Differential Effects of COVID-19 School Closures on Students' Learning: Preregistration

Tim Fütterer[a], Tony C. A. Tan[b], Rolf V. Olsen[b], and Sigrid Blömeke[b]

[1]Hector Research Institute of Education Sciences and Psychology, University of Tübingen

[2]Centre for Educational Measurement, University of Oslo

A PDF printout of OSF preregistration

Tübingen, Germany

Thursday, 11 August 2022

Author Note

Tim Fütterer https://orcid.org/0000-0001-5399-9557

Tony C. A. Tan https://orcid.org/0000-0001-6632-3791

Rolf V. Olsen https://orcid.org/0000-0002-9621-4083

Sigrid Blömeke https://orcid.org/0000-0001-6068-4932

Correspondence concerning this preregistration should be addressed to Dr Tim Fütterer, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany. Email: tim.fuetterer@uni-tuebingen.de

Preregistration

Title

Differential Effects of COVID-19 School Closures on Students' Learning

Authors

Tim Fütterer, Tony C. A. Tan, Rolf V. Olsen, Sigrid Blömeke

Keywords

COVID-19, student achievement, social inequality, Norway, socioeconomic status, school closures

Description

Preface: This study uses Norway's national registers to investigate the associations between COVID-19 school closures and students' learning outcomes. Register datasets overcome sampling issues by preserving information about the entire Norwegian population. Due to our source data's large sizes, we expect statistical significance for most inferential parameters. In addition to preventing questionable practices such as *p*-hacking, this preregistration serves to enhance research transparency by declaring our research questions and methodological approaches before key variables become available to the authors.

School closures resultant from the COVID-19 pandemic in 2020 and 2021 represented a sudden and unexpected disruption of students' learning in schools. "School systems had to rapidly improvise to ensure some continuity in the education of children and adapt their teaching methods to a situation in which, in the space of a day, the setting in which education took place moved from the school to the home for most children and the mode of instruction shifted from face-to-face contact between pupils and their teachers/instructors to some form of remote or distance learning, often supervised by parents" (Thorn & Vincent-Lancrin, 2021, p. 13). Findings from previous studies suggest that school closures had a negative effect on student achievement ($d = -0.005$ SD to $d = -0.05$ SD per week), especially for students with low socioeconomic status (Hammerstein et al., 2021). In their meta-analysis, Betthäuser et al. (2022) found an overall negative effect of school closures on student learning early in the pandemic across 34 studies (Cohen's $d = -0.17$). In summary, the results indicate that measures to maintain learning during school closures that began in March 2020 were not effective. Although these findings appeared robust as many studies examined large samples, used administrative or test data, and often employed methods enabling causal inferences (e.g., difference-in-difference approach), most of the prior studies contained weaknesses in their underlying data (Thorn & Vincent-Lancrin, 2021). For example, convenience samples were often used (e.g., Clark et al., 2021), data were not representative of the underlying population (e.g., Kuhfeld et al., 2020), or data were obtained from survey methods such as web-based surveys, mobile apps, or telephone interviews result in distorted samples and/or low response rates (e.g., van der Velde et al., 2021). Moreover, although some consensus has emerged (e.g., greater learning loss among students with lower SES), uncertainties remain among mixed findings across different subjects. Lastly, school closures had prevented achievement assessments from taking place. The lack of data presents educational researchers with additional challenge of generating evidence on the impact of school closures. Against this background, there is a need for both studies that se the impact of school closures on student achievement using enhanced methodology based on high quality data (e.g., representative data allowing us to draw conclusions about the entire population including family circumstances that are important for learning at home) and studies on how to accommodate systemic missing data.

We aim to conduct two studies using Norway's register data between 2009 and 2020. In Study 1 (number of archival entries ≈ 12.3 million), we will present a Bayesian approach to estimating missing exam data in 2020 using Year 10 students' teacher-assigned grades and exam grades from the previous 10 years, with particular focus on mathematics, Norwegian and English. In Study 2 (number of archival entries ≈ 5.6 million), we will use Year 8 and 9 students' national test data in reading and mathematics as measures for students' learning. Since identical tests are used in Year 8 and 9, learning growths can be operationalised using difference scores. We will use difference-in-difference (DiD) approaches to compare learning progression (i.e., gains or losses) between the cohorts effected by school closures due to COVID-19 and previous cohorts. We therefore aim to provide robust findings that allow causal inferences on the effects of school closures (e.g., duration) on students' learning progression. This study focuses on the school closure elasticity of student achievement (i.e., percentage change in student academic achievement in response to percentage change in school closure). This study aims to gain insights into the importance of teaching and learning in schools. In addition, we aim to shed light into the differential effects of school closures by looking at important background variables of learning at home (e.g., parental income and education status, and housing conditions in terms of floor areas per person). Findings from this study will assist future policy-formation by quantifying educational costs resultant from major social measures.

The full preregistration is available in PDF on our OSF project page.

Study Information

Research questions

Our overall research question is: How did school closures affect students' learning? We plan on approaching this overarching theme through two studies. The first study explores how statistical methods can address the systemic missing of important information through this research question (RQ):

RQ1 How can missing data in Year 10 exam grades in 2020 be estimated using the Bayesian inference approach?

In the second study, we examine differential effects of school closures on the estimated student achievement (Year 10, see RQ1) and national test data (Year 8 and 9) in reading and mathematics before and during the school closures. Specifically, we wish to investigate the following research questions:

RQ2 What impact had school closures had on students' learning outcomes?

RQ3 How were students' learning outcomes related to students' socioeconomic status, household, and family characteristics?

RQ4 How did learning outcomes during closure differ between low and high achievers?

Hypothesis

We will consider all effects to be meaningful for both the period that includes the first school closures in 2020 (Period 1) and the period that includes the second school closures in 2021 (Period 2).

H1 Regarding RQ2: We expect small effect sizes $d = -0.2/52 = -0.0038$ SD per week or smaller in learning loss in Period 1.

Notes: For a justification of H1 see the Effect Size section. For Period 2, we will conduct exploratory analyses (see the Exploratory Analyses section). The term learning loss does not suggest a decline in total amount of learning, but smaller growth sizes compared to equivalent cohorts before COVID school closures. Since identical national tests are implemented in Year 8 and Year 9 (see the Variables section), we expect cohorts prior to 2020 to have advantages in learning growth. In addition, we expect to see differential effects of school closures on students' learning growth.

H2 Regarding RQ3: Effects of school closures on students' learning are moderated by students' SES in Period 1, with lower SES students showing more learning loss from school closures than their higher SES peers.

Justification for H2: Previous studies on Period 1 highlighted the critical role SES played for the extent of learning loss. Although Norwegian residents usually enjoy high standards of living, such privilege is distributed unevenly when comparing the mean of the 5th against that of the 95th percentile (OECD, 2019), suggesting unequal relationship between school closures and SES. More specifically, we conjecture that the less favourable students' SES, the greater their learning loss.

For Period 2, we will conduct exploratory analyses (see the Exploratory Analysis section), but we expect similar differential effects as for Period 1 such that SES significantly moderates the relationship between school closures and learning loss.

H3 Regarding RQ4: Effects of school closures on students' learning are moderated by students' previous achievement in Period 1. Low achievers show more learning loss due to school closures than high achievers.

Justification for H3: It is well known that low and high achievers share dissimilar metacognitive capabilities. As high achievers are usually better at self-regulated learning, we expect high achievers to be affected less by distance learning than low achievers. For Period 2, we will conduct exploratory analyses (see the Exploratory Analysis section), with similar differential effects from Period 1. We expect low achievers to show more learning loss due to school closures in Period 2 than high achievers.

Data Description

Datasets

This project sources its data from the Norwegian national register. This data source captures information about the entire Norwegian population dating back to the early 1900s through each individual's unique national ID number. Under a secured IT environment, we obtain national statistics on Norwegian residents' education (e.g., person's highest level of education, academic attainment record), employment (e.g., working hours per week), income (e.g., after-tax income), COVID-19 specific information (e.g., COVID infection rates), housing conditions (e.g., floor areas, number of persons per household), as well as family relations and composition (e.g., kinship, number of siblings). Most importantly, entries across different datasets can be linked through person IDs, enabling us to match students with their parents' education and income data as well as their housing conditions such as floor space. Furthermore, municipality-level (e.g., duration of school closures) and school-level data (e.g., school resources, student composition) can be linked to student-level outcomes (e.g., teacher-assigned grades, exam grades, and national test grades).

Data Availability

The datasets underlying this project were provided by Statistics Norway (SSB) and the Norwegian Ministry of Education (UDIR) by permission. Researchers can gain access to these datasets by submitting written applications to SSB and by following instructions on UDIR website respectively. The Norwegian national register contains large amount of private and sensitive information. Research institutes must provide sufficient justification and undergo a rigorous application process for data access. The Norwegian government requires data access to be granted only to registered users and within a secured IT environment that fully logs every operation.

Data Access

The dataset can be accessed using secured IT infrastructure only.

Data Identifiers

No persistent, unique identifier of the datasets is available.

Access Date

The second author received access credential to the register data in April 2022. However, we do not have access to key independent (e.g., durations of school closures, housing condition data) and dependent variables (e.g., 2021 grades and national test data) at the time of preregistration lodgement. Retrieval applications will be submitted in August 2022, with expected delivery in autumn 2022.

Data Collection Procedures

Norwegian national register is maintained by Statistics Norway (SSB). SSB is the national statistical institute of Norway and the main producer of official statistics. SSB is responsible for collecting statistics related to the economy, population, and society at national, regional and local levels. Information related to school characteristics was managed by the Ministry of Education and Training (UDIR) through the School Information System (GSI). Our project team received access to both data sources in April 2022. No further data collection is conducted.

Codebook

Table 1 describes key variables used in our study. A full codebook can be obtained from our OSF page.

Variables

Manipulated Variables

Manipulation, blinding, and randomisation is not applicable to any unit of analyses in this study due to its archival data design.

Measured Variables

The SSB collects person-level information from all Norwegian residents. The education database covers graduation statistics since 1970 and national test results since 2007. The population register includes information about household composition and family relations between 1975 and 2005, as well as housing conditions since 1990. Wealth and income data are available since 1993, including cash support information since 1999 and employment statistics since 2000. The UDIR's GSI database contains information about (primary and lower secondary) schools in Norway since 1992, including COVID-related restrictions and measures in 2020 and 2021.

Table 1 summarizes key variables used in this study. Full variable descriptions can be found in the codebook from our OSF page once all authors have gained data access.

Regarding Study 2

We will create the dichotomous variables `condition1` and `condition2`, which will encode whether the analysed achievement trends between Year 8 and Year 9 of the students relate to the first period of school closures (`condition1` = 1) or not (`condition1` = 0) or whether the achievement trends relate to the second period of school closures (`condition2` = 1) or not (`condition2` = 0).

Key variables to answer our research questions and to test our hypotheses are:

1. The dependent variables are students achievement in reading and mathematics (scores) measured by national tests (nr. 10 in Table 1).

2. We will use the newly computed variables `condition1` as independent variables to test whether there is an effect of school closures in Period 1 on student achievement without considering detailed varying durations of school closures.

3. We will use the duration of school closures (nr. 11 in Table 1) as our main independent variable to answer research Question 2.

4. We will use after-tax income per consumption unit (nr. 13 in Table 1) as our main moderator representing student's SES to answer RQ3.

5. We will use teacher-assigned grades for mathematics and written Norwegian (nr. 7 in Table 1) as our main moderators representing student's previous achievement to answer RQ4.

Figure 1 provides a visual illustration of the temporal locations of key variables relative to student achievement measures and school closures.

Unit of Analysis

In Study 1, we will focus on Year 10 students as this cohort represents the end of Norway's compulsory education (*grunnskole*) years. Afterwards, students have the option to continue into either vocational (*yrkesfaglig opplæring*) or academic (*studieforberedende opplæring*) streams based on their academic performance (*grunnskolepoeng*, grade point average [GPA]). Resultantly, academic achievement data for Year 10 contain a large number of common subjects and minimal missing data due to their compulsory nature. Based on our current knowledge of the data, we will delete subjects designed for returning adults (e.g., ENGV) as well as subjects in special-purpose school that do not follow the standardized schooling system (e.g., ENGM). According to the second author's previous studies, we estimate that we will exclude one to two percent of the cases from our analyses.

In Study 2, we will focus on Year 9 students as they repeat the national reading and mathematics tests from a year ago, forming a pre-test–intervention–post-test design.

Missing Data

The first study involves two categories of missing data: missing by design and sporadic missing. Missing by design were the result of random allocation of candidates into mathematics, Norwegian, and English written exams (2/3 missing probability; missing completely at random [MCAR]). Sporadic missing refers to small scale absence due to non-recording of some information such as students' education attainment and/or demographic data. Multiple imputation (MI) will be used to impute both types of missings thanks to its ability to calculate parameter standard errors (van Buuren, 2018, p. 25). Under the advisory of van Buuren (2018, p. 43), 10 draws will be conducted from the posterior distribution using R package mice (van Buuren & Groothuis-Oudshoorn, 2011). Most relevant to Study 1 are the systematic missing of exam grades in 2020 and 2021 resultant from Norway's COVID measures. Dealing with these missing values is the core mission of Study 1. We will use a Bayesian approach to infer the plausible values of missing exam grades (see the Analyses section).

The second study contains sporadic missingness. Although Norway's national tests are compulsory in principle, schools do have the ability to grant exemptions following specific guidelines, leading to missing values in assessment outcome variables. Furthermore, the past five years witnessed a sharp increase in exemption rates to 30 to 40 percent, likely due to student disadvantages. Since SES are observed variables, the missing process in the national tests can be modelled using MI under the missing at random (MAR) assumption.

Should evidence emerge suggesting the moderating role of SES in national test participation, this study would further investigate the differences between students who have received exemptions (the drop-out group) and those who have not (remain group). More specifically, we will conduct logistic regressions for both reading and mathematics by creating a dichotomous variable (`attr`; 0 = participated in national tests, 1 = exempted from national tests) and predict `attr` using the independent variables (e.g., school closures) and moderators (e.g., SES, low/high achievers). We expect national test exemptions to be correlated with students' background variables (e.g., the most disadvantaged students may have been excluded to protect them from unfair testing; Eisner et al., 2019).

We pay special attention to the missing data processes. If the probability of missing national tests is correlated only with observable variables such as students' SES and not related to the test grades, assessment outcome data can be considered to be missing at random (MAR). MAR is not an implausible assumption for Norway since great government efforts are aimed to decouple educational achievement from social disadvantages. Resultantly, multiple imputation (MI) would guarantee unbiasedness if the regressions incorporate all covariates associated with the MAR process (van Buuren, 2018; Graham, 2012). An alternative to the MAR process is missing not at random (MNAR), under which the probability of missing national tests is correlated with the test grade itself (for instance, poor performing students are discouraged from participating in national tests). This possibility is explicitly prohibited by Norwegian laws and regulations and therefore will not be pursued by this study. Lastly, the robustness of the MI procedure will be verified using sensitivity analyses (see robustness testing section).

Statistical Outliers

In principle, we define outliers as values that are three standard deviations (SDs) away from their means. However, we will exercise discretion in determining the reasonableness of each value. For example, after marking income data 3 SDs away from the national average, we exclude unusually low figures (e.g., 1 Norwegian krone) as outliers due to their inconsistency with Norway's social safety net but retain unusually high income as meaningful financial records. Invalid or impossible entries such as negative floor areas will also be marked as outliers and overwritten as missing values.

Sampling Weights

Sampling weights and stratification do not apply to this study since the national register represents the entire Norwegian population.

Knowledge of Data

Prior Publication/Dissemination

We worked on no publications, working papers, or conference presentations based on the dataset we will use for this study.

Prior Knowledge

As at submission, the second author has commenced preparing the register data into the research-ready format. The second author has also generated descriptive statistics for all variables contained in the register database to assist other researchers with variable selection. This overview is merely administrative in nature with minimal insight into the fine structures of the variables. All co-authors do not have computer access to the datasets at submission. Public record from the Norwegian Directorate for Education and Training (UDIR, 2021) showed no major differences between 2019 and 2021 in student's reading and mathematics achievement, either between genders, immigration backgrounds or parental educational levels. We are also aware that the national test exemption rates have increased by 30 to 40 percent over the past five years. The authors report no other knowledge about the data, particularly relating to the variable relationships needed to answer our research questions.

Analyses

Statistical Models

Study 1

Bayesian Approach. In answering Study 1's research question, we propose a Bayesian procedure for estimating the missing 2020 exam data. Since educational assessment maps students' learning outcomes (L) onto a numerical scale (M), a valid and reliable assessment inventory should show high degrees of agreement between L and M, that is, $\mathbb{P}(M \mid L)$ ("given the learning, how likely these grades appear") shall be close to 1. Stakeholders, on the other hand, are more interested in knowing "given the grades, how likely there is learning" (i.e., $\mathbb{P}(L \mid M)$). These two interests can be linked via the Bayes formula:

$$\mathbb{P}(L \mid M) = \frac{\mathbb{P}(M \mid L)\,\mathbb{P}(L)}{\text{normalizing constant}}, \tag{1}$$

where $\mathbb{P}(L)$ is a "prior belief" of learning that is to be updated by the exam results. In this study, teacher-assigned grades can serve as the prior. Properties of $\mathbb{P}(M \mid L)$ can be ascertained from earlier years' exam papers as exam questions are reused. The normalizing constant can be approximated using Markov chain Monte Carlo (MCMC) methods should manual computation become infeasible.

Study 2

Descriptive Statistics. In a first step, we plan to gain insight into what conditions did Norwegian students study at home during school closures and to get an overview of characteristics of the schools and municipalities, we will report descriptive statistics (e.g., arithmetic mean, standard deviation) of the independent and dependent variables as well as the moderators on student, school and municipality level. In a second step, we will test whether the schools differ descriptively and statistically significantly in key variables (e.g., SES) over the last five years (i.e., 2015 to 2019) before COVID-19 led to measures such as school closures. For testing differences, we will conduct analyses of variance (ANOVA) for each year separately. We will assume that schools differ with respect to a key variable if differences are evident in five or three years between 2015 and 2019. In a third step, we will describe students excluded and not excluded from the national tests separately (e.g., age, sex, SES).

Difference-in-Difference Approach. We will analyze students' learning progression (i.e., gain or loss) between 2019 (students' achievement in national tests [reading and mathematics] in Year 8) and 2020 (students' achievement in national tests [reading and mathematics] in Year 9) using a difference-in-difference (DiD) approach (Angrist & Pischke, 2009) conducted in R (e.g., Brumback, 2021) similar to the analytical approach that was used by Engzell et al. (2021). That is, we will compare students' learning progression between 2019 and 2020 (Period 1 [i.e., period of 1st school closures]) with that in the five previous periods (e.g., 2018 to 2019). We will use the five periods before 2020 because from 2014 on results in the national tests are measured in score points.

First, we will calculate differences between students' national test grades of Year 8 and Year 9 using the following equation: $\Delta y_i^{(k+1)-k} = y_i^{k+1} - y_i^k$, where y_i^k is the achievement (i.e., grade) of a student i in the national tests in year k ($i, k \in \mathbb{N}$). The year $k+1$ refers to the year where a student was a Year 9 student and k refers to the year where the same student was a Year 8 student. We will calculate the difference scores $\Delta y_i^{2015-2014}$, $\Delta y_i^{2016-2015}$, $\Delta y_i^{2017-2016}$, $\Delta y_i^{2018-2017}$, $\Delta y_i^{2019-2018}$ (five difference scores before COVID-19 school closures in 2020) and $\Delta y_i^{2020-2019}$ (difference score regarding the first school closure in 2020).

Second, we will compare these difference scores using a regression specification. More precisely, we will use two-level models with cross-level interactions (i.e., linear mixed-effect models) to account for students nested within schools. Equation 2 shows our basic model (using mathematics as an example; using only the level 2 variables dur [and not condition 1] as an example; using no covariates) that we will use to test our first hypothesis H1. Equation 3 shows or advanced model (using mathematics as an example; using additional covariates; if necessary, additional individual-level and school-level control variables will be included here).

Level 1

$$\Delta y_{ij} = \beta_{0j} + \varepsilon_{ij}, \tag{2.1}$$

$i, j \in \mathbb{N}$, i representing Student i, j representing School j and with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. An index ij references students within schools. In this example Δy_{ij} is $\mathrm{np\_math9}_{ij} - \mathrm{np\_math8}_{ij}$.

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} dur_j + u_{0j}. \tag{2.2}$$

Note that dur is indexed by j as dur varies at the school level. $u_{0j} \sim \mathcal{N}(0, \sigma^2_{u_0})$, $u_{3j} \sim \mathcal{N}(0, \sigma^2_{u_3})$, and $u_{4j} \sim \mathcal{N}(0, \sigma^2_{u_4})$. For the periods before COVID-19 school closures dur is expected to be always 0.

Total composite formula

$$\Delta y_{ij} = \gamma_{00} + \underbrace{\gamma_{01} dur_j}_{\text{RQ2}} + u_{0j} + \varepsilon_{ij} \tag{2.3}$$

The slope γ_{01} is of interest to answer RQ2.

Level 1

$$\Delta y_{ij} = \beta_{0j} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} + \beta_{3j} atipcu_{ij} + \beta_{4j} stp\_math_{ij} + \varepsilon_{ij}, \tag{3.1}$$

$i, j \in \mathbb{N}$, i representing Student i, j representing School j and with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\varepsilon})$. An index ij references students within schools. In this example Δy_{ij} is np_math9$_{ij}$ − np_math8$_{ij}$.

Level 2

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} dur_j + u_{0j} \\ \beta_{3j} &= \gamma_{30} + \gamma_{31} dur_j + u_{3j} \\ \beta_{4j} &= \gamma_{40} + \gamma_{41} dur_j + u_{4j}. \end{aligned} \tag{3.2}$$

Note that dur is indexed by j as varies at the school level. $u_{0j} \sim \mathcal{N}(0, \sigma^2_{u_0})$, $u_{3j} \sim \mathcal{N}(0, \sigma^2_{u_3})$, and $u_{4j} \sim \mathcal{N}(0, \sigma^2_{u_4})$. For the periods before COVID-19 school closures dur is expected to be always 0.

Total composite formula

$$
\begin{aligned}
\Delta y_{ij} &= \gamma_{00} + \gamma_{01} dur_j + u_{0j} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} \\
&+ (\gamma_{30} + \gamma_{31} dur_j + u_{3j}) atipcu_{ij} + (\gamma_{40} + \gamma_{41} dur_j + u_{4j}) stp_m ath_{ij} + \varepsilon_{ij} \\
&= \gamma_{00} + \underbrace{\gamma_{01} dur_j}_{\text{RQ2 (check, if it is robust, when using covariates)}} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} + \gamma_{30} atipcu_{ij} \\
&+ \underbrace{\gamma_{31} dur_j atipcu_{ij}}_{\text{cross-level interaction (RQ3)}} + \gamma_{40} stp\_math_{ij} + \underbrace{\gamma_{41} dur_j stp\_math_{ij}}_{\text{cross-level interaction (RQ4)}} \\
&+ \underbrace{u_{3j} atipcu_{ij} + u_{4j} stp\_math_{ij} + u_{0j} + \varepsilon_{ij}}_{\text{random part}}
\end{aligned} \tag{3.3}
$$

The slopes of the cross-level interactions between the duration of school closures and students' SES γ_{31} and students' previous achievement γ_{41} represent the moderations used to answer RQ3 and RQ4.

We will adjust the models according to the research question regarding the centering strategy of variables, following the recommendations of Enders and Tofighi (2007). We will run all models separately for the two subjects (i.e., reading and mathematics).

Remarks in this section refer to the second study. The key idea of the DiD approach is that, in the absence of treatments, the trends of the dependent variable (here, the national test results) should be comparable between the treated and the untreated group (i.e., school closures, Angrist & Pischke, 2009). Since birth years from a decade ago shall not anticipate the arrival of COVID-19, the underlying student characteristics such as IQs and learning capabilities should remain stable across the cohorts. That is, in the absence of COVID-19, we shall expect no significant differences between the amount of learning growth of the 2020 students and that of the 2019 students. Should such comparability be broken, the "treatment" (i.e., school closure) must carry the explanatory power for such differences. As the validity of the DiD approach relies on the assumption of equal trends, we will use the five periods before COVID-19 to compare $\Delta y_i^{(k+1)-k}$. If the $\Delta y_i^{(k+1)-k}$ before COVID-19 are not different, it is reasonable to assume that $\Delta y_i^{2020-2019}$ would have continued to be the same in the absence of COVID-19. Similarly, we compare whether the trends are equivalent with respect to the different levels of students' SES (i.e., atipcu) and students' previous achievement (e.g., stp_math) before COVID-19.

To gain more insight into students' family situation and household, as well as the circumstances of learning at home, and thus identify significant factors for student achievement, we will explore additional individual-level and school-level variables (see Table 1) using this model Equation (2.1) as a starting point (see section exploratory analyses).

For all our analyses we plan to use R (R Core Team, 2022). If we reach the limits of R during the analyses (e.g., multi-level analyses), then we will also use the statistical program Mplus (Muthén & Muthén, 1998–2017).

Nesting of Data. Study 2 presents data that are hierarchical in nature. Different nesting structures can be considered for our analyses (e.g., students are nested in classes, in schools, or in families), depending on the research questions ("What is of interest?") and on the level at which key independent variables reside (Scott et al., 2013). In our study, school closure duration is the key independent variable, therefore suggesting two possible nesting structures (a) students nested in municipalities, or (b) students nested in schools. Option (a) makes sense as municipalities in Norway are responsible for making policies related to lower secondary schools (Norwegian Ministry of Local Government and Modernisation, n.d.) including COVID directories. As scientific knowledge about COVID accumulates, large municipalities such as the capital Oslo initiated differentiated school closure requirements depending on a colored alert level. By the later stage of the pandemic, it makes better sense to use schools as the Level 2 unit since variation started to emerge within the same municipality. We therefore decide to adopt students nesting in schools for our analyses. We will assess the significance of the nesting structure using intraclass correlations (ICC_1, Lüdtke et al., 2009).

Effect Size

In Study 2, learning growth is operationalized as the difference in national test results between Year 8 and Year 9. We then express effect sizes in percentiles using the SD-based metrics Cohen's d (Cohen, 1988), with all effect sizes being considered informative (see explanations in hypotheses). We expect negative effects of school closures on student achievement. Prior studies signaled a learning loss of approximately $d = -0.10$ SD in Period 1 for reading and mathematics (Hammerstein et al., 2021). We further expect the effect sizes to be small since Norway's national tests measure youth's basic competencies such as numeracy and literacy—skills that should have stabilized by Year 9 and not immediately sensitive to school closures. We will use earlier findings as benchmarks when judging the importance of effect sizes. Betthäuser et al. (2022) found a learning loss of $d = -0.17$, 95% CI $[-0.22,\ -0.13]$ early in the pandemic, equivalent to 42 percent of average learning during a school year in the absence school closures (teachers typically can attain between $d = 0.20$ and $d = 0.40$ per year; Hattie, 2009). Hammerstein et al. (2021) found a learning loss of $d = -0.005\ SD$ to $-0.05\ SD$ per week, which can be interpreted as an average summer learning loss. Based on these findings and the fact that Norway's national tests are general competency-based in contrast to curriculum-based (e.g., in the Netherlands), we expect a small effect size of approximately $d = -0.2/52 = -0.0038\ SD$ per week or smaller for Period 1.

Statistical Power

Since Norway's national register data preserve the entire population with minimal missing values, we expect our study to have sufficiently high statistical power with significance output for all inferential parameters. The challenge following the analyses is to interpret which magnitudes of effect sizes are of practical relevance and are important (see section effect sizes).

Inference Criteria

Classical statistical tests operate based on sampling distributions and are not directly applicable to our studies involving the entire population. We are nevertheless able to employ the conventional Type I error criterion of $\alpha = .05$ and the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) procedure to adjust for multiple comparisons.

Assumption Violation / Model Non-convergence

Variables with severely non-normal distributions such as income will be subject to log transformations in order to enhance normality. Since no latent construct is involved in current studies and our data size is sufficiently large, we do not expect model identification problems or non-convergence risks.

Reliability and Robustness Testing

We will take several measures to verify the robustness of our results.

Study 1

We propose two different methods for handling missing values. The first and primary approach is multiple imputation (see the missing data section). We then wish to corroborate MI with the full information maximum likelihood (FIML) approach (van Buuren, 2018; Graham, 2012) thanks to its ability to produce unbiased estimates once all variables associated with missingness are included in the estimation (Graham, 2012; Schafer & Graham, 2002). In executing FIML for independent and dependent variables, we will include demographic and SES variables in addition to those that are already included in the substantive model (saturated correlates models, Graham, 2003).

Study 2

We would like to first of all implement robustness checks at the variable level. In order to gain deeper insight into different facets of SES (APA, 2017; Avvisati, 2020; O'Connell, 2019), students' social capital can be operationalized using alternative measures such as parents' education or using an index approach similar to that in PISA studies (OECD, 2019). Similarly, students' previous achievement in Norwegian (stp_norw) and mathematics (stp_math) can be corroborated using their attainment record in social (stp_socs) and natural sciences (stp_nats), respectively. In addition, we will use different methods for missing data treatment. Missing data treatment described in Study 1 can also be applied to Study 2.

Next, our robust checks focus on the modelling level. Our central approach (see analyses) is highly influenced by the Engzell et al. (2021) paper. An alternative regression design can be found in Angrist and Pischke (2009) for validating effect estimations (see also Brumback, 2021). Angrist and Pischke's (2009) DiD regression models may include time-varying covariates and are implemented using long data format (see, in particular, pp. 236—239). Furthermore, hierarchical data structure can be re-constructed as students nested in 428 Norwegian municipalities as an alternative to schools. Should the data suggest variations at another level that we are not currently aware of, $ICC_1 \geq .05$ (LeBreton & Senter, 2008), we will include this level in our analyses as another robustness check.

If further opportunities to check the robustness of our results arise in the process of the analyses (e.g., suggested by reviewers), we will carry out these checks if they are reasonable (e.g., the costs do not exceed the benefits of the analyses).

Exploratory Analysis

In principle, we will conduct further exploratory analyses if we can theoretically derive them in a plausible way.

First, we plan to also analyze Period 2. Analogous to Period 1, we will calculate the difference in Year 8 and Year 9 student achievement in the national tests $\Delta y_i^{2021-2020}$. We will explore how Period 2 relates to the other periods prior to COVID-19 school closures and to

Period 1. To do so, we will use the same methodological approaches (i.e., DiD). On the one hand, a cumulative nature of learning can be assumed (Hammerstein et al., 2021; see also Shuell, 1986), which is why potential learning losses could become greater in the long run. In addition, recent research indicate that at least learning deficit early in the pandemic persist over time (Betthäuser et al., 2022). On the other hand, it can be assumed that measures (e.g., government-initiated training programs, providing free digital devices) were introduced after the first school closures. Furthermore, it can be assumed that all stakeholders have become familiar with distance learning. That is students and parents may have known better how to learn at home in the case of later school closures in Period 2 than in Period 1.

Second, we may compare learning progressions across subjects. Previous studies on the effects of school closures on student achievement reported few differences between reading and mathematics (Hammerstein et al., 2021). It is unclear, however, whether similar patterns apply to Norway due to its unique education and assessment systems, such as repeating identical tests one year apart and focusing on general competencies rather than on curricula. One's general literacy and numeracy levels are reasonably expected to have stabilized by Year 8—if differential effects were to happen, we expect them to be larger in mathematics than in reading because numeracy depends more heavily on purposeful training than language skills.

Third, we may explore further variables. That is, we aim to map COVID-19 school closures, the socio-economic status, the family situation, and household of the students that may have been particularly significant for learning at home during school closures as comprehensively as possible. For instance, we may explore additional individual-level (e.g., immigrant status, COVID-19 infection, number of siblings; see Table 1), school-level variables (e.g., school type, proportion of immigrants; see Table 1), or household related variables (e.g., floor space per person in the household, working hours of parents; nr. 15 in Table 1).

In general, if we are advised by experts during manuscript preparation (e.g., by reviewers) to include additional variables that make sense in terms of content and make the findings more robust, and we have these variables available in the dataset, then we will include these variables in our model and check whether the findings remain robust.

References

Angrist, J. D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

APA. (2017). *Report of the APA task force on socioeconomic status*. American Psychological Association. https://www.apa.org/pi/ses/resources/publications/task-force-2006.pdf

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-scale Assessments in Education*, *8*, Article 8. https://doi.org/10.1186/s40536-020-00086-x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://www.jstor.org/stable/2346101

Betthäuser, B. A., Bach-Mortensen, A., & Engzell, P. (2022). A systematic review and meta-analysis of the impact of the COVID-19 pandemic on learning. *SocArXiv*. https://doi.org/10.31235/osf.io/g2wuy

Brumback, B. A. (2021). *Fundamentals of causal inference with R*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003146674

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press. https://stefvanbuuren.name/fimd/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Clark, A. E., Nong, H., Zhu, H., & Zhu, R. (2021). Compensating for academic loss: Online learning and student performance during the COVID-19 pandemic. *China Economic Review*, *68*, Article 101629. https://doi.org/10.1016/j.chieco.2021.101629

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Eisner, N. L., Murray, A. L., Eisner, M., & Ribeaud, D. (2019). A practical guide to the analysis of non-response and attrition in longitudinal research using a real data example. *International Journal of Behavioral Development*, *43*(1), 24–34. https://doi.org/10.1177/0165025418797004

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989x.12.2.121

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, *118*(17), 1–7. https://doi.org/10.1073/pnas.2022376118

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. https://doi.org/10.1207/s15328007sem1001_4

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer. https://doi.org/10.1007/978-1-4614-4018-5

Hammerstein, S., König, C., Dreisörner, T., & Frey, A. (2021). Effects of COVID-19-related school closures on student achievement—A systematic review. *Frontiers in Psychology*, *12*, Article 746289. https://doi.org/10.3389/fpsyg.2021.746289

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

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

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus stuatistical analysis with latent variables: User's guide* (Version 8). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

O'Connell, M. (2019). Is the impact of SES on educational performance overestimated? Evidence from the PISA survey. *Intelligence*, *75*, 41–47. https://doi.org/10.1016/j.intell.2019.04.005

OECD. (2019). *PISA 2018 results (volume II): Where all students can succeed.* https://doi.org/10.1787/b5fd1b8f-en

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989x.7.2.147

Scott, M. A., Simonoff, J. S., & Marx, B. D. (2013). *The SAGE handbook of multilevel modeling*. SAGE.

Shuell, T. J. (1986). Cognitive conceptions of learning. *Review of Educational Research*, *56*(4), 411–436. https://doi.org/10.3102/00346543056004411

Thorn, W., & Vincent-Lancrin, S. (2021). *Schooling during a pandemic*. OECD Publishing. https://doi.org/10.1787/1c78681e-en

van der Velde, M., Sense, F., Spijkers, R., Meeter, M., & van Rijn, H. (2021). Lockdown learning: Changes in online foreign-language study activity and performance of dutch secondary school students during the COVID-19 pandemic. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.712987

Links to Additional Resources and Information

- Statistics Norway
  - National tests
  - Variable explanation (e.g., after-tax income per consumption unit)
- Ministry of Education
  - School Information System
- Norwegian Tax Administration
  - Norwegian identification number
- OECD:
  - Education recovery after COVID-19
  - Building on COVID-19's innovation momentum for digital, inclusive education
  - Schooling disrupted, schooling rethought: How the COVID-19 pandemic is changing education
  - How learning continued during the COVID-19 pandemic
- American Educational Research Association
  - How education fared during the first wave of COVID-19 lockdowns? International evidence