Preregistration

Study Information

Title

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

Authors

Tim Fütterer, Tony Tan, Rolf Vegar Olsen, and Sigrid Blömeke

Keywords

COVID-19, student achievement, social inequality, Norway, SES

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 that allow 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 between different subjects. Lastly, school closures have 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 analyse 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 ($Number\ of\ archival\ entries\ \approx 12.3$ 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 ($Number\ of\ archival\ entries\ \approx 5.6$ million), we will use Year 8 and 9 students' national exam data in reading and mathematics as measures for students' learning. Since identical tests are used in Year 8 and 9, learning changes can be operationalized via 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. That is, the focus of this study is school closure elasticity of student academic achievement (i.e., how responsive was the student academic achievement to a change in the duration of school closures [percentage change of student achievement, when schools were closed for an additional day]). 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.

Study Information

Research Questions*

Our overall research question is: How did pandemic-induced 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:

(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 exam 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 have school closures had on students' learning outcomes?

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

(RQ4) How do low and high achieving students differ in their learning outcomes?

Hypotheses

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

H1 *regarding RQ2: We expect small effect sizes of approximately* $d = -\frac{0.2}{52} = -0.0038\ SD$ *per week that illustrate learning loss period 1.*

Notes: For a justification of hypothesis H1 see the section *effect sizes*. For period 2, we will conduct explorative analyses (see section *exploratory analysis).* We speculate that the effect sizes regarding period 2 are smaller than for period 1 (approximately $d = -\frac{0.1}{52} = -0.0019\ SD$per week). The term *learning loss* means that student achievement does not necessarily decline from Year 8 to Year 9 in periods with school closures, but at least does not increase as much as in periods without school closures. That is, because identical national exams are used in Year 8 and Year 9 (see section *variables*). This means that in years without COVID-19 and thus without school closures, we assume a learning gain when comparing students' achievement (i.e., grades in national exams; see section *variables*).

In addition, we expect to see differential effects of school closures on students' learning.

H2 *regarding RQ3: Effects of school closures on students' learning are moderated by students' SES in period 1. That is, students with lower SES show more learning loss due to school closures than students with higher SES.*

Justification of our hypothesis H2: Previous studies on period 1 suggest that the SES of students is critical for the extent of learning loss. Furthermore, Norway is a country characterized by a high average SES of its students, but at the same time shows a heterogeneity (i.e., range between mean of 5th percentile to mean of 95th percentile) comparable to other countries (with different high SES) at this high level (OECD, 2019). This means that the relationship between SES and school closures can be assumed equally for Norway. The less favorable students' SES, the greater the learning loss.

For period 2, we will conduct explorative analyses (see section *exploratory analysis*), but we expect similar differential effects as for period 1. That is, we assume that students with lower SES show more learning loss due to school closures in period 2 than students with higher SES.

H3 *regarding RQ4: Effects of school closures on students' learning are moderated by students' prior achievement in period 1. Low achievers show more learning loss due to school closures than high achievers.*

Justification of our hypothesis H3: It is well known that low and high achievers differ. For example, it is known that high achievers are better at self-regulated learning. Therefore, we assume that high achievers were less affected by distance learning (e.g., because they have self-regulatory strategies) than low achievers. For period 2, we will conduct explorative analyses (see section *exploratory analysis*), but we expect similar differential effects as for period 1. That is, we assume that low achievers 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 numbers. Under a secured IT environment (TSD), we obtain national statistics on Norwegian residents' education (e.g., person's highest level of education, achievement in national exams), employment (e.g., total expected/agreed working hours per week incl. secondary jobs), income (e.g., after-tax income per consumption unit, EU scale), COVID-19 specific information (e.g., SARS-CoV-2 infection rates), housing/household conditions (e.g., floor space, number of persons living per household), as well as family background (e.g., number of siblings). Most importantly, entries across different datasets within the register can be linked through person ID, 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 information (e.g., teacher-assigned grades, and exam 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 on https://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning and by following instructions on UDIR website https://login.udir.no/LoggInn/logginn respectively.

Data Security

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 (TSD) that fully logs every operation.

Data Identifiers

No persistent, unique identifier of the dataset 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 exam 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 (https://www.ssb.no/en). Information related to school characteristics was managed by the Ministry

of Education and Training (Norwegian Directorate for Education and Training [UDIR]) through the Primary School Information System (GSI; https://gsi.udir.no) database. 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 will be upload to our OSF page (https://osf.io/t6myh/?view_only=85ac0580daf54c44979de1b9ffe0c011).

Variables

Manipulated Variables

Manipulation is not applicable to any unit of analyses in this study due to its archival data design. Neither is blinding or randomization relevant to this study.

Measured Variables*

SSB collects data at the individual level for all individuals who have lived in Norway. Data comprises registers of education and diplomas since 1970 respectively 2000 including national exams since 2007, population registers including information on household conditions and family relations since 1975 respective 2005, employment registers since 2000, wealth and import register since 1993, housing register since 1990 and cash support register since 1999. GSI data comprises information about schools from 1992 through 2021 including information about COVID-19-related restrictions and measures in 2020 and 2021 for all primary schools.

Some key variables are summarized in Table 1. We are going to provide a codebook showing all variables that will be included in our study when all authors have access to the data and when we know the dataset better.

Table 1

Overview of the Variables (or categories of Variables) to be Used in the Studies

Nr	Construct	Level	Variable Name (NO)	Variable Name (EN)	Operationalization [database]	Function (Study)
1	Identifier	1	lopenr_person	idper	Person ID (character)	Matching ID (1, 2)
2	Identifier	2	Skolekommune	scmu	School municipality (character). This variable is used to link students to the municipalities of their schools.	Matching ID (1, 2) Cluster Variable (1, 2)
3	Identifier	2	Løpenummer organisasjonsnummer	idsc	School ID (character)	Matching ID Cluster Variable (1, 2)
4	Identifier	1	Løpenummer far	idfa	Person ID of father (character). This is necessary to associate SES and family background to students.	Matching ID (1, 2)
5	Identifier	1	Løpenummer mor	idmo	Person ID of mother (character). This is necessary to associate SES and family background to students.	Matching ID (1, 2)
6	Identifier	n.u.	Fagkode for fag i grunnskole	idsub	Subject code for subjects in primary and lower secondary school (character). Is necessary to link the grades to the subjects.	Matching ID (1, 2)
7	Student academic achievement	1	Skriftlig eksamenskarakter	e_math e_engw e_norw	Written exams grades for mathematics, Norwegian, and English (ordered categorical; 1 = very low competence, 2 = low competence, 3 = fairly good competence, 4 = good competence, 5 = very good competence, 6 = superior competence, Regulations for the education act, 2006, §3-5). Students are sampled into taking *either* mathematics, written Norwegian, or written English with *equal* probability. Written exams were cancelled between 2020 and 2022 due to COVID-19. Only teacher grades (Variable Nr 9) are available as outcome measures.	Dependent variable (1)
8	Student academic achievement	1	NPLES, NPREG	e_read8 e_math8 e_read9 e_math9	National written exams for are used to evaluate students' reading, arithmetic, and English proficiency. The exams, which are given in the fall of the fifth, eighth, and ninth grades, are a component of formative assessment. We will use data from Year 8 and 9 students. The results of the tests are measured in score points. We will use a scale with which it is possible to describe the development of students' skills and to compare between different tests and years. The average in the scale is set at 50, with standard deviation 10. The scale is based on Item response theory (IRT). Further information on the national exams is given here:	Dependent variable (2)

					https://www.ssb.no/en/utdanning/grunnskoler/statistikk/nasjonale-prover	
9	Student academic achievement	1	Muntlig eksamenskarakter	e_engo e_noro	Oral exam grades for Norwegian and English (ordered categorical; 1 = very low competence, 2 = low competence, 3 = fairly good competence, 4 = good competence, 5 = very good competence, 6 = superior competence, Regulations for the education act, 2006, §3-5). Oral exams consist of not only English and Norwegian, but also mathematics, social and natural sciences, and other electives. Students are randomly assigned to *one* subject only for their oral exams. Oral exams were cancelled between 2020 and 2022.	Dependent variable (1)
10	Student academic achievement	1	Standpunkt	math engw engo norw noro	Teacher-assigned grades for mathematics, written and oral Norwegian, written and oral English (ordered categorical; 1 = very low competence, 2 = low competence, 3 = fairly good competence, 4 = good competence, 5 = very good competence, 6 = superior competence, Regulations for the education act, 2006, §3-5).	Dependent variable (1)
12	Duration of school closures	2	(we are currently not aware of this variable name)	dur	School closure number of days	Independent variable (2)
13	School closures (other variables to get a more detailed insight into school closures.)	2	-	-	e.g., school closure yes/no and how often, teachers on sick leave due to COVID-19, students on sick leave due to COVID-19, students with less special support, students with less language support…. Further information can be obtained, among other things, from the questionnaire used for the school closures in 2021: https://gsi.udir.no/View?show=form&formId=90969&languageId=1&fromApp=false&includeMetaContent=true	Independent variable (2)
14	Students' SES	1	*belopm* Inntekt etter skatt per forbruksenhet (EU-skala)	atipcu	After-tax income per consumption unit (EU-scale) Numeric (Index: Total household taxable and non-taxable income, minus taxes, divided on the number of consumption units in the household. The number of consumption units is calculated by using the 'modified' OECD scale or the EU scale, where the first adult is given a value of 1, any additional adult is given the value of 0.5, and each child is given a value of 0.3. The number of consumption units in a household consisting of two adults and two children is thus 2.1, according to this method https://www.ssb.no/a/metadata/conceptvariable/vardok/3363/en)	Moderator (2, RQ3)

15	Students' SES (other variables that allow tooperationalize SES differently)	1	Personens høyeste utdanningsnivå (Koden NUS2000; 1970, deretter årlig fra og med 1980)	phle	e.g., person's highest level of education (NUS2000 Code; 1970, then annually from 1980). NUS: Norwegian Standard Classification of Education ranging from 0 (no education) to 8 (a research degree). Used to assess the educational level of the parents. If there is information from the mother and father, then the highest value is used. Type of education (by 2000 standard) The Norwegian Standard Classification of Education (NUS2000) includes Norwegian education codes and corresponding codes from The International Standard Classification of Education ISCED. Further information are provided here: https://www.ssb.no/en/utdanning/norwegian-standard-classification-of-education and here https://www.ssb.no/klass/klassifikasjoner/36/	Moderator (2, RQ3)
			Fars høyeste utdanning da personen var 16 år; Mors høyeste utdanning da personen var 16 år		e.g., father's and mother's highest education when the person was 16 years old	
16	Students' family background and situation at home (likely related to stundents' SES)	1	-	-	e.g., refugee background (flyktningbakgrunn; dichotomous [0 = no refugee background, 1 = refugee background]), Immigration category (innvandringskategori [dichotomous (0 = no immigrant, 1 = immigrant)]), number of persons per household (per_18plus_i_hushnr), floor space in square metres of household, new variable (floor space per person living in the household), number of a student's siblings, and any other variables that may have been relevant to learning at home during school closings.	Moderator (2, RQ3) or Control variables (2)
17	Student academic achievement	1	Skalapoeng	tg1 tg2	Term grades, Term 1 and Term 2. Term 1 runs between August and January, Term 2 runs between February and July. All subjects attract a final grade at end end of Term 1 and Term 2. Numeric (ordered categorical; 1 = very low competence, 2 = low competence, 3 = fairly good competence , 4 = good competence, 5 = very good competence, 6 = superior competence, Regulations for the education act, 2006, §3-5). This variable is used in the second study to identify low achieving and high achieving students in the class directly before the phase of school closures.	Maybe moderator (2, RQ4)
18	Students sex	1	Kjønn	sex	Dichotomous (0 = boy, 1 = girl)	Control variable (1, 2) Description of students (1, 2)

19	Students age[a]	1	-	age	Calculated from the month (Fødselsår og -måned) and year (Fødselsår) of birth	Covariate (1, 2) Description of students (1, 2)
20	Country of birth	1	Fødeland	cob	Nominal (e.g., 1 = Norway)	Description of students (1, 2)
21	School type	2	UTD	types	School type	Description of students (1, 2) Control Variable (2)
Variables that may be used (e.g., to merge data or to describe individuals)						
22	Identifier	n.u.	lopenr_familienr/lopenr_husholdning	idfam/idhus	Family/household running number (cha)	Matching ID
23	Identifier	2	Løpenummer organisasjonsnummer	idsc	School ID (character)	Matching ID
24	Identifier	2	Bostedskommune (per 1. oktober)	idmu	Municipality of residence ID (character; October 1st)	Matching ID
25	Identifier	n.u.	lopenr_familienr/lopenr_husholdning	idfam/idhus	Family/household running number (cha)	Matching ID

Note. Because we do not yet have a complete overview of all variables in the data set, we may still add single variables to the study, for example, to better describe the students or to more fully answer our research questions. Further information about many variables available: https://www.ssb.no/a/metadata/conceptvariable/vardok/3363/en; https://gsi.udir.no/.
NO = Norwegian, EN = English, n.u. = not used.

Regarding study2

We will create the dichotomous variables *condition1* and *condition2*, which will encode whether the analyzed 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). In addition, we will use the variables *tg* to compute the new variable *achiever*, which dichotomously encodes which student is a low-achiever (=0) and which is a high-achiever (=1). For this, we will calculate the arithmetic mean of the last two term grades separately for the subjects Norwegian (*mtgn*) and in mathematics (*mtgm*) available before the Year 8 national exam grades. Low achiever: mtgn respectively mtgm ≤ 2; high achiever mtgn respectively mtgm ≥ 5.

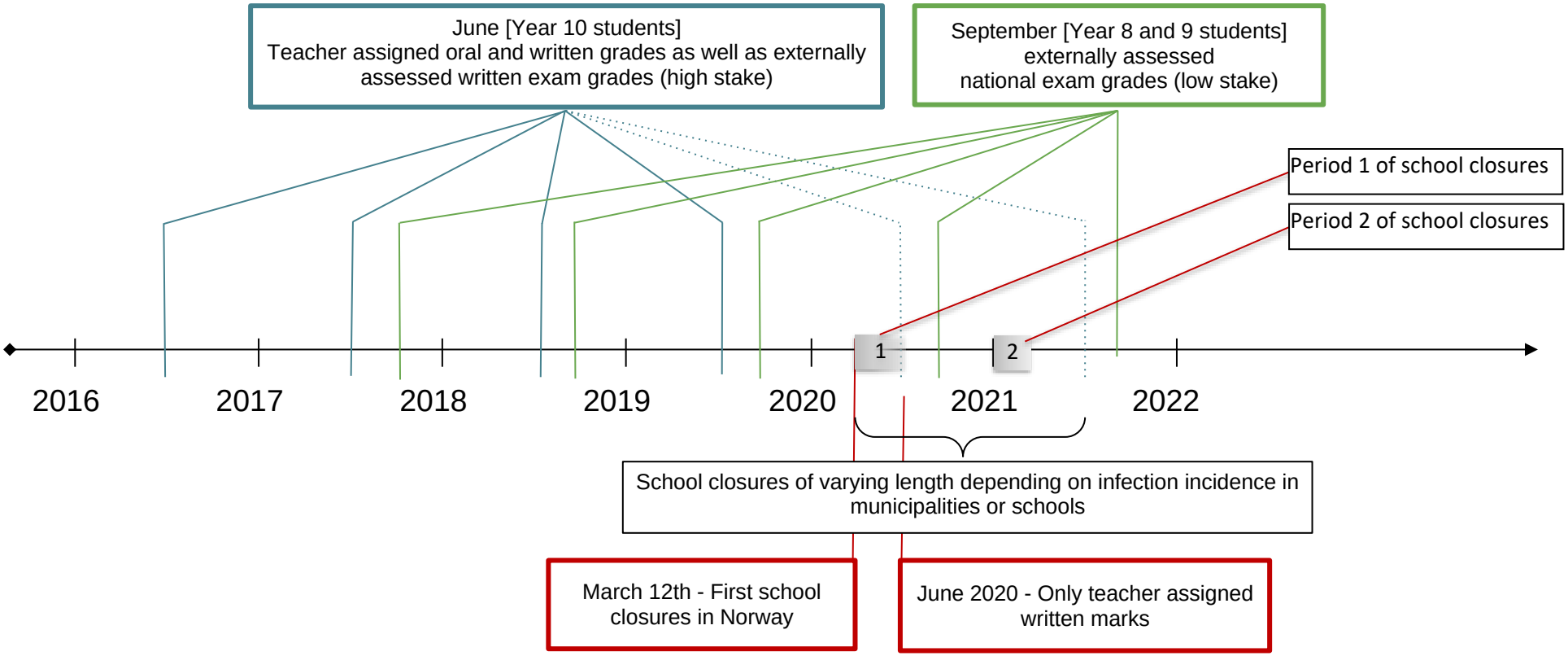
Our key variables to answer our research questions and to test the hypotheses are as follows:

1. The dependent variables are students achievement in reading and mathematics (scores) measured by national exams (nr. 8 in Table 1).
2. We will use the new 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. 12 in Table 1) as our main independent variable to answer research qustion 2.
4. We will use after-tax income (i.e., tax free salary) per consumption unit (nr. 14 in Table 1) as our main moderator representing student's SES to answer research question 3.
5. We will use the new computed variables *mtgn* and *mtgm* as our main moderators representing student's prior achievement to answer research question 4.

To get an overview of where the key variables should be placed over time, Figure 1 illustrates the chronological order of student achievement measures and school closures.

Figure 1

Overview of Available Student Achievement Data and School Closure Periods



Note. Dotted lines indicate that this data was not collected due to school closures. First school closures: 12 March 2020 for six weeks, then irregular opening according to warning levels (e.g., yellow and red flag). From 15 May 2020 at the latest, all schools should have reopened. Second school closures: Varying school closings at the school level, January 4, 2021 through March 12, 2021.

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, after which, students have the freedom to continue into either vocational (*yrkesfaglig opplæring*) or academic (*studieforberedende opplæring*) streams based on their academic performance (*grunnskolepoeng*, grade point average [GPA]) as well as personal interests. Resultantly, academic achievement data for Year 10 are highly regular thanks to the large number of common subjects and contain minimum missing data as a result of Year 10's high stake in determining one's education directions (i.e., vocational or academic). Based on our current knowledge of the data, we know that we will exclude a few students from our analyses. We will delete cases with special pathways like adults who want to come back to the educational system and to finish their grunnskole (e.g., ENGV). We know the amount of these students is about 0.7%. In addition, we will drop students that are not at regular schools and thus do not follow the standardized schooling system (e.g., alternative schools that have an exemption and students from those schools are not required to take testing or that follow an alternative curriculum, e.g., ENGM). We know the amount of these students is about 0.3%. Due to such peculiarities, we estimate that we will exclude one to two percent of the students from the analyses.

In study 2, we will focus on Year 9 students as these students are completing the national literacy (i.e., reading) and numeracy (i.e., calculating) tests for the second time after Year 8.

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 high stake exams in mathematics, Norwegian, and English (⅔ of a column missing; missing completely at random [MCAR]). Sporadic missing were "small scale missings" due to non-response or non-recording of some 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), 50 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 high stake test grades in 2020 and 2021 because the tests at the end of Year 10 were cancelled due to school closures. Dealing with these missing values is the core of research question 1 in study 1. Our idea is, to use a Bayesian approach to compensate for the absence of these values (see analyses section).

In the second study, we face sporadic missing (MCAR, missing at random [MAR] or missing not at random [MNAR]). In addition, we will have to deal with missing values in the dependent variable of students who were exempted from taking the national exams. Although the national exams are obligatory for all students in Year 8 and 9 (and also in Year 5 which is not used in this study) each school can give exemptions according to specified conditions (for further information see: https://www.ssb.no/en/utdanning/grunnskoler/statistikk/nasjonale-prover). We know that the exemption has sharply increased by between 30% and 40% over the past five years (i.e., MAR or MNAR). We suspect that there is a relationship between approved exemptions from participation in the national exams and disadvantaged students (e.g., visible in SES). If such a relationship is shown, this is meaningful for our study, because we need to make sure that we do not get a bias in the results. We would certainly get a bias in this case if we would apply casewise deletion as a measure to deal with these missing values.

To check if there are differences between students who have been exempted from national exams (drop-out group) and students who have not been exempted (non-drop-out group) regarding key background variables like SES measures (i.e., moderator and control variables) and to be able to describe such differences, we will use *t*-tests. Regardless of whether the groups differ, we will include students from both the drop-out and the non-drop-out group in the analyses. As current research shows that approaches in which missing values are handled with MI (van Buuren, 2018) provide unbiased estimates if all variables associated with missingness are included in the estimation (Graham, 2012; van Buuren, 2018), we will use MI analogous to the first study to treat missing values of the drop-out group. We will make sure that these variables for which differences between the drop-out group and non-drop-out group are evident are used as predictors of missing values to be imputed. We also use MI to treat sporadic missing.

We will conduct additional approaches to treat missing values (see section of robustness testing).

Statistical Outliers

In principle, we define outliers as values that are three standard deviations above or below the mean. However, we will check on a case-by-case basis whether a value can be considered valid or not. We will use the example variables *income* and *floor area* to illustrate how we will treat outliers in individual cases: Income information, data points falling 3 standard deviations away from their means are marked for additional examination. Unusually low-income figures are inconsistent with Norway's social safety net and will therefore be considered outliers whereas high income will be retained as meaningful financial records. Invalid or impossible entries such as negative floor areas will also be

marked as outliers. That is, valid outliers will be retained. Non-valid outliers are treated as missing values.

Sampling Weights

The Norwegian register represents the entire population, rather than a sample. Sampling weights and stratified sampling design does not apply to this study.

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*

By the time of the submission of this preregistration, the second author has begun to settle the infrastructure. That is, he currently prepares the data so that it can be used for research purposes (e.g., transformation from a long to a wide format). Further, the second author has created an overview of the variables including descriptive statistics (e.g., means and standard deviations, number of missing values) to allow for strategic selection of variables to answer research questions. However, we still do not have a full overview of all the data available and the structure of all variables. All co-authors do not have access to the data until the preregistration is submitted (i.e., no access to the TSD technically exists). Based on data from the Norwegian Directorate for Education and Training (UDIR, 2021), we know that student's achievement based on national exams in reading and calculation show no major differences on average between 2019 and 2021 or across different genders, immigration backgrounds or parents' educational level. We also know, that the exemption (i.e., the children who are exempt from the national exams, although the tests are actually compulsory for all pupils) has increased by between 30 and 40% over the past five years. Beyond this information, we have no further prior knowledge about the data. In particular, we have no prior knowledge about the relationships of variables needed to answer our research questions.

Analyses

Statistical Models*

Regarding study 1

Bayesian Approach

Regarding our first research question (Study 1), we propose a Bayesian procedure to estimate students' achievement data that went missing due to exam cancellation. Our idea is the following: educational assessment maps students' learning outcomes (L) to a numerical scale (M). A valid and reliable assessment inventory should show high degrees of agreement between L and M, that is, $P(M|L)$ close to 1 ("given the learning, how likely these marks appear"). Society, however, is more interested in knowing "given the marks, how likely there is learning". These two questions can be linked via the Bayes formula (Formula 1):

Formula 1

$$P(L|M) = \frac{P(M|L)P(L)}{normalising\ constant},$$

where $P(L)$ 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 $P(M|L)$can be ascertained from earlier years' exam papers given exam papers' reusability and stability. The normalising constant can be approximated using Markov chain Monte Carlo (MCMC) methods.

Regarding 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 both student, schools and municipality level. In a second step, we will test whether the schools differ descriptively and statistically significantly in key variables (e.g., SES measures) over the last ten years (i.e., 2009 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 more years between 2009 and 2019.

Attrition Analyses (Logistic Regressions)

In a second step, we assume that the proportion of students excluded from the national exams has increased during the pandemic. Furthermore, it is a reasonable assumption that this exclusion had effects on the average of students' academic achievement (e.g., the most disadvantaged students may have been excluded to protect them from unfair testing; Eisner et al., 2019). Therefore, to assess if the missing is dependent on key independent and moderator variables used in this study, we will conduct an attrition analysis for both subjects reading and mathematics. To do this, we will create a dichotomous variable (attr; 0 = *participation in national academic achievement exam*, 1 = *no participation in national academic achievement exam*) and predict *attr* using the independent variables (e.g., school closures) and moderators (e.g., SES, low/high-achievers) as predictors (logistic regressions).

Difference-in-Difference Approach

We will analyse students' learning progression (i.e., gain or loss) between 2019 (students' achievement in national exams [reading and mathematics] in Year 8) and 2020 (students' achievement in national exams [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 exams are measured in score points (https://www.ssb.no/en/utdanning/grunnskoler/statistikk/nasjonale-prover).

First, we will calculate differences between students' national exam 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 of a student i in the national exams 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 1st school closures 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. Formula 2 shows our basic model, that we will use to test our hypotheses (if necessary, additional individual-level and school-level control variables will be included here).

Formula 2

Level 1

$$\Delta y_{ij} = \beta_{0j} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} + \beta_{3j} \text{atipcu}_{ij} + \beta_{4j} achiever_{ij} + \varepsilon_{ij},$$

$i, j \in \mathbb{N}$, i representing student i, j representing school j and with $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. An index ij references students within schools.

Level 2

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

$$\beta_{3j} = \gamma_{30} + \gamma_{31} dur_j + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41} dur_j + u_{4j}$$

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

Total composite formula

$$\begin{aligned}\Delta y_{ij} &= \gamma_{00} + \gamma_{01} dur_j + u_{0j} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} + (\gamma_{30} + \gamma_{31} dur_j + u_{3j}) \text{atipcu}_{ij} \\ &\quad + (\gamma_{40} + \gamma_{41} dur_j + u_{4j}) achiever_{ij} + \varepsilon_{ij}\end{aligned}$$

$$\begin{aligned}= \gamma_{00} + \underbrace{\gamma_{01} dur_j}_{\text{RQ2}} + \beta_{1j} sex_{ij} + \beta_{2j} age_{ij} + \underbrace{\gamma_{30} \text{atipcu}_{ij} + \gamma_{31} dur_j \text{atipcu}_{ij}}_{\text{cross-level interaction (RQ3)}} + \gamma_{40} achiever_{ij} \\ + \underbrace{\gamma_{41} dur_j achiever_{ij}}_{\text{cross-level interaction (RQ4)}} + \underbrace{u_{3j} \text{atipcu}_{ij} + u_{4j} achiever_{ij} + u_{0j} + \varepsilon_{ij}}_{\text{random part}}\end{aligned}$$

The regression weight γ_{01} is of interest to answer research question 2, the regression weights of the cross-level interactions between the duration of school closures and students' SES γ_{31} and status of achiever γ_{41} represent the moderations used to answer research questions 3 and 4. As our predictor of main interest (i.e., duration of school closures) is located at level 2, we will apply grand mean centering (Enders, 2013). Regarding level 1 predictors, we will follow centering recommendations of Enders (2013). We will run all models separately for the two subjects (i.e., reading and mathematics).

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 (Formula 2) as a starting point (see section exploratory analyses).

For all our analyses we plan to use R (R Core Team, 2021). 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

We will face a nested data structure. Different nesting structures of the data could be considered in the analyses (e.g., students nested in classes, nested in schools, or students nested in families). The levels of nesting that should be considered in the analyses are derived from the research questions (what is of interest?) and according to which level the key independent variables (that explains variance in the dependent variable) are to be located (Simonoff et al., 2013). In our study, the duration of school closures is the key independent variable. With respect to this variable, two central nesting structures are conceivable: 1. students nested in municipalities, 2. students nested in schools. Municipalities as level 2 makes sense as municipalities in Norway are responsible for lower secondary schools (Norwegian Ministry of Local Government and Modernisation) and therefore for the regulations of the school closures during COVID-19 (e.g., period [i.e., time point and duration]; treatment of interest). For this reason, municipalities would have to be used as level 2. However, for Oslo municipality, for example, we know that at some point Oslo municipality made differential decisions about school closures within Oslo by regions. Therefore, schools as level 2 also makes sense. as it is likely that schools show variance in our independent variables of interest. We will use the nesting structure students in schools for the analyses. To classify the importance of the nesting structure for the results of the analyses, we will evaluate the intraclass correlation coefficient (ICC[1]; Lüdtke et al., 2009).

Effect Size

The following comments refer to the second study. The key idea of difference-in-difference approaches is that trends of a key dependent variable (in this study student achievement in national exams) would be the same in an untreated and treated group in the absence of the treatment (i.e., school closures; Angrist & Pischke, 2009). In this study, we assume that there is a learning gain without treatment. We assume that this trend is weakened by school closures. We will use effect sizes to quantify trends in students' achievement (i.e., differences in national exam grades between Year 8 and Year9). We will express effect sizes on the scale of percentiles using the standard-deviation based

metrics Cohen's d (Cohen, 1988). In general, we consider all effect sizes to be informative (see explanations of the hypotheses). Based on prior studies, we expect negative effects of school closures in period 1 on student achievement. On average, across several studies from different countries, learning losses, for instance, of $d = -0.10\ SD$ have been shown for the period 1 (i.e., after the first 7 to 8 weeks of the first school closures in 2020) for both reading and mathematics. However, the size of the expected effect in this study is not easy to estimate as the national exams target basic skills such as reading or numeracy. These skills should have stabilized by Year 9 students and should not be as sensitive to school closures as curriculum-based tests. In general, we will be guided by previous findings on the effects of school closures as benchmarks when judging the importance of effect sizes regarding learning gain or loss. Betthäuser et al. (2022) found a learning loss of $d = -0.17$, $95\ c.i.[-0.22, -0.13]$ early in the pandemic (i.e., for school closures in 2020) which can be interpreted as 42% of average learning during a school year without school closures (teachers typically can attain between d = 0.20 to d = 0.40 growth per year; Hattie, 2009). Hammersteint 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 considering that the tests in this study (i.e., national exams) are not likely to be as sensitive as curriculum-based tests, we expect small effect sizes of approximately $d = -\frac{0.2}{52} = -0.0038\ SD$ per week of learning loss for period 1.

Statistical Power

As we use data from the entire population, in which only isolated students will have missing values, the statistical power will be high enough that all of our statistics will be statistically significant. Thus, the challenge following the analyses is to interpret which magnitudes of effect sizes are of practical relevance and, accordingly, are important (see section on effect sizes).

Inference Criteria

As we examine the entire population, classical statistical inference tests and the inference cut-off values are rather obsolete. However, as soon as we get into the situation of making inferences about the population, we will use the standard $p < .05$ criteria in all our analyses for determining if our tests indicate evidence against the null hypothesis. The Benjamini-Hochberg (1995) correction procedure adjusts for multiple comparisons.

Assumption Violation/ Model Non-convergence

If the data violate the assumptions (e.g., no normal distribution), then we will take established measures to deal with these violations (e.g., logarithmise). In principle, we do not assume that our models will not converge because, for example, we do not model latent constructs and thus sufficient information is put into the models to estimate the parameters and our models will be identified.

Reliability and Robustness testing

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

Regarding study 1

We will use two different methods to deal with missing values (i.e., missing by design and sporadic missing). The first approach, which will also be the one we primarily use for our analyses, is multiple imputation (see *missing data* section). The second approach will be the Full Information Maximum Likelihood (FIML) approach (Graham, 2012; van Buuren, 2018). We will use FIML because research shows that approaches in which missing values are handled with FIML provide unbiased estimates if all variables associated with missingness are included in the estimation (Graham, 2012; Schafer & Graham, 2002). We will use FIML for both independent and dependent variables. Furthermore, we will include auxiliary variables (e.g., age or variables on which students who were exempt from national exams and students who were not exempt differ) in the FIML estimation (saturated correlates models: Graham, 2003) unless these variables are already included in the model as control variables.

Regarding study 2

All of the following comments on robustness analyses with respect to the second study are approaches that we may use to estimate the robustness of our results.

First, we will use different measures of school closures (see *measured variables* section).

Second, we will use different measures to operationalise SES (see *measured variables* section).

Third, we will use three different methods to deal with missing values. The first and second approaches are identical to those described for study 1. As a third approach, we will validate results on subsamples (i.e., casewise deletion) and include only those students from the non-drop-out group (i.e., students who were not exempted from the national exams) in the analyses. If students who were exempted and students who were not exempted differ on key variables (e.g., SES), we would not be surprised if the results differ using the casewise deletion approach.

Fourth, our main independent variable is the duration of school closures. Our dependent variable is student achievement. We assume that the duration of school closures is a causal effect. However, to support this assumption we are thinking of using instrumental variables (IV) methods (Imbens & Rubin, 2015). Currently, we have not found a good instrument yet. We considered the *infection rates of students* as an instrument. It is very likely that infection rates have been the justification for policy decisions on school closures in Norway. Therefore, we assume that infection rates are strongly correlated with the duration of school closures. At the same time, individual student learning and ultimately achievement should be potentially related to individual COVID-19 disease but not with the mean infection rates at the municipality level. However, if a mean infection rate is high, more teachers are infected as well. In addition, class composition varies more due to ever on-going infections among students which again disturbs the whole instruction process. For this reason, infection rates may not be an optimal instrument. Whether we can apply an IV approach ultimately depends on whether we are able to find a valid instrument and have the data for this instrument (e.g., student infection rates). When we know the data better, then we will consider whether there are additional or better tools, if any, so that we can test the presumed causal relationship between the duration of school closures and student achievement. If there is no suitable instrument, we will not use an IV approach.

Fifth, we will also conduct two-level models using municipalities for level 2: ~60.000 students (level 1) nested in 428 municipalities (level 2). If we find that the treatment shows (also) variance at another level we are not aware of right now, then we will use this level in our analyses as another robustness check if ICC(1) ≥ .05 (LeBreton & Senter, 2008).

Sixth, we will conduct a within-family design, which discards all variation between families by introducing a separate intercept for each group of siblings identified in our data.

Seventh, we will use propensity score matching and possibly entropy weighting approaches to match treatment and control groups on a wider range of individual-level and school-level characteristics. Doing this, we will use, for instance, sex, parental education, prior performance, interactions between them, and school-level covariates like school-type or proportion of immigrants. Propensity of treatment weights involves first estimating the probability of treatment using a binary response (logit) model and then reweighting observations so that they are balanced on this propensity across comparison and treatment groups. The entropy balancing procedure instead uses maximum-entropy.

Eight, 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).

Ninth, we will use different operationalizations for students' SES and students' prior achievement. SES can be conceptualised with several alternative approaches (American Psychological Association, Task Force on Socioeconomic Status, 2007; Avvisati, 2020; O'Connell, 2019). Therefore, we plan to use different measures to assess SES, to gain deeper insight into the importance of different facets of SES for the impact of school closures on learning achievement. That is, if possible, we will also use other approaches like the index of economic, social and cultural status (ESCS; e.g., nr. 15 in Table 1) as it is used in PISA studies (OECD, 2019). For the operationalization of X, first, we will not only use the last two term grades for Norwegian and mathematics, but we will use all available term grades from relevant subjectsof the last two years to calculate mean values of prior achievement (GPA). In addition, we will follow the idea of Engzell et al. (2021) who constructed a variable to assess prior achievement from all test results in the previous year. Engzell et al. created a composite score and split this into low, middle, and high achievers. In our study, as an additional approach, we will use the national exam grades in Year 8 to operationalize low and high achievers.

Exploratory Analysis

In principle, we will conduct further exploratory analyses if we can theoretically derive them in a plausible way. For instance, we may also test curvilinear associations (i.e., quadratic models) to examine whether there is an optimal duration of school closures for student achievement. In the manuscript, all such further analyses—if they are carried out—will be explicitly marked as exploratory analyses.

The most important exploratory analyses relate to the analyses on period 2. Analogous to period 1, we will calculate the difference in Year 8 and Year 9 student performance on the national exams: $\Delta y_i^{2021-2020}$ (difference score regarding 2nd school closures in 2021). We will explore how period 2 relates to the other periods prior to COVID-19 school closures and to period 1 in particular. To do so, we will use the same methodological approaches (i.e., DiD). In contrast to period 1, there are hardly any robust findings on the long-term effects of school closures on student achievement, making it difficult to estimate the expected effect sizes. 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. On the other hand, it can be assumed that measures (e.g., government-initiated training programs) were introduced after the first school closures. In addition, it can be assumed that all stakeholders have become familiar with distance learning (i.e., students and parents may know better how to learn at home in the case of later school closures than in the case of the first school closures).

As recent research indicate that at least learning deficit early in the pandemic persist over time (Betthäuser et al., 2022) and guided by an average learning gain within a school year of $d = +0.20$ on average as a lower bound, assume effect sizes of approximately $d = -\frac{0.1}{52} = -0.0019\ SD$ per week for period 2.

A second important exploratory approach will be the comparison between subjects. Previous studies on the effects of school closures on student achievement have not identified (or at least little) significant differences between reading and mathematics. Thus, we do not expect to see differences in between reading and mathematics. However, in our study, basic skills such as reading and numeracy are tested in Year 8 and Year 9 using identical tests. Reading and numeracy are skills that should have already stabilized in Year 8. If we do see differences in effects between subjects, we would expect them to be larger in mathematics than in reading because numeracy is a skill that requires more frequent training.

A third important exploratory approach is the inclusion of additional variables in the analyses. That is, we aim to map COVID-19 school closures and the socio-economic status (SES) of the students as comprehensively as possible. That is, we will use the duration of school closures (nr. 12 in Table 1) as our main independent variable and after-tax income (i.e., tax free salary) per consumption unit (nr. 14 in Table 1) as our main moderator representing student's SES. However, to gain more detailed insights, we will look at different variables that are suitable to measure school closures (nr. 13 in Table 1) that can plausibly be assumed to be related to student achievement. Furthermore, we will use different variables to capture family background that may have been particularly significant for learning at home during school closures (e.g., floor space per person in the household, working hours of parents, number of siblings; nr. 16 in Table 1). Therefore, 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 (e.g., immigrant status, COVID-19 infection, number of siblings; see Table 1) and school-level variables (e.g., school type, proportion of immigrants; see Table 1) using this model (Formula 2) as a starting point (see section exploratory analyses). Which variables can be included together in models depends, among other things, on the multicollinearity of the variables. Basically, we will try to include the variables of a content area (e.g., regarding students SES) together and to include as many variables as possible (i.e., as long as the model converges and no multicollinearity shows up). If multicollinearity is shown, we will also include variables separately in several models.

In addition, for instance, teacher assigned grades are available for 2020 and 2021 for more subjects than mathematics, Norwegian, and English. Therefore, we might explore these variables at these two

measurement points and also their changes over time before and during COVID-19 for languages (i.e., Norwegian and English; oral and written grades), sciences (i.e., mathematics and natural sciences; written grades), and arts (i.e., social sciences and religion; written grades). In terms of different subjects, we are not aware of any study that analysed such a broad set of subjects. This means that there are only isolated findings on the impact of school closures on student achievement for subjects other than mathematics. The findings on subjects other than mathematics are mixed. For example, van der Velde et al. (2021) found a positive effect of $d = +9.25\ SD$ for French, and Clark et al. (2021) found a general negative effect of $d = -0.22\ SD$ across subjects such as mathematics, English, politics, and history (for an overview see, for instance, Hammerstein et al., 2021). As positive effects of school closures have tended to be shown in studies in which a specific online learning methods or online-learning software were relevant (e.g., Clark et al., 2021), we assume a negative effect of school closures on student achievement in all subjects based on the numerous negative effects in previous studies.

Furthermore, we assume that students' gender may be related to their learning during school closures. Therefore, we might analyse differences in students' learning between male and female students.

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

American Psychological Association, Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on socioeconomic status*. American Psychological Association.

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

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-Scale Assessments in Education*, *8*(1), 8. https://doi.org/10/gpqhgb

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.2307/2346101

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* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/g2wuy

Brumback, B. A. (2021). *Fundamentals of causal inference with R* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003146674

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*, 101629. https://doi.org/10.1016/j.chieco.2021.101629

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

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/ggk7qm

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), e2022376118. 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/c2zf2d

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*, 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.

Imbens, G., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

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/frx8kx

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/bjc3vf

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/gmhw9s

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

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/bhx874

Shuell, T. J. (1986). Cognitive conceptions of learning. *Review of Educational Research*, *56*(4), 411–436. https://doi.org/10/bz957v

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

Thorn, W., & Vincent-Lancrin, S. (2021). *Schooling during a pandemic: The experience and outcomes of schoolchildren during the first round of COVID-19 lockdowns*. OECD Publishing. https://doi.org/10.1787/1c78681e-en

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

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*, 712987. https://doi.org/10.3389/feduc.2021.712987

Wolf, B., & Harbatkin, E. (2022). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, 1–28. https://doi.org/10/gqkw55

Important Links for Further Information

- SSB: https://www.ssb.no/en/utdanning/grunnskoler/statistikk/nasjonale-prover
- SSB: https://www.ssb.no/a/metadata/conceptvariable/vardok/3363/en
- GSI: https://gsi.udir.no/
- https://www.skatteetaten.no/en/person/foreign/norwegian-identification-number/
- OECD: https://oecdedutoday.com/education-recovery-after-covid/
- OECD: https://www.oecd-ilibrary.org/education/building-on-covid-19-s-innovation-momentum-for-digital-inclusive-education_24202496-en
- OECD: https://read.oecd-ilibrary.org/view/?ref=133_133390-1rtuknc0hi&title=Schooling-disrupted-schooling-rethought-How-the-Covid-19-pandemic-is-changing-education
- OECD: https://www.oecd-ilibrary.org/sites/bbeca162-en/index.html?itemId=/content/publication/bbeca162-en
- AERA: https://www.aera.net/Events-Meetings/How-Education-Fared-During-the-First-Wave-of-COVID-19-Lockdowns-International-Evidence