Technical appendix

Additional analyses and background information for Het basisonderwijs 1 jaar na het begin van Corona, wat was de impact op toetsresultaten van de kernvakken?

A. Frey[†], R. van der Leeuw[‡], & M.D. Verhagen[†]

[†]University of Oxford, [‡] Leeruniek

June, 2021

This technical appendix has been created to support the analyses contained in 'Het basisonderwijs 1 jaar na het begin van Corona, wat was de impact op toetsresultaten van de kernvakken?' and is largely based on the Supplementary Information that was part of the study "Learning loss due to school closures during the COVID-19 pandemic", published in the *Proceedings of the National Academy of Sciences of the United States of America*. The original document was amended in parts to reflect updates to the most recent analysis, but otherwise follows the Supplementary Information where possible.

Contents

List of Figures

List of Tables

1 Study context

Education in the Netherlands is based on a common school up to the age of 12, after which students are placed on separate tracks (1; 2). Schooling is compulsory from age 5 to 16, but the majority of children start at the age of 4 (1). The Dutch system combines a high degree of school autonomy with a centralized system for school funding and accountability (1; 2; 3). The system ultimately dates back to the early 20th century, and arose as a compromise to give schools equal access to state funding regardless of denomination (3; 4). To this day, most schools are denominational, predominantly Roman Catholic or Protestant. Schools are run by local school boards, and the right to establish a school is enshrined in the constitution, once certain basic criteria are met (1). All schools are funded by the Ministry of Education, Culture and Science ("Ministry of Education" henceforth), with schools that serve disadvantaged students receiving a larger budget per capita (5). The Dutch system achieves a high degree of both efficiency and equity as measured by performance on international student assessments (2). The Netherlands, while close to the OECD average in school spending and in Reading performance, places among Europe's top performers in Maths (6).

During the first wave of the COVID-19 pandemic, the government pursued a so-called "intelligent lockdown," relying on voluntary cooperation and allowing ordinary life to continue as far as possible (7; 8; 9). School closures were one of few strictly enforced non-pharmaceutical interventions. However, their duration were short compared to most other OECD countries. As shown in Fig. 1, schools closed on March 16 and reopened eight weeks later, on May 11. While students initially attended classes every other day, in-person schooling returned to normal activity from June 8, before closing once more between January and mid February. Arguably, the Netherlands was unusually well prepared for remote learning: the country leads the world in broadband penetration (10; 11) with more than 90% of households enjoying broadband access even among the poorest quartile (12). Adding to this advantage, the response of national and local governments was swift: in March 2020, the Ministry of Education devoted 2.5 million euros to online learning devices for students in need (13), and this scheme was extended

with another 3.8 million in June (14), with similar initiatives at a local level. Towards the end of the second lockdown, the Ministry of Education announced a 8.5 billion euro stimulus package to assist schools in closing learning gaps. Approximately 5.8 billion euro was earmarked for primary and secondary education.

Despite these efforts, anecdotal evidence suggests that primary school teachers had limited prior experience of or preparation for distance learning, asking much in terms of flexibilty and adaptive power from both schools and teachers. In contrast to older students who can be expected to shoulder some of the responsibility for their study themselves, primary school study is more dependent on continuous instruction from a teacher. Being deprived of classroom instruction meant that the responsibility for structuring the school day and creating a supportive work environment at least in part fell on parents and household support functions as well. Many teachers created instruction packets with physical assignments and handouts that parents had to collect from schools during the first shutdown. There is limited data on how much instruction actually took place online, and how many hours of effective school work students were able to achieve. However, evidence from Germany suggests that students reduced their study time by as much as half (15). Survey evidence from the Netherlands also indicates that there were considerable disparities in help with schoolwork and learning resources (16), and high levels of dissatisfaction with remote learning (7). This evidence mirrors that from several other countries, showing important disparities in children's conditions for learning from home (17; 18; 19; 20). Although research on the matter is still missing, anecdotal evidence suggests schools and teachers were better prepared to handle the second shutdown.

2 Data sources

Three features of the Dutch education system make this study possible. The first is the student monitoring system, LVS, which provides our test score data. This system comprises a series of mandatory tests that are taken twice a year throughout a child's

primary school education (age 6–12). The second feature is the weighted system for school funding, which until recently obliged schools to collect information on the family background of all students. Third is the fact that some schools rely on third-party service providers to curate data and provide analytical insights. It is not uncommon that such providers generate anonymized datasets for research purposes. We teamed up with the Mirror Foundation (`https://www.mirrorfoundation.org/`), an independent research foundation associated with one such service provider, who gave us access to a fully anonymized dataset of students' test scores. In the following we describe the student monitoring system, the student background data, and our data partner.

2.1 Student monitoring system (LVS)

The measures of student performance that we use are gathered from the student monitoring system or *leerlingvolgsysteem* (LVS), which is a distinguishing feature of Dutch primary education (21). The LVS is one of several components introduced to uphold quality and accountability despite the country's high degree of school autonomy (2). Diagnostic tests are administered to all students twice a year, normally in the middle of the school year in January/February and at the end of the school year in June. By continuously assessing students and tracking their performance longitudinally, the system helps educators tailor their instruction to the needs of a particular cohort and identify students in demand of extra support. The LVS was first developed by the National Institute for Educational Measurement, CITO, in the 1990s. CITO was originally founded as a non-profit organization in the 1960s, but is today a commercial enterprise with several international branches. In the Netherlands, CITO testing services are developed and sold on a "semicommercial" basis (21), which means that the Ministry of Education serves as the main funder of CITO and appoints its board director. Schools decide on whether to purchase their service using education funds that are public. Since 2014, it is mandatory for all primary schools to use an LVS, with CITO being the leading provider holding a large majority of the market share.

2.2 Student background data

Data on student background are collected by schools as part of the national system of weighted student funding or *gewichtenregeling*. Primary education in the Netherlands is operated as a voucher system, where funding is provided to schools by the Ministry of Education on a per-student basis (5; 2). Since 1985, an additional contribution toward each student depends on their social background in an effort to reduce social inequality and raise bottom performance. The amount of funding that a school gets is proportional to the socioeconomic composition of the student body, with schools with a higher proportion of disadvantaged students receiving more funds per student. To support this system, schools are legally required to collect data on parental background when a student first starts school or transfers between schools. The number of indicators used to determine school funding has changed throughout the history of the system, but between 2006 and 2019, parental education was the sole indicator (22; 23). In 2019, responsibility for determining funding weights was transferred to the central government, using a wider set of indicators and administrative data stored by Statistics Netherlands. As this information is only available at a school level, our main analysis relies on both the individual-level data on parental education collected by schools, as well as the newer indicator of student disadvantage derived from administrative data.

2.3 Data partner (Mirror Foundation)

To access test scores and student background data, we entered a partnership with the Mirror Foundation (`https://www.mirrorfoundation.org/`), an independent non-profit research foundation set up to support educational research initiatives. The Mirror Foundation enabled us to access a fully anonymized dataset of 15% of primary schools in the Netherlands. The schools are users of a data analytics platform that provides school boards with timely insights based on LVS and other data that are kept by the schools. All schools in the Netherlands are mandated to use a digital interface for student monitoring, and some also subscribe to services offering more extensive functionality and

independent analysis—as is the case for the schools in our sample. The dataset was generated by anonymizing existing school records from the schools' LVS and done at the schools' instruction, whereby the latter act as 'data controller' in the definition of the EU's General Data Protection Regulation (GDPR). Anonymization was done with the explicit and stated objective of supporting academic research. All analysis was carried out in accordance with the GDPR and at no point did the authors have access to data that would allow the identification of individuals.

3 Variables

3.1 Outcomes

3.1.1 Curricular tests

Achievement is measured via the LVS system using performance on standardized tests developed by CITO. Tests are taken across three main subject areas: Maths, Spelling, and Reading, the first two of which are mandatory. Each test lasts up to one hour per subject. Maths comprises abstract problems involving the four arithmetic operations—addition, subtraction, multiplication, division—as well as applied problems based on concrete tasks. The applied tasks evaluate the student's facility with concepts such as time or currency. In Spelling, a series of words is presented verbally and the student demonstrates that he or she has mastered the spelling rules by writing the words down correctly. Reading assesses the student's ability to understand written texts, including both factual and literary content. The student is presented with a series of texts and at the end of each, he or she gets to answer a set of multiple-choice questions. All tests are psychometrically validated by CITO and translated to national performance benchmarks expressed as percentile scores for a given grade and test (24; 25; 26). However, as the translation keys provided by the test producer are actually based on smaller samples than that at our disposal, we further re-norm the distribution within our sample. That is, we pool results across all study years and impose a uniform distribution separately

by subject, grade, and testing occasion: mid-year vs end-of-year. Our main outcome is a composite score that takes the average of non-missing percentile scores across the three subject areas. We also display performance on each separate test and, in supplementary analyses in Section 7.1, require a student to have valid scores in all three subjects. The reliability of these tests is excellent (**?**).

3.2 Covariates

3.2.1 School grade

Schooling in the Netherlands is mandatory from age 5, but the first three grades feature limited didactic material and are comparable to kindergarten (2). Since we have to compare students performance during the middle-of-the-year test in 2021 to their performance in 2020, prior to the pandemic, we require students to have been tested across two academic years. Our analysis therefore follows students from grade 5 and until the penultimate grade of primary schooling, grade 7. The final grade 8 is dedicated to transitioning to secondary education and is shorter than the other grades. The designation of grades differs from international standards, where the ages we study would correspond to grades 1–4 of elementary school. To avoid confusion with international standards we choose to label grades by the modal age of students in the latter of the two grades, which correspond to the ages 9, 10, and 11.

3.2.2 Parental education

Information on parental education is collected from parents by schools as part of the weighted student funding system (5). The classification is therefore the one designated by the Ministry of Education to determine school funding weights. The variable takes on three values: *high* if at least one parent has a degree above lower secondary education; *low* if both parents have a degree above primary education but neither has one above lower secondary; and *lowest* if at least one parent has no degree above primary education and neither has a degree above lower secondary. The three groups make up, respectively,

92%, 4%, and 4% of the student body and our sample (Fig. 2). The school funding weights based on surveys of parental education were replaced by a new system based on administrative data in 2019 (Section 2.2). Nevertheless, the earlier information collected by schools remains available and we rely on it in our main analysis for two reasons. First, the new funding weights are only made available at a school level and therefore do not allow us to distinguish the socioeconomic background of individual students. Secondly, survey data on education are likely to be superior in some respects, especially for immigrant parents whose credentials often do not register in official statistics.

3.2.3 Prior performance

To assess prior performance, we take all available tests from the previous year and calculate a percentile score similarly to our main outcome measures. We then create a categorical variable by calculating a student's average rank across all non-missing values and splitting the variable into three equal-sized groups. By basing this information on data collected in the previous year, we avoid the mechanical correlation that would obtain if prior performance had been measured at baseline in the same year as we assess student progress. Doing so is known to introduce regression to the mean which can lead to various statistical artifacts (27).

3.2.4 Immigrant background

In the Netherlands today, immigrant minorities make up a significant share of the student body (2). Unfortunately we lack an individual-level indicator of immigrant background. Instead, we measure the proportion of non-Western inhabitants in a school's neighborhood using administrative data. 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. Although this measure reflects the composition of the neighborhood rather than the student body, the two are likely to be correlated given that residential proximity is one of the most important determinants of school choice in the Netherlands (28; 29).

3.2.5 School disadvantage

In recent years, there has been increasing debate about the reliance on parental education as the sole indicator of socioeconomic disadvantage and determinant of school funding weights in the Dutch system (22; 23). Following a prolonged investigation, the practice was therefore replaced in 2019 by one where the Ministry of Education determines school funding with the help of administrative data held by Statistics Netherlands (30). The factors considered in the new measure include the educational level of both the mother and the father as before; but also the country of origin of the parents, the duration of the mother's residence in the Netherlands, and whether parents have taken part in debt restructuring (*schuldsanering*) (31). We use both measures in the main analysis, and report specific results per school weight in Section 7.4.

3.2.6 Sex

This information is collected by the schools in conjunction with parental education and is available from school records.

4 Analytical strategy

4.1 Differences in analytical strategy compared to earlier work

As discussed in the main text, our key interest is to examine the effect of both school shutdowns, as well as dedicated efforts to mitigate learning losses by teachers, schools and the government, on student performance on standardized tests. For an extensive discussions of the results based on the first shutdown, we refer the reader to the original paper (**?**). As highlighted in the current study, we encountered additional complexities when evaluating learning losses over the entire period of the pandemic, spanning both shutdowns. Firstly, the mid-year test in 2021 was severely delayed as a result of the second school shutdown only ending in mid-February. This left students with more time to prepare for the tests, complicating the selection of an appropriate control group.

Secondly, the current study spans over an entire calendar year, but over two school years. Because of this, we are comparing student test scores across multiple grades, and thus have less unique years and grade groups at our disposal. We discuss these additional complexities in kind below.

4.2 Identification strategy

Estimating the effect of school closures on student achievement raises several challenges. A naive approach would be to compare average national test scores following school closures to average national test scores in a previous year. However, this ignores the considerable fluctuation in performance that can occur due to changes in student composition or other factors from one year to the next. It is therefore vital that achievement measures are collected both before and during the COVID-19 pandemic, so that progress in this period can be compared to progress during the same period in previous years. Still, if not all students are tested during the pandemic, differences in the composition of test takers from the earlier to the latter test may bias estimates. In our analysis, we only include students who take both the mid-year test in 2020, before schools closed, and the mid-year test in 2021, one year into the pandemic. This is, in effect, a differences-in-differences design (32).

$$\Delta y_i = \alpha + \delta T_i + \epsilon_{ij}, \tag{1}$$

where $\Delta y_i = y_i^{year=t} - y_i^{year=t+1}$ is an individual student's relative movement in the achievement ranking from the initial mid-year test prior to the onset of the pandemic to the subsequent mid-year test in the following year, T_i is an indicator for the treatment year 2020/2021, and ϵ_{ij} is an i.i.d. error term clustered at the school level. The coefficient δ thus captures overall learning loss since the start of the pandemic. This specification deals with the fact that the composition of test takers may differ between both years by ensuring that only students present at both occasions contribute to the estimation. Compared to the analyses done in (**?**), the yearly trend and days-between-tests variables

have been omitted from the main specification. This is because we require information across two grades to generate our outcome variable of interest since it spans two grades, depleting the number of available years. More importantly, the support of the testing dates for the mid-year test in 2021 does not coincide with the support in previous years—see the extensive discussion in the main text. This means that any estimate of the days-between-tests coefficient will be based on incomparable domains between treatment and control —i.e. the variable will be considerably larger in the treatment group than the control group. For these reasons, and as discussed in the main text, we opted for a pure estimate of δ without the baseline controls included in (**?**). In 7.2 we evaluate whether we find similar results applying the above approach to the 12-month period including only the first shutdown (end-of-year 2019 to end-of-year 2020).

To deal with differences in the composition of students between treatment and comparison years, we pursue several strategies. The first is simply to include a set of student characteristics $\mathbf{X}_i$. We first use this setup including one variable at a time to assess heterogeneity in the treatment effect, interacting each student characteristic X_i with the treatment indicator T_i:

$$\Delta y_i = \alpha + \beta X_i + \delta_0 T_i + \delta_1 T_i X_i + \epsilon_{ij}, \tag{2}$$

where X_i is one of: parental education, student sex, or prior performance. We also estimate separate models for each school grade. In the next step, we add all student covariates jointly as control variables:

$$\Delta y_i = \alpha + \mathbf{X}_i'\beta + \delta T_i + \epsilon_{ij}, \tag{3}$$

where $\mathbf{X}_i$ is a vector containing parental education, student sex, and prior performance. Our dataset includes not only student covariates but also school characteristics, and potential interactions between variables. A flexible way to adjust for high-dimensional variation is through weighting schemes that ensure that characteristics are balanced between comparison and treatment group. Rosenbaum and Rubin (33) show

how a large set of potential confounders can be reduced to a single propensity score, capturing the conditional probability of treatment. This approach proceeds in two steps: first by estimating the probability of treatment conditional on all observed covariates, and second by estimating the main outcome equation while balancing on the propensity score. The propensity score $\hat{p}(\mathbf{X}_i)$ is estimated from a logistic regression of the treatment indicator on the set of covariates. It is possible to incorporate it in several ways but we use it to construct a set of regression weights (34):

$$\mathbf{E}[\Delta y(0) \widehat{\mid} T = 1] = \frac{\sum_{\{i|T=0\}} \Delta y_i v_i}{\sum_{\{i|T=0\}} v_i}, \tag{4}$$

where the weight v_i of each observation is related to the propensity score through the equation $v_i = \frac{\hat{p}(\mathbf{X}_i)}{1-\hat{p}(\mathbf{X}_i)}$. In this balancing procedure, we adjust for a large set of covariates including interaction terms between all individual variables as well as school disadvantage, ethnic composition, and school denomination (Section 7.5). Across these models, we adjust for compositional differences using balancing weights while including the vector $\mathbf{Z}_i$ for testing year and date as standard regression controls.

Both approaches—regression adjustment and propensity score weighting—represent different ways of achieving balance on observed covariates but are vulnerable to unobserved sources of heterogeneity. In additional analyses, we make use of the fact that students are nested within schools to estimate a fixed-effects design (35). This allows us to adjust for any time-invariant confounding at the school level, whether due to observed or unobserved sources of heterogeneity. The fixed-effects design can be written:

$$\Delta y_i = \sum_{j=1}^{J} \alpha_j J_{ij} + \mathbf{X}_i' \beta + \delta T_i + \epsilon_{ij}, \tag{5}$$

where $J_{ij} = \mathbf{1}_{J_i=j}$ is a binary indicator equal to one if unit i belongs to cluster j, and zero otherwise. In our analysis the J_{ij} group identifiers are school level indicators grouping all students within the same school.

We estimate all models in the R statistical computing environment using packages listed in Section 4.6 below.

4.3 Effect size conversions

4.3.1 Percentiles and standardized effects

Our effect sizes are expressed on the scale of percentiles. In educational research it is common to use standard-deviation based metrics such as Cohen's d (36):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_p}, \tag{6}$$

where $\bar{x}_1 - \bar{x}_2$ is the difference in means between treatment and comparison groups and σ_p is the pooled standard deviation. To convert between treatment effects on the percentile scale and standardized effects, we rely on the lesser known U_3 metric also proposed by Cohen (37), which describes the overlap between two distributions. Specifically, U_3 is defined as the proportion of the comparison group exceeded by the upper half of cases in the treatment group. Conversion between U_3 and d can be done with the following equation:

$$d = \Phi^{-1}(U_3), \tag{7}$$

where Φ^{-1} is the inverse cumulative standard normal distribution. While this conversion applies to the normal case, U_3 is defined such that it is invariant to any rank-preserving transformation. Hence, we can apply the same conversion to our percentile scores under the assumption that they came from an underlying normal distribution.

For two identical distributions with no difference in means, the upper half of cases in the treatment group will exceed exactly half of the cases in the comparison group. In this case, $U_3 = 0.50$ and $d = \Phi^{-1}(0.50) = 0$. A difference of -3 percentiles in the treatment group vs the comparison group implies that $U_3 = 0.50 - 0.03 = 0.47$. Hence, the standardized effect size equivalent becomes $d = \Phi^{-1}(0.47) = -0.075$. More generally, with "small" or "medium" effect sizes in the range $d \in [-0.5, 0.5]$, Cohen's U_3 implies a conversion factor of 0.025 standard deviations per percentile.

4.4 Addressing differences in testing moment

As has been discussed extensively in the main text, the support of testing dates for the end-of-year test 2021 does not coincide with the support in earlier years. In effect, students had on average six weeks of additional preparation time prior to taking the test during the treatment year than the control group (see Figure 6 in the main text). There is very little guidance to generate assumptions on such an additional time window, since most existing work exploits interruptions to education on a much smaller scale. To remain consistent with our earlier calculations, we assume a weekly percentile gain of approximately 0.37 pp / week which would imply an expected improvement in test scores of 2.0 pp. We note however that this estimate is conditional, among others, on choices made by educators at the time. Estimates by the Worldbank, OECD and other work suggest the range to fall between 0.3 and 0.4 pp (**?** 38).

4.5 Group comparisons

We make the following assumption when reporting on de-meaned treatment effects (Figure 9 in the main text): that the positive effect of extra time is the same across students but might differ per subject. Under this assumption, the difference between treatment effect for subsets of the population can be compared to the overall treatment effect to give an indication of the difference in learning loss between groups. For instance, if an overall treatment effect of -1pp was observed for all students, and an effect of -2pp was observed for girls and +0pp for boys, under the mean effect assumption the true difference between the two groups would still be 2 pp. We provide these estimates such that policymakers can get an indication of the group differences over the total period compared to those found for the first shutdown. We do not provide formal statistical tests of the difference in group differences.

4.6 Statistical software

Analyses were done in `R` (version 4.0.3). We used the `estimatr` package to cluster standard errors at the school level, as well as to include school and family fixed effects. We used the `lme4` package for the multilevel analyses. To adjust our sample using matching and weighting techniques, we relied on the `WeightIt` and `cobalt` packages. We are thankful to the `R` community, in particular to the `tidyverse`, `broom`, and `lubridate` data wrangling libraries, and the `data.table` library that helped us greatly speed up data processing. All computations were done on a machine running Mac OSX 10.17.7.

- `Estimatr`: `https://cran.r-project.org/web/packages/estimatr/estimatr.pdf`
- `lme4`: `https://cran.r-project.org/web/packages/lme4/lme4.pdf`
- `WeightIt`: `https://cran.r-project.org/web/packages/WeightIt/WeightIt.pdf`
- `Cobalt`: `https://cran.r-project.org/web/packages/cobalt/cobalt.pdf`
- `Tidyverse`: `https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf`
- `broom`: `https://cran.r-project.org/web/packages/broom/broom.pdf`
- `lubridate`: `https://cran.r-project.org/web/packages/lubridate/lubridate.pdf`
- `data.table`: `https://cran.r-project.org/web/packages/data.table/data.table.pdf`

5 Quality control

5.1 Representativeness

We obtained access to the data through the Mirror Foundation, and an educational analytics provider with which the foundation collaborates (Section 2.3). Selection into the sample is thus mediated through a school's use of the educational analytics service. There might be concerns whether this selection is non-random, and hence how well our sample represents the universe of schools in the Netherlands. In Fig. 2 we evaluate this question by inspecting the distribution of observable characteristics in our sample and the population. The sample distribution mirrors the population on most observables: school size, denomination, urbanity, parental education, school disadvantage, and neighbourhood composition, with only very minor differences.[1] Importantly, the relative representation of school type (e.g. public or Christian) is near identical to that in the population, as is the distribution of parental education within schools. Schools in our sample are also close to the population distribution on the newer composite indicator of school disadvantage (Section 2.2 and 3.2).

This leaves the possibility that our sample is selected on unobserved characteristics. In particular, it is possible that adopters of the analytics service are especially invested in digital infrastructure and data-based accountability. Although all schools are mandated to use a digital interface for student monitoring, many existing solutions are simple data management tools with limited analytical functionality. To the extent that the schools in our sample are more invested in digital infrastructure, they may have been better equipped to cope with online learning which could lead us to underestimate the impact of pandemic school closures. Another consideration is ability to pay, given that the platform service is offered on a paid subscription basis. However, the cost of the service is minor relative to a school budget: 1,500 euro annually, which corresponds to 3% of a single teacher's salary (50,000 euro) or less than 0.1% of a typical school budget (2 million euro). As the main determinant of school funding until 2019 was the parental

[1] The most visible difference for neighbourhood composition is due to us only having access to categorical information on schools' share of foreign residents.

education of the student body, the fact that this does not differ markedly between the sample and the population (Fig. 2) corroborates that economic considerations are not a major determinant of service uptake.

5.2 Missing data

We define our analytical sample as all observations in the relevant grades in our database for which there is valid information on sex and parental education. Table 1 shows missing data on test scores and ability scores across all class years for the two school years used in the analysis, 2019–2020 and 2020–2021. About 6% lack performance data from the previous year for both school years. There is also considerable difference in missingness between subjects, with missing values being higher in Reading than in either Maths or Spelling. This is due to the regulation of the student monitoring system, which only requires schools to report student achievement in Maths and Spelling (Section 3.1). In general, missingness across test scores is higher in the treatment year (2020–2021), but considerably lower than during the end-of-year test in 2020, on which the initial analysis was based. One year into the pandemic, most students are being tested, even if those tests occur later in time than during previous years (see Figure 6 in the main text). Still, this missingness will bias our estimates if it is selected on the outcome variable, that is, if only those students who tend to over- or underperform from one year to the next are being tested. Reassuringly, Table 2 shows that the missing data is almost perfectly balanced by prior performance: the composition of top, middle, and bottom performers is similar in comparison and treatment years.

6 Descriptive statistics

Table 2 shows summary statistics for the full sample broken up by comparison and treatment groups. Because of the large sample size, most differences are statistically significant even when they are quantitatively unimportant. There are no substantive differences between comparison and treatment groups.

7 Additional results

7.1 Regression tables

In Table 3–7, we display regression results underlying Fig. 7 in the main manuscript, as well as additional analyses by subject and subgroup. Table 3 displays the main effect reported in main manuscript Fig. 7, and separate results by subject domain. Table 4 shows results by parental education for the composite score as reported in main text Fig. 7, and for separate subjects. Table 5 does the same for student sex and Table 6 does so for prior performance. Table 7 displays separate analyses by grade. Tables 8 and 9 show results by school-level disadvantage and neighbourhood composition. In Table 10, we report additional regression results simultaneously controlling for all individual-level covariates: sex, parental education, prior performance. This does little to change the treatment effects, which is unsurprising given that treatment status is largely unrelated to student observables (Table 2). In Table 11 we restrict the sample to only those students with a valid score in all three subjects, again with similar results. This last set of results is presented visually in Fig. 3.

7.2 Consistency of results with previous work

To evaluate the impact of the necessary changes to the research design relative to the original study, we evaluated model (1) using the end-of-year tests in the school year 2019-2020 and compared it to the end-of-year tests in school year 2018-2019, using the same sample as previously. In other words: we applied the same analytical approach where we analyze a 12-month period, but now spanning only the first shutdown. The results are shown in Table **??** and indicate a similar overall effect size of -2.84 percentile points which is slightly lower than that found in the original research when including full controls, but overall is of a similar size. This lends credence to the analytical approach for 12-month periods to identify possible losses.

7.3 Sample comparison to previous work

As discussed extensively in our earlier work, there was considerable attrition in testing directly after the first shutdown but almost none during the second shutdown (Table 1. This means we can evaluate any additional differentials between the sample used at the time and the full sample. As a comparison, we provide the main results of our analysis over the 12-month period using only the sample which included scores for the end-of-year 2020 tests on which prior work was based. Results are presented in Table 13. As can be seen, results are similar albeit slightly higher compared with the full population set. This implies that the sample used in the first analysis might have even provided conservative estimates of learning loss, since that sample did slightly better over the full 12-month period than those who did not participate in the end-of-year 2020 test. Note that in the original paper we addressed non-random attrition which led to larger effect sizes, which is in line with this finding. There might also be reason to believe that additional exposure to testing could have improved the performance over the total period of those students who took the end-of-year test.

7.4 Treatment effect by school weight

While school closures were deployed nationwide, the circumstances surrounding online learning were largely a matter for individual schools to handle. It is therefore likely that the response differed considerably at the school level. Fig. 4 reports estimates from a mixed-effects model that lets the estimated learning loss differ between schools with a different school weight. A school's weight reflects the level of social disadvantage of the student population, with higher weights reflecting higher levels of social disadvantage than lower weights (see Section 3.2. The results reveal considerable variation in effects by school weight, with the least disadvantaged schools performing about 2pp better than schools with a more disadvantaged student population during the treatment year. Note that the effect is particularly pronounced for schools with a very low student weight (around 20) until those with a medium complexity (around 30), after which the negative

association seems to level off.

7.5 Covariate balancing

In Table 10, we report regression results including individual-level control variables. To adjust for a larger set of observables, including school characteristics and potential higher-order interactions, in Fig 5–6 we further implement propensity score weighting (Section 4.2). In these analyses we include the same individual-level covariates as earlier—sex, parental education, prior performance—but also two- and three-way interactions between them, a student's school grade, and school-level covariates: school disadvantage, and neighborhood ethnic composition. Fig. 5 shows that the propensity score weighting method achieves a sample that is balanced on the relevant characteristics. Fig. 6 displays our main results using each weighting method. Regardless of weighting schemes, both estimates of learning loss are highly similar and correspond closely to our main specification as reported in Fig. 7 of the main manuscript.

7.6 School fixed effects

Another way to address selective loss to follow-up is by introducing school-level fixed effects (Section 4.2). This design discards all variation between schools which might have biased our results if, for example, schools that perform worse in previous years are over-represented in the treatment year. Table 14 shows results adding school fixed effects, while Table 15 does so for the interaction by parental education. In Fig. 7 we display the results for these and other subgroups visually. This plot confirms that our other results remain similar, and all qualitative conclusions remain unchanged.

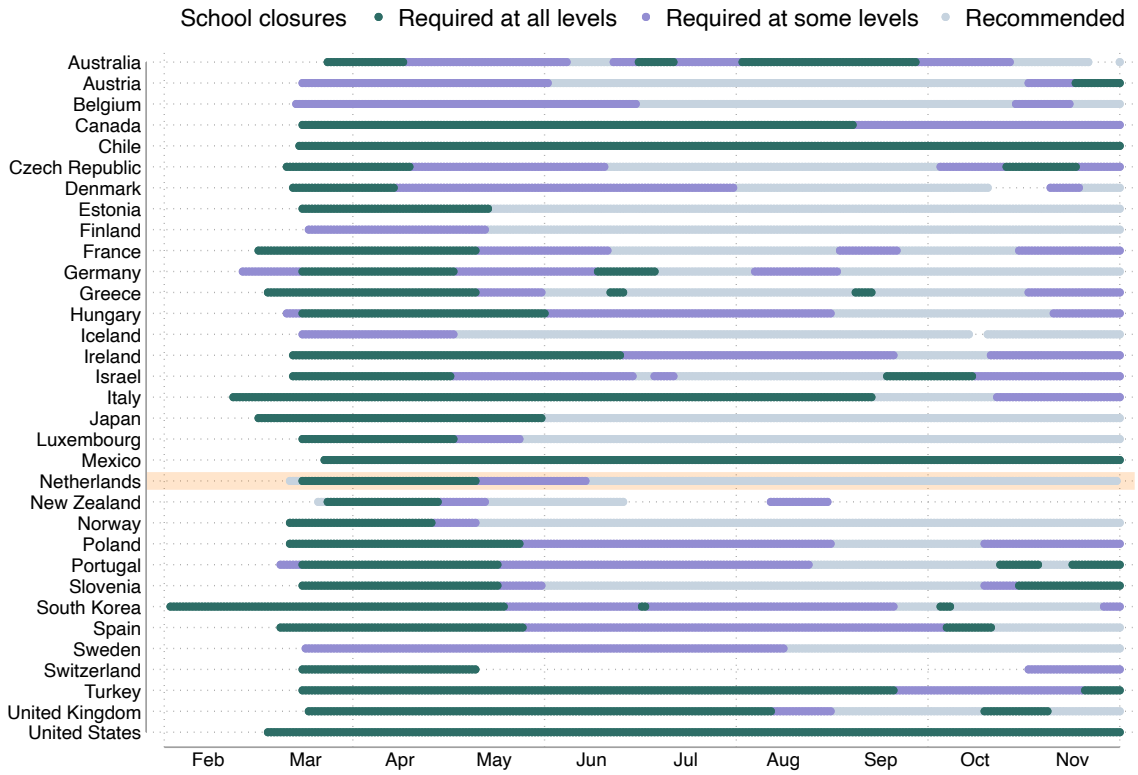


Figure 1: **School closures in the OECD.** The graph shows the onset and duration of school closures in 33 OECD countries through November 2020, with the Netherlands marked in orange. Source: Oxford COVID-19 Government Response Tracker (`https://covidtracker.bsg.ox.ac.uk/`).

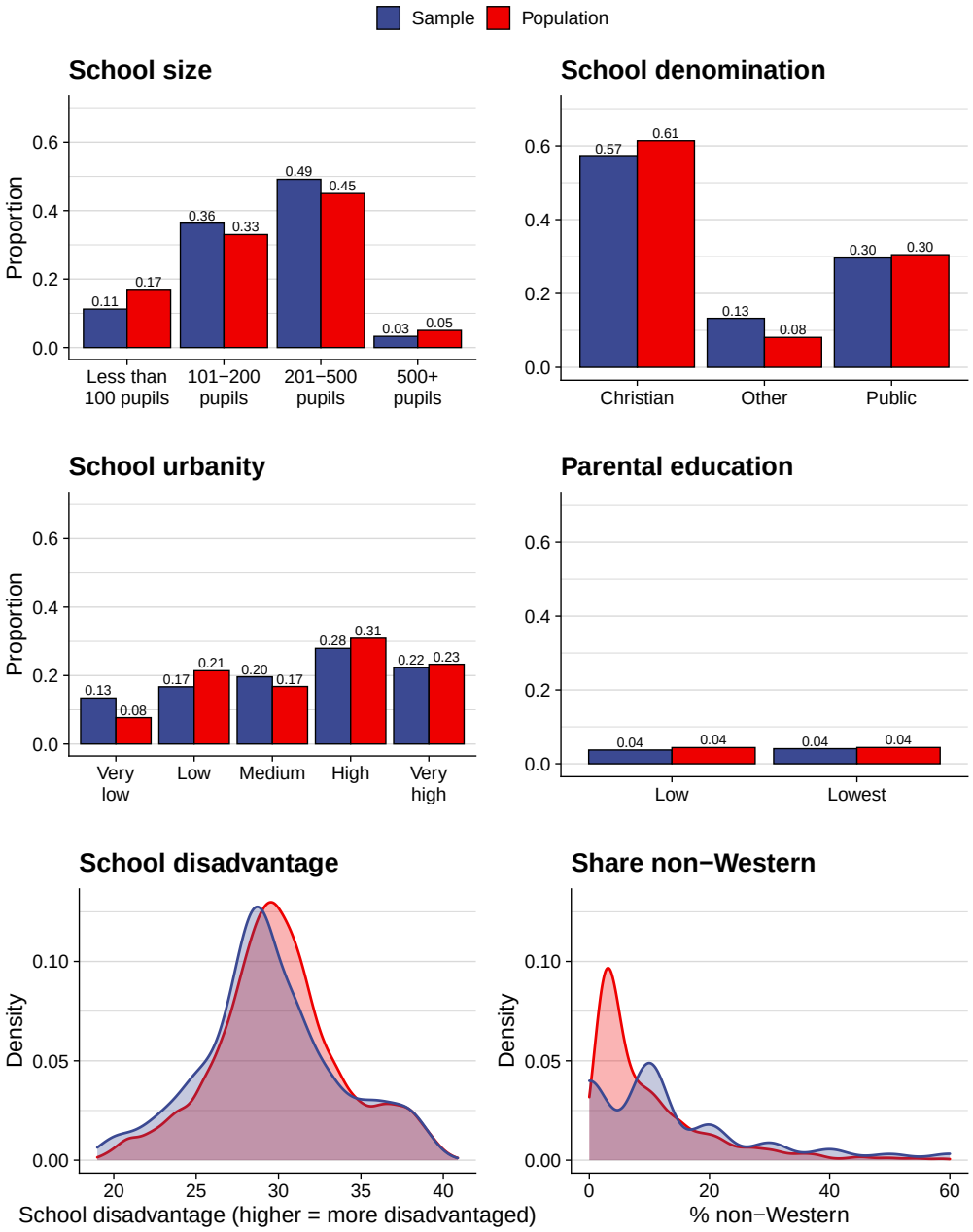


Figure 2: **Representativity of the sample.** The graph compares the distribution of school characteristics in our sample, shown in blue, with that of the universe of primary schools in the Netherlands, shown in red. Source: Onderwijsinspectie (`https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata`), CBS Statline (`https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurten-2004-2019`).

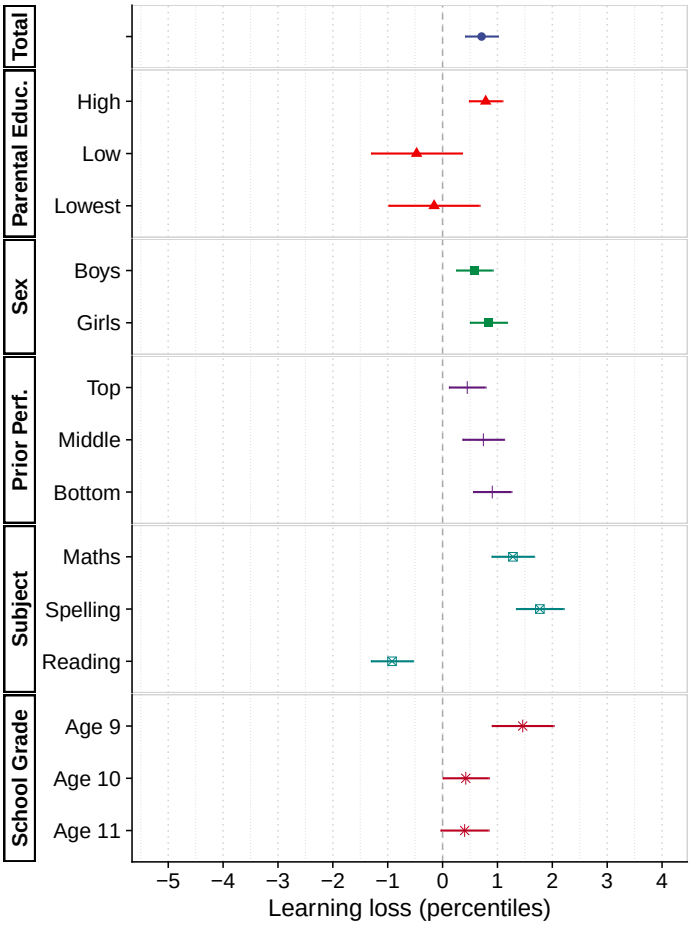
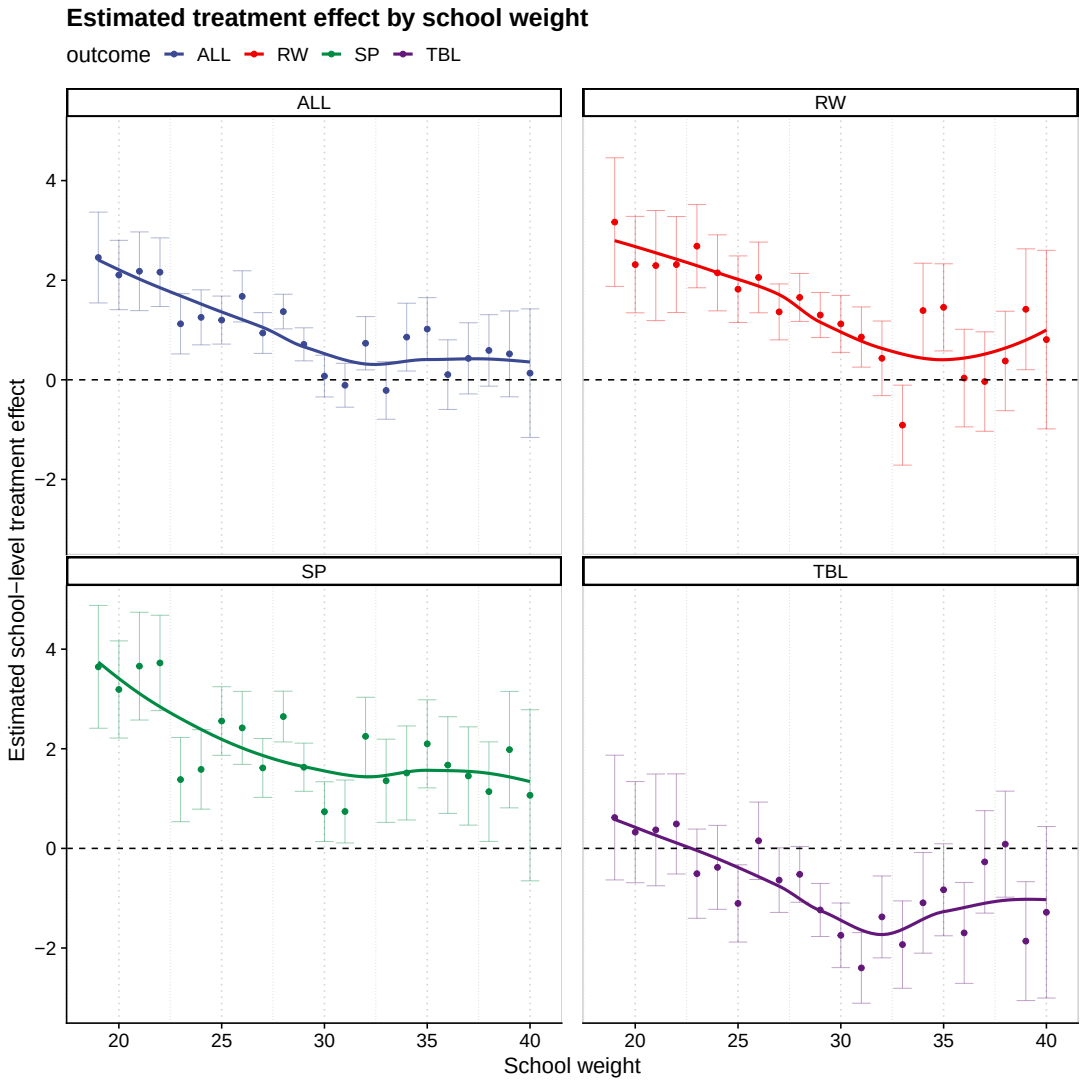


Figure 3: **Results with complete subject scores.** The graph shows results from a specification identical to our main analysis except the sample is restricted to students with complete scores in all subjects.



(a) School-level treatment effects

Figure 4: **School-level effects.** Treatment effect by school-weight level. Low school-weight implies a non-complicated student body. As can be seen, treatment effects steadily decline as the school weight moves from 20 to around 30, after which the treatment effect stabilises. Overall, the difference between the lowest school weight and the highest school weight is between 2-3pp.

Figure 5: **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 the estimated propensity of treatment.

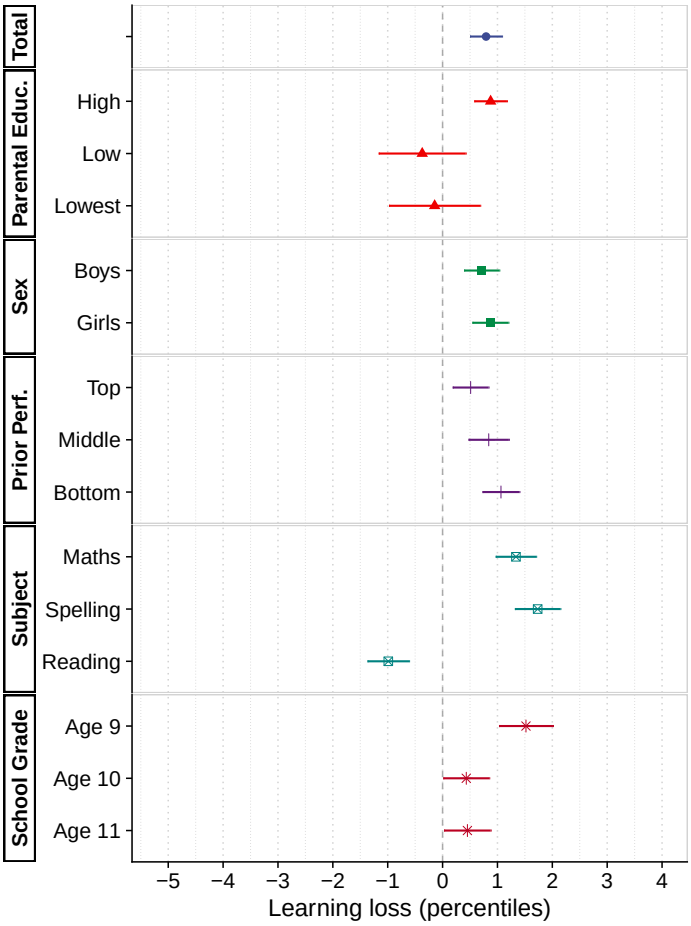


Figure 6: **Results with covariate balancing.** The graph shows results using our main specification while balancing treatment and comparison years on the estimated propensity of treatment.

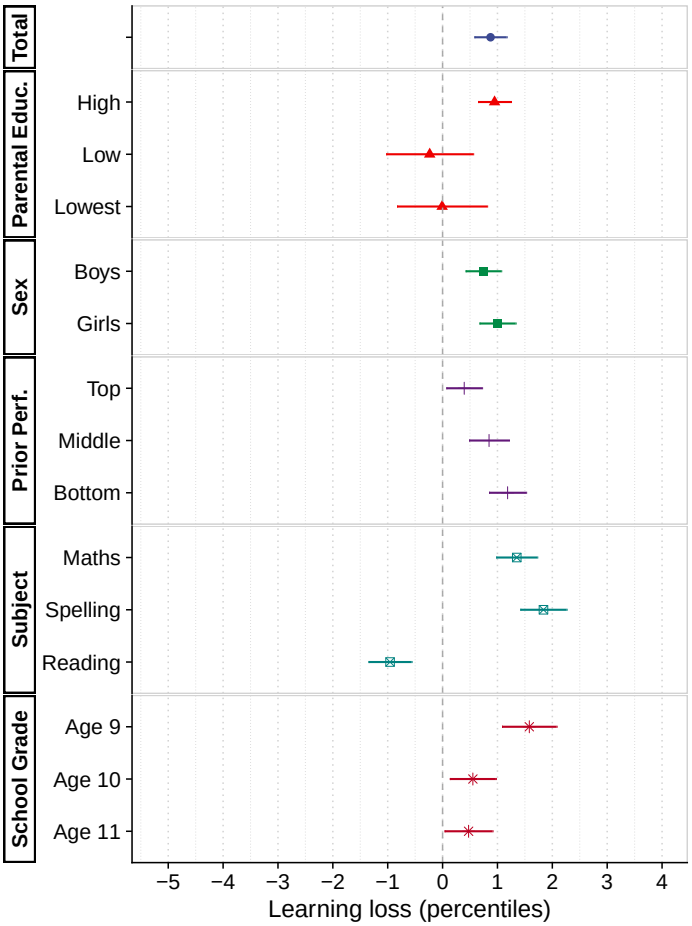


Figure 7: **School fixed effects.** The graph shows results combining our difference-in-differences with school fixed effects. This analysis discards all variation between schools by introducing a separate intercept for each school, thus adjusting for any heterogeneity across schools.

Table 1: Missingness table

	2019-2020			2020-2021		
School Age	9	10	11	9	10	10
Prior Performance	0.06	0.06	0.05	0.06	0.05	0.05
Composite	0.07	0.06	0.05	0.09	0.08	0.07
Maths	0.07	0.07	0.06	0.11	0.10	0.09
Reading	0.23	0.08	0.07	0.31	0.14	0.11
Spelling	0.07	0.07	0.06	0.11	0.12	0.11
Learning Readiness	0.12	0.16	0.20	0.18	0.23	0.29
N	29587	30193	30864	27853	28131	28853

Table 2: Summary statistics

	Control			Treated			
Variable	N	Mean	SD	N	Mean	SD	p-value
$\Delta Composite$	85093	0.02	12.72	77794	0.9	13.47	<0.001 (F)
$\Delta Maths$	84456	0.04	17.35	76274	1.45	17.96	<0.001 (F)
$\Delta Reading$	79328	0.18	20.49	69079	-0.72	20.9	<0.001 (F)
$\Delta Spelling$	84503	-0.04	18.96	75118	1.78	19.3	<0.001 (F)
Parental Education	85569			80037			<0.001 (X^2)
... high		0.92			0.93		
... low		0.04			0.03		
... lowest		0.04			0.04		
Sex	85569			80037			0.255 (X^2)
... Female		0.5			0.5		
... Male		0.5			0.5		
Prior Performance	85569			80037			0.003 (X^2)
... top		0.33			0.33		
... middle		0.35			0.35		
... bottom		0.32			0.33		
School Grade	85569			80037			0.699 (X^2)
... Age 9		0.32			0.33		
... Age 10		0.33			0.33		
... Age 11		0.34			0.34		
School Disadvantage	85569	28.81	4.22	80037	28.75	4.2	0.003 (F)
School Denomination	85569			80037			0.55 (X^2)
... Christian		0.58			0.58		
... Other		0.13			0.14		
... Public		0.29			0.29		
% Non-Western	85569	0.16	0.16	80037	0.15	0.16	0.075 (F)

	Composite	Maths	Reading	Spelling
Treatment	0.87***	1.41***	−0.90***	1.82***
	(0.15)	(0.19)	(0.19)	(0.21)
(Intercept)	0.02	0.04	0.18	−0.04
	(0.11)	(0.13)	(0.13)	(0.15)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	162887	160730	148407	159621
RMSE	13.08	17.64	20.68	19.12
N Clusters	1096	1094	1088	1095

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

Table 3: Main effects by subject

	Composite	Maths	Reading	Spelling
Treatment	0.95***	1.53***	−0.86***	1.89***
	(0.15)	(0.19)	(0.20)	(0.22)
Treat x Par. Educ. (low)	−1.27**	−1.52**	−0.90	−1.40*
	(0.40)	(0.53)	(0.65)	(0.60)
Treat x Par. Educ. (lowest)	−1.00*	−1.95***	−0.22	−0.61
	(0.42)	(0.57)	(0.61)	(0.67)
Parental Educ. (low)	−0.94***	−1.30***	−1.34***	−0.27
	(0.25)	(0.32)	(0.39)	(0.37)
Parental Educ. (lowest)	0.32	0.17	0.24	0.48
	(0.26)	(0.36)	(0.37)	(0.39)
(Intercept)	0.04	0.08	0.22	−0.05
	(0.11)	(0.13)	(0.14)	(0.15)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	162887	160730	148407	159621
RMSE	13.08	17.64	20.68	19.12
N Clusters	1096	1094	1088	1095

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

Table 4: Main effects by students' parental education and subject

	Composite	Maths	Reading	Spelling
Treatment	0.75^{***}	1.19^{***}	-0.92^{***}	1.69^{***}
	(0.16)	(0.21)	(0.23)	(0.23)
Treat x Female	0.25	0.43^{*}	0.05	0.25
	(0.15)	(0.20)	(0.25)	(0.21)
Female	-0.70^{***}	0.05	-0.98^{***}	-1.20^{***}
	(0.09)	(0.13)	(0.15)	(0.13)
(Intercept)	0.37^{**}	0.01	0.67^{***}	0.56^{***}
	(0.12)	(0.14)	(0.16)	(0.16)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	162887	160730	148407	159621
RMSE	13.08	17.64	20.68	19.11
N Clusters	1096	1094	1088	1095

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

Table 5: Main effects by student sex and subject

	Composite	Maths	Reading	Spelling
Treatment	0.53^{**}	0.90^{***}	-0.67^{**}	1.19^{***}
	(0.17)	(0.22)	(0.24)	(0.23)
Treat x Prior Perf. (middle)	0.35^{*}	0.59^{*}	-0.43	0.77^{**}
	(0.17)	(0.23)	(0.29)	(0.26)
Treat x Prior Perf. (bottom)	0.57^{**}	0.85^{***}	-0.31	0.99^{***}
	(0.19)	(0.26)	(0.30)	(0.27)
Treat x Prior Perf. (bottom)	0.57^{**}	0.85^{***}	-0.31	0.99^{***}
	(0.19)	(0.26)	(0.30)	(0.27)
Prior Perf. (middle)	3.67^{***}	3.07^{***}	3.82^{***}	4.10^{***}
	(0.12)	(0.16)	(0.19)	(0.17)
Prior Perf. (bottom)	7.10^{***}	5.95^{***}	7.80^{***}	7.53^{***}
	(0.13)	(0.17)	(0.19)	(0.20)
(Intercept)	-3.52^{***}	-2.92^{***}	-3.65^{***}	-3.86^{***}
	(0.12)	(0.15)	(0.16)	(0.16)
R^2	0.05	0.02	0.02	0.03
Adj. R^2	0.05	0.02	0.02	0.03
Num. obs.	162887	160730	148407	159621
RMSE	12.74	17.46	20.45	18.84
N Clusters	1096	1094	1088	1095

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

Table 6: Main effects by prior performance and subject

Table 7: Results by grade and subject

	Composite	Maths	Reading	Spelling
Age 9				
Treatment	1.61^{***}	2.64^{***}	-0.51	2.14^{***}
(0.25)	(0.31)	(0.35)	(0.36)	
(Intercept)	-0.31	-0.24	0.62^{*}	-1.01^{***}
	(0.20)	(0.25)	(0.27)	(0.29)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	52948	52298	42049	52181
RMSE	14.35	19.34	22.70	21.00
N Clusters	1095	1092	944	1095
Age 10				
Treatment	0.51^{*}	1.12^{***}	-1.12^{***}	1.41^{***}
	(0.21)	(0.27)	(0.27)	(0.29)
(Intercept)	0.08	0.06	-0.48^{*}	0.67^{*}
	(0.17)	(0.21)	(0.21)	(0.26)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	54078	53348	52094	52955
RMSE	12.63	17.53	20.47	17.86
N Clusters	1092	1087	1081	1089
Age 11				
Treatment	0.52^{*}	0.52^{*}	-0.97^{***}	1.90^{***}
	(0.22)	(0.26)	(0.27)	(0.33)
(Intercept)	0.28	0.29	0.46^{*}	0.20
	(0.17)	(0.20)	(0.20)	(0.26)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	55861	55084	54264	54485
RMSE	12.22	15.98	19.16	18.38
N Clusters	1091	1089	1081	1088

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

	Composite	Maths	Reading	Spelling
Treatment	0.91***	1.53***	−0.78***	1.61***
	(0.17)	(0.23)	(0.24)	(0.25)
Treat x share non-Western (middle)	−0.13	−0.40	−1.14	1.15
	(0.48)	(0.51)	(0.61)	(0.66)
Treat x share non-Western (high)	−0.05	−0.31	0.27	0.25
	(0.39)	(0.52)	(0.46)	(0.54)
Share non-western (middle)	0.30	0.38	0.54	0.18
	(0.32)	(0.35)	(0.41)	(0.44)
Share non-western (high)	1.07***	1.09**	0.87**	1.13**
	(0.27)	(0.35)	(0.32)	(0.37)
(Intercept)	−0.23	−0.23	−0.08	−0.29
	(0.13)	(0.16)	(0.17)	(0.18)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	162887	160730	148407	159621
RMSE	13.07	17.64	20.68	19.11
N Clusters	1096	1094	1088	1095

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

Table 8: Main effect by non-Western student population and subject

	Composite	Maths	Reading	Spelling
Treatment	2.36***	2.42***	0.58	4.04***
	(0.49)	(0.71)	(0.58)	(0.78)
Treat x Schoolweight (23-25)	−1.05	−0.15	−1.40	−2.03*
	(0.63)	(0.88)	(0.82)	(0.93)
Treat x Schoolweight (26-28)	−0.90	−0.47	−0.66	−1.60
	(0.56)	(0.79)	(0.69)	(0.89)
Treat x Schoolweight (29-31)	−2.10***	−1.24	−2.37***	−3.11***
	(0.56)	(0.78)	(0.67)	(0.88)
Treat x Schoolweight (32-34)	−2.07**	−2.31*	−1.88*	−2.49*
	(0.68)	(0.91)	(0.82)	(0.98)
Treat x Schoolweight (35-37)	−2.11**	−2.17*	−1.55	−2.71*
	(0.72)	(0.97)	(0.87)	(1.07)
Treat x Schoolweight (38+)	−2.09*	−1.97	−1.82	−2.70
	(1.04)	(1.44)	(1.26)	(1.47)
School weight (23-25)	−1.41**	−1.85**	−0.56	−1.48*
	(0.45)	(0.57)	(0.58)	(0.65)
School weight (26-28)	−2.36***	−3.01***	−2.05***	−1.92**
	(0.39)	(0.51)	(0.50)	(0.59)
School weight (29-31)	−2.07***	−2.81***	−1.83***	−1.48*
	(0.38)	(0.50)	(0.48)	(0.58)
School weight (32-34)	−2.39***	−2.88***	−2.50***	−1.63*
	(0.49)	(0.62)	(0.62)	(0.68)
School weight (35-37)	−1.33**	−1.76**	−1.69**	−0.50
	(0.49)	(0.64)	(0.59)	(0.73)
School weight (38+)	−1.36*	−1.94*	−0.73	−1.34
	(0.63)	(0.81)	(0.64)	(0.95)
(Intercept)	1.89***	2.47***	1.78***	1.39**
	(0.34)	(0.45)	(0.42)	(0.52)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	162887	160730	148407	159621
RMSE	13.05	17.61	20.66	19.10
N Clusters	1096	1094	1088	1095

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

Table 9: Main effect by school disadvantage and subject

	Composite	Maths	Reading	Spelling
Treatment	0.81^{***}	1.35^{***}	-0.95^{***}	1.75^{***}
	(0.14)	(0.18)	(0.19)	(0.20)
Female	-0.64^{***}	0.22^{*}	-0.99^{***}	-1.15^{***}
	(0.06)	(0.09)	(0.10)	(0.09)
Parental Educ. (low)	-3.31^{***}	-3.55^{***}	-3.62^{***}	-2.81^{***}
	(0.19)	(0.23)	(0.29)	(0.25)
Parental Educ. (lowest)	-2.23^{***}	-2.54^{***}	-2.06^{***}	-2.02^{***}
	(0.20)	(0.26)	(0.28)	(0.28)
Prior Perf. (middle)	3.93^{***}	3.45^{***}	3.72^{***}	4.54^{***}
	(0.09)	(0.12)	(0.14)	(0.13)
Prior Perf. (bottom)	7.67^{***}	6.67^{***}	7.97^{***}	8.27^{***}
	(0.10)	(0.13)	(0.14)	(0.14)
(Intercept)	-3.26^{***}	-3.15^{***}	-2.95^{***}	-3.49^{***}
	(0.11)	(0.14)	(0.16)	(0.16)
R^2	0.06	0.02	0.02	0.03
Adj. R^2	0.06	0.02	0.02	0.03
Num. obs.	162887	160730	148407	159621
RMSE	12.71	17.44	20.43	18.83
N Clusters	1096	1094	1088	1095

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

Table 10: Main effects with controls

	Composite	Maths	Reading	Spelling
Treatment	0.71^{***}	1.28^{***}	-0.92^{***}	1.77^{***}
	(0.15)	(0.20)	(0.20)	(0.22)
(Intercept)	0.12	0.06	0.20	0.10
	(0.11)	(0.14)	(0.14)	(0.15)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	145238	145238	145238	145238
RMSE	12.77	17.54	20.69	19.04
N Clusters	1088	1088	1088	1088

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

Table 11: Main effects, complete subject scores only

	Composite	Maths	Reading	Spelling
Treatment	-2.84^{***}	-3.38^{***}	-2.34^{***}	-2.35^{***}
	(0.24)	(0.29)	(0.29)	(0.33)
(Intercept)	0.73^{***}	0.85^{***}	0.14	0.94^{***}
	(0.18)	(0.22)	(0.22)	(0.24)
R^2	0.01	0.01	0.00	0.00
Adj. R^2	0.01	0.01	0.00	0.00
Num. obs.	87499	80722	69617	77506
RMSE	13.93	17.93	20.82	18.78
N Clusters	854	834	797	833

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

Table 12: Overall learning loss, by subject, including first shutdown only (end-of-year to end-of-year tests

	Composite	Maths	Reading	Spelling
Treatment	1.11^{***}	1.74^{***}	-0.75^{***}	2.07^{***}
	(0.18)	(0.23)	(0.22)	(0.27)
(Intercept)	0.24	0.19	0.48^{**}	0.14
	(0.13)	(0.16)	(0.16)	(0.18)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	0.00
Num. obs.	103618	102364	95141	101825
RMSE	13.00	17.56	20.60	19.10
N Clusters	875	868	859	871

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

Table 13: Overall learning loss by subject, only for students who also took the 2020 end-of-year test

	Composite	Maths	Reading	Spelling
Treatment	0.87^{***}	1.35^{***}	-0.96^{***}	1.84^{***}
	(0.15)	(0.19)	(0.20)	(0.22)
R^2	0.04	0.04	0.03	0.04
Adj. R^2	0.04	0.03	0.02	0.03
Num. obs.	162887	160730	148407	159621
RMSE	12.84	17.40	20.50	18.84
N Clusters	1096	1094	1088	1095

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

Table 14: Main effects with school fixed effects

	Composite	Maths	Reading	Spelling
Treatment	0.95^{***}	1.47^{***}	-0.93^{***}	1.90^{***}
	(0.15)	(0.19)	(0.21)	(0.22)
Treat x Par. Educ. (low)	-1.18^{**}	-1.47^{**}	-0.80	-1.27^{*}
	(0.40)	(0.53)	(0.64)	(0.59)
Treat x Par. Educ. (lowest)	-0.96^{*}	-1.88^{***}	-0.13	-0.61
	(0.42)	(0.56)	(0.61)	(0.67)
Parental Educ. (low)	-0.50^{*}	-0.79^{*}	-0.83^{*}	0.08
	(0.24)	(0.31)	(0.38)	(0.34)
Parental Educ. (lowest)	0.14	0.07	0.02	0.22
	(0.24)	(0.35)	(0.37)	(0.37)
R^2	0.04	0.04	0.03	0.04
Adj. R^2	0.04	0.03	0.02	0.03
Num. obs.	162887	160730	148407	159621
RMSE	12.84	17.40	20.49	18.84
N Clusters	1096	1094	1088	1095

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

Table 15: School inequality with school fixed effects

References

[1] J Scheerens, M Ehren, P Sleegers, R de Leeuw, Country background report for the Netherlands (OECD Review on evaluation and assessment frameworks for improving school outcomes) (2012).

[2] J Zapata, et al., *Education Policy Outlook: Netherlands*. (Organisation for Economic Co-operation and Development OECD), (2014).

[3] HA Patrinos, School choice in the Netherlands (CESifo DICE Report 2/2011, ifo Institute for Economic Research, Munich) (2011).

[4] JM Ritzen, J Van Dommelen, FJ De Vijlder, School finance and school choice in the netherlands. *Economics of Education Review* **16**, 329–335 (1997).

[5] HF Ladd, EB Fiske, Weighted student funding in the Netherlands: A model for the US? *Journal of Policy Analysis and Management* **30**, 470–498 (2011).

[6] A Schleicher, *PISA 2018: Insights and interpretations*. (OECD Publishing), (2018).

[7] M de Haas, R Faber, M Hamersma, How COVID-19 and the Dutch 'intelligent lockdown' changed activities, work and travel behaviour: Evidence from longitudinal data in the Netherlands. *Transportation Research Interdisciplinary Perspectives*, 100150 (2020).

[8] ME Kuiper, et al., The intelligent lockdown: Compliance with COVID-19 mitigation measures in the Netherlands (PsyArXiv) (2020).

[9] P Tullis, Dutch cooperation made an 'intelligent lockdown' a success. *Bloomberg Businessweek* (2020).

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

[11] Statistics Netherlands (CBS), The Netherlands leads Europe in internet access (https://www.cbs.nl/en-gb/news/2018/05/the-Netherlands-leads-europe-in-internet-access) (2018).

[12] 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).

[13] FM Reimers, A Schleicher, *A framework to guide an education response to the COVID-19 Pandemic of 2020*. (OECD Publishing), (2020).

[14] 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/) (2020).

[15] E Grewenig, P Lergetporer, K Werner, L Woessmann, L Zierow, COVID-19 and educational inequality: How school closures affect low-and high-achieving students (IZA Discussion Paper 13820, Institute of Labor Economics, Bonn.) (2020).

[16] T Bol, Inequality in homeschooling during the corona crisis in the Netherlands: First results from the LISS panel (SocArXiv) (2020).

[17] A Andrew, et al., Inequalities in children's experiences of home learning during the COVID-19 lockdown in England. *Fiscal Studies* **41**, 653–683 (2020).

[18] A Bacher-Hicks, J Goodman, C Mulhern, Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. *Journal of Public Economics* **193**, 104345 (2020).

[19] C Bansak, M Starr, COVID-19 shocks to education supply: how 200,000 US households dealt with the sudden shift to distance learning. *Review of Economics of the Household*, in press (2021).

[20] MM Jæger, EH Blaabæk, Inequality in learning opportunities during COVID-19: Evidence from library takeout. *Research in Social Stratification and Mobility* **68**, 100524 (2020).

[21] KF Vlug, Because every pupil counts: the success of the pupil monitoring system in the Netherlands. *Education and Information Technologies* **2**, 287–306 (1997).

[22] D Fettelaar, E Smeets, Mogelijke indicatoren van schoolgewichten: Onderzoek naar de voorspellende waarde (ITS Institute of Applied Social Sciences, Radboud Universiteit Nijmegen) (2013).

[23] G Driessen, De wankele empirische basis van het onderwijsachterstandenbeleid. *Mens en Maatschappij* **90**, 221 (2015).

[24] A de Wijs, F Kamphuis, F Kleintjes, M Tomesen, Wetenschappelijke verantwoording: Spelling voor groep 3 tot en met 8 (Cito) (2010).

[25] H Feenstra, F Kamphuis, F Kleintjes, R Krom, Wetenschappelijke verantwoording: Begrijpend lezen voor groep 3 tot en met 6 (Cito) (2010).

[26] J Janssen, N Verhelst, R Engelen, F Scheltens, Wetenschappelijke verantwoording: Rekenen-wiskunde voor groep 3 tot en met 8 (Cito) (2010).

[27] J Jerrim, A Vignoles, Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society Series A* **176**, 887–906 (2013).

[28] L Borghans, BH Golsteyn, U Zölitz, Parental preferences for primary school characteristics. *The BE Journal of Economic Analysis & Policy* **15**, 85–117 (2015).

[29] N Ruijs, H Oosterbeek, School choice in Amsterdam: Which schools are chosen when school choice is free? *Education Finance and Policy* **14**, 1–30 (2019).

[30] H Posthumus, et al., Herziening gewichtenregeling primair onderwijs: Fase I (Statistics Netherlands) (2016).

[31] H Posthumus, S Scholtus, J Walhout, De nieuwe onderwijs achterstandenindicator primair onderwijs: Samenvattend rapport (Statistics Netherlands) (2019).

[32] JD Angrist, JS Pischke, *Mostly Harmless Econometrics: An empiricist's companion.* (Princeton University Press), (2008).

[33] PR Rosenbaum, DB Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).

[34] GW Imbens, JM Wooldridge, Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**, 5–86 (2009).

[35] JM Wooldridge, *Econometric analysis of cross section and panel data.* (MIT press), (2010).

[36] MA Kraft, Interpreting effect sizes of education interventions. *Educational Researcher* **49**, 241–253 (2020).

[37] J Cohen, *Statistical power analysis for the behavioral sciences.* (Academic press), (2013).

[38] HG van de Werfhorst, Inequality in learning is a major concern after school closures. *Proceedings of the National Academy of Sciences* **118** (2021).