.11)
Year (std.)	0.70^{***}	0.21^{**}	0.94^{***}	0.97^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.48^{***}	0.54^{***}	0.20^{**}	0.64^{***}
	(0.05)	(0.05)	(0.07)	(0.09)
(Intercept)	0.18^{**}	0.08	0.76^{***}	-0.06
	(0.07)	(0.08)	(0.09)	(0.12)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	358407	352656	272382	343151
RMSE	11.03	14.85	18.78	17.00
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S5. Main effects with controls.

	Composite	Maths	Reading	Spelling
Treatment	-3.12^{***}	-2.97^{***}	-3.22^{***}	-2.98^{***}
	(0.16)	(0.19)	(0.21)	(0.24)
Parental Educ. (low)	-0.74^{***}	-0.75^{***}	-0.98^{***}	-0.44^{*}
	(0.11)	(0.14)	(0.19)	(0.18)
Parental Educ. (lowest)	-0.32^{*}	-0.20	-1.33^{***}	0.35
	(0.13)	(0.15)	(0.18)	(0.26)
Female	0.26^{***}	0.67^{***}	-0.91^{***}	0.70^{***}
	(0.04)	(0.05)	(0.07)	(0.06)
Prior Perf. (middle)	0.55^{***}	0.60^{***}	0.36^{***}	0.64^{***}
	(0.05)	(0.07)	(0.09)	(0.08)
Prior Perf. (bottom)	1.06^{***}	1.06^{***}	0.82^{***}	1.20^{***}
	(0.06)	(0.07)	(0.10)	(0.10)
Grade 5	-0.01	-0.10	-0.32	0.31
	(0.12)	(0.14)	(0.18)	(0.23)
Grade 6	0.02	0.01	0.18	0.01
	(0.11)	(0.14)	(0.17)	(0.19)
Grade 7	0.08	-0.06	0.69^{***}	-0.07
	(0.13)	(0.17)	(0.19)	(0.22)
Year (std.)	0.69^{***}	0.21^{**}	0.82^{***}	0.99^{***}
	(0.06)	(0.07)	(0.10)	(0.09)
Days between tests (std.)	0.48^{***}	0.55^{***}	0.21^{**}	0.64^{***}
	(0.05)	(0.05)	(0.07)	(0.09)
(Intercept)	0.07	-0.23	1.21^{***}	-0.39
	(0.11)	(0.14)	(0.15)	(0.20)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	358407	352656	272382	343151
RMSE	11.03	14.85	18.77	16.99
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S6. Main effects, complete subject scores only.

	Composite	Maths	Reading	Spelling
Treatment	-3.05^{***}	-3.05^{***}	-3.38^{***}	-2.73^{***}
	(0.17)	(0.22)	(0.22)	(0.27)
Year (std.)	0.68^{***}	0.35^{***}	0.96^{***}	0.73^{***}
	(0.07)	(0.09)	(0.09)	(0.11)
Days between tests (std.)	0.49^{***}	0.56^{***}	0.24^{***}	0.68^{***}
	(0.05)	(0.06)	(0.07)	(0.10)
(Intercept)	0.84^{***}	0.54^{***}	1.12^{***}	0.87^{***}
	(0.06)	(0.07)	(0.08)	(0.12)
R^2	0.01	0.00	0.00	0.00
Adj. R^2	0.01	0.00	0.00	0.00
Num. obs.	259206	259206	259206	259206
RMSE	10.53	14.94	18.79	16.99
N Clusters	929	929	929	929

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S7. Main effects in near-complete schools.

	Composite	Maths	Reading	Spelling
Treatment	-3.18^{***}	-2.98^{***}	-3.26^{***}	-3.13^{***}
	(0.19)	(0.23)	(0.25)	(0.27)
Year (std.)	0.73^{***}	0.21^{*}	0.88^{***}	1.12^{***}
	(0.08)	(0.09)	(0.12)	(0.12)
Days between tests (std.)	0.39^{***}	0.51^{***}	0.14	0.47^{***}
	(0.06)	(0.07)	(0.09)	(0.11)
(Intercept)	0.82^{***}	0.62^{***}	1.23^{***}	0.77^{***}
	(0.08)	(0.08)	(0.10)	(0.15)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	248346	244227	190227	235817
RMSE	11.06	14.88	18.79	16.97
N Clusters	594	594	591	593

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S8. Social inequality in near-complete schools.

	Composite	Maths	Reading	Spelling
Treatment	-3.07^{***}	-2.87^{***}	-3.20^{***}	-3.00^{***}
	(0.18)	(0.23)	(0.25)	(0.27)
Treat x Par. Educ. (low)	-1.33^{***}	-1.10^{*}	-1.18^{*}	-1.84^{***}
	(0.31)	(0.43)	(0.53)	(0.48)
Treat x Par. Educ. (lowest)	-1.42^{***}	-1.86^{***}	-0.40	-1.74^{***}
	(0.35)	(0.46)	(0.47)	(0.51)
Parental Educ. (low)	-0.20	-0.28	-0.53^{*}	0.20
	(0.15)	(0.22)	(0.27)	(0.26)
Parental Educ. (lowest)	0.40^{*}	0.35	-0.89^{***}	1.42^{***}
	(0.19)	(0.21)	(0.26)	(0.33)
Year (std.)	0.73^{***}	0.21^{*}	0.88^{***}	1.12^{***}
	(0.08)	(0.09)	(0.12)	(0.12)
Days between tests (std.)	0.39^{***}	0.51^{***}	0.14	0.47^{***}
	(0.06)	(0.07)	(0.09)	(0.11)
(Intercept)	0.81^{***}	0.62^{***}	1.29^{***}	0.70^{***}
	(0.08)	(0.09)	(0.10)	(0.15)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	248346	244227	190227	235817
RMSE	11.06	14.88	18.79	16.97
N Clusters	594	594	591	593

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S9. Main effects with school fixed effects.

	Composite	Maths	Reading	Spelling
Treatment	-3.17^{***}	-2.92^{***}	-3.29^{***}	-2.93^{***}
	(0.17)	(0.20)	(0.22)	(0.24)
Year (std.)	0.71^{***}	0.22^{**}	0.94^{***}	0.98^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.36^{***}	0.43^{***}	0.04	0.52^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
School fixed effects	✓	✓	✓	✓
R^2	0.03	0.02	0.01	0.03
Adj. R^2	0.02	0.01	0.01	0.03
Num. obs.	358407	352656	272382	343151
RMSE	10.94	14.79	18.71	16.78
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S10. Social inequality with school fixed effects.

	Composite	Maths	Reading	Spelling
Treatment	-3.07^{***}	-2.81^{***}	-3.23^{***}	-2.79^{***}
	(0.17)	(0.20)	(0.23)	(0.24)
Treat x Par. Educ. (low)	-1.30^{***}	-1.06^{**}	-0.95^{*}	-1.93^{***}
	(0.28)	(0.38)	(0.48)	(0.45)
Treat x Par. Educ. (lowest)	-1.33^{***}	-1.76^{***}	-0.49	-1.68^{***}
	(0.32)	(0.41)	(0.43)	(0.48)
Parental Educ. (low)	-0.10	-0.34^{*}	-0.20	0.29
	(0.10)	(0.15)	(0.20)	(0.17)
Parental Educ. (lowest)	0.22^{*}	0.16	-0.47^{*}	0.78^{***}
	(0.11)	(0.15)	(0.20)	(0.17)
Year (std.)	0.71^{***}	0.22^{**}	0.94^{***}	0.98^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
Days between tests (std.)	0.36^{***}	0.43^{***}	0.05	0.52^{***}
	(0.06)	(0.07)	(0.09)	(0.09)
School fixed effects	✓	✓	✓	✓
R^2	0.03	0.02	0.01	0.03
Adj. R^2	0.02	0.01	0.01	0.03
Num. obs.	358407	352656	272382	343151
RMSE	10.94	14.79	18.70	16.78
N Clusters	937	937	930	936

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S11. Main effects with family fixed effects.

	Composite	Maths	Reading	Spelling
Treatment	-3.26^{***}	-3.03^{***}	-3.14^{***}	-3.19^{***}
	(0.11)	(0.16)	(0.22)	(0.18)
Year (std.)	0.70^{***}	0.16^{*}	0.94^{***}	1.07^{***}
	(0.05)	(0.07)	(0.10)	(0.07)
Days between tests (std.)	0.35^{***}	0.35^{***}	0.06	0.62^{***}
	(0.05)	(0.06)	(0.09)	(0.07)
Family fixed effects	✓	✓	✓	✓
Num. obs.	145363	143140	114290	139595
R^2	0.24	0.23	0.27	0.26
Adj. R^2	0.04	0.03	0.02	0.05
Num. groups: family	30490	30467	29436	30394

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S12. Social inequality with family fixed effects.

	Composite	Maths	Reading	Spelling
Treatment	-3.16^{***}	-2.93^{***}	-3.12^{***}	-3.02^{***}
	(0.12)	(0.16)	(0.23)	(0.19)
Treat x Par. Educ. (low)	-1.34^{**}	-0.79	-0.61	-2.50^{***}
	(0.45)	(0.63)	(0.89)	(0.76)
Treat x Par. Educ. (lowest)	-1.74^{***}	-2.10^{***}	-0.06	-2.97^{***}
	(0.42)	(0.58)	(0.83)	(0.72)
Year (std.)	0.70^{***}	0.16^{*}	0.94^{***}	1.07^{***}
	(0.05)	(0.07)	(0.10)	(0.07)
Days between tests (std.)	0.35^{***}	0.35^{***}	0.06	0.61^{***}
	(0.05)	(0.06)	(0.09)	(0.07)
Family fixed effects	✓	✓	✓	✓
Num. obs.	145363	143140	114290	139595
R^2	0.24	0.23	0.27	0.26
Adj. R^2	0.04	0.03	0.02	0.05
Num. groups: family	30490	30467	29436	30394

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$