Exploring the Short-Term and Long-Term Effects of the COVID-19 School Closures on Reading Fluency in Southern Germany: A Longitudinal Causal Effects Analysis

Elisabeth Barbara Kraus[1], Johannes Wild [2], Anita Schilcher[2], and Sven Hilbert[3]

[1] Chair for Computational Modeling in Psychology; LMU Munich; Germany

[2] Chair for Didactics of the German Language and Literature; University of Regensburg; Germany

[3] Chair for Educational Data Science; University of Regensburg; Germany

Author Note

CRediT: Elisabeth Kraus: Conceptualization, Methodology, Formal analysis, Data curation, Writing-Original Draft preparation, Johannes Wild: Data curation, Writing-Original draft preparation, Project administration, Anita Schilcher: Writing- Reviewing and Editing, Funding acquisition, Project lead, Sven Hilbert: Writing- Reviewing and Editing, Supervision.

Elisabeth Barbara Kraus, ORCID: 0000-0001-8007-0321. Johannes Wild, ORCID: 0000-0001-7731-1614. Sven Hilbert, ORCID: 0000-0001-5808-8357.

We have no conflict of interest to disclose. Correspondence concerning this article should be addressed to: Elisabeth Kraus, LMU Munich, Akademiestr. 7, 80799 München, Germany, Email: e.kraus@psy.lmu.de

Abstract

The coronavirus-related school closures in Bavaria, as in Germany as a whole, lasted from March to June 2020 and were therefore on an average scale internationally. This study investigates the impact of these closures on reading fluency in elementary students. To evaluate the short-term (one month) and long-term (one year) effects of third-grade school closures on reading fluency, we use a longitudinal causal effects approach. The study involved 9,083 students in Bavaria, divided into two cohorts: one affected by third-grade closures, and a control cohort who experienced closures in second grade one year earlier. Data were collected over five measurement points, using the Salzburger Lesescreening (SLS) to assess reading fluency. The study employed a Difference-in-Differences approach with propensity score weighting to estimate causal effects, accounting for confounders like participation in a reading intervention, gender, educational background, and migration status. Pre-COVID data and pre-trend analysis on the study sample indicated a linear increase in SLS raw scores over time. The affected cohort showed no significant deviations from this trend during the pandemic, suggesting neither short-term nor long-term impacts on reading fluency. However, variations in teaching efforts and other COVID-related effects could have influenced these results. In conclusion, findings indicate that school closures due to COVID-19 have not impacted reading fluency as one example of a consolidated ability, in third grade students in Southern Germany.

Keywords: reading fluency, COVID-19, school closures, causal analysis, elementary school, Germany

Exploring the Short-Term and Long-Term Effects of the COVID-19 School Closures on Reading Fluency in Southern Germany: A Longitudinal Causal Effects Analysis

As governments worldwide imposed lock-downs and social distancing measures, over 1.6 billion learners in more than 190 countries faced school closures at the peak of the crisis in April 2020 (Oecd, 2020). To curb the coronavirus' spread, elementary schools in Bavaria, Germany, were closed from March to June 2020. Teachers, students, and parents were largely unprepared (König et al., 2020), especially regarding digital tools for distance learning. Initial estimates foresaw significant performance losses and long-term repercussions (Bao et al., 2020). Meta-analyses and literature reviews generally support this notion but reveal diverse outcomes depending on countries, domains, the time perspective, and socioeconomic factors (Donnelly & Patrinos, 2022). The overall picture, thus, remains inconclusive and influenced by various factors, such as country and study design - preventing to causally relate school closures to learning losses. Causal effects necessitate longitudinal studies with robust and comprehensive control conditions (Stuart Mill, 1843), reducing the impact of confounding influences on estimated effects as much as possible. This study strives to fulfill these requirements and employs the Rubin causal model (Rubin, 2005) to assess the causal influence of school closures on reading fluency in elementary students.

Background

To provide a broad overview of the pandemic's effects on student learning, we summarize findings from national and international research starting with meta-analyses and reviews that involve different domains and grades and ending with a more specific look at reading, especially reading fluency in elementary schools.

Global Overview of COVID-related Learning Losses

Assuming a broad perspective, the overall picture on learning losses after COVID-related school closures reveals significant and substantive learning losses across various countries and domains. Hammerstein et al. (2021) reviewed studies across eight countries and found that most reported negative effects of COVID-19-related school closures on student achievement, with a median effect size of approximately -0.10 standard deviations (SD) for mathematics and reading. Another meta-analysis by König and Frey, 2022 also found a robust negative average effect of -0.175 SD, indicating even larger significant learning losses across various subjects including reading, mathematics, and the social sciences. An extensive review by Betthäuser et al., 2023 across 15 countries reported an average negative effect of -0.14 SD. Similar effects were found by multiple other authors, such as Zierer, 2021 (-0.11 to -0.17 SD), Engzell et al., 2021 (-0.08 to -0.14 SD), Di Pietro, 2023 (-0.17 SD), Storey and Zhang, 2021 (-0.15 SD), and Donnelly and Patrinos, 2022 (-0.05 to -0.29 SD). However, effect sizes varied substantially between countries. Additionally, the duration of school closures varied significantly between countries, making it challenging to accurately compare the true influence of school closures on learning losses across different contexts.

Country-specific Variations

In some countries, schools were closed for as few as eight weeks (e.g., Australia), while others experienced closures extending beyond 34 weeks (e.g., Brazil, India) (see Zalewska et al., 2023, for an overview of the duration of school closures in Europe). Consequently, a more detailed country-specific analysis is necessary when assessing the effects of school closures. The difference in school closure durations directly influenced effective learning time. Several reviews and meta-analyses have indicated that effective learning time in distance learning environments was significantly reduced compared to regular school lessons (Crosson & Silverman, 2022; Lockl et al., 2021; Pilonieta et al., 2023; Schult et al., 2022b), and it is widely accepted that less effective learning time is associated with slower learning progress (Everaert et al., 2017).

The impact of school closure duration on learning loss is reflected in comparative studies such as Kennedy and Strietholt (2023) and international comparisons. For instance, American, Italian, and Brazilian schools were closed for periods ranging from 4.5 to 10 months (UNESCO, 2022), and studies from these countries consistently report substantial learning losses (Alves et al., 2022; Contini et al., 2022; Kuhfeld, Soland, & Lewis, 2022; Kuhfeld, Soland, Lewis, et al., 2022). In contrast, schools in Australia were closed for only eight weeks, and schools in Sweden were mostly not closed at all. Reported learning losses in these countries were very low or absent (Gore et al., 2021; Hallin et al., 2022).

Moreover, different countries adopted varied strategies to cope with school closures. Generally, European countries seem to have managed the school closures better than non-European countries, as indicated by generally lower learning losses (compare e.g., Ardington et al., 2021; Birkelund & Karlson, 2023; Depping et al., 2021). Similarly, Australia, where distance learning had been more established compared to Europe, reported no significant impact on reading performance and mathematics among primary school students (Gore et al., 2021). Therefore, considering both the geographic region and the duration of school closures is essential when interpreting the effects of school closures on learning losses.

Demographics-related Variations

Next to the origin of the study, also the background of the studied population played into the effects of school closures on learning losses. Independent of the geographic origin of the studies, socio-economic status (SES), socio-cultural capital, gender, and sometimes race moderated influences of the school closures on learning losses. As one example, Domingue et al., 2022 observed slower growth in reading fluency in US-districts with a higher proportion of Black students in grades one and two. Under the assumption that race, SES, and educational support by parents were confounded in the samples studied, these findings align with those from Bao et al. (2020), which suggest that daily reading aloud at home could mitigate about 42 percent of the potential loss in reading fluency. For the Netherlands, Oostdam et al., 2024 showed that the effects of the SES at the individual level were attenuated by negative SES effects at the school level, with only small differences in achievement scores between cohorts. In line with this, the impact of school closures on reading comprehension varied significantly with SES. Students from higher-SES backgrounds generally fared significantly better than those from lower-SES backgrounds (Depping et al., 2021; Engzell et al., 2021; Hammerstein et al., 2021; Kuhfeld, Soland, & Lewis, 2022; Segers et al., 2023). Also in Germany, learning losses in reading comprehension were weakly associated with the students' socio-cultural capital (Schult et al., 2022b). Using Progress in International Reading Literacy Study (PIRLS) data, Ludewig et al., 2022 found significant negative differential effects for gender (higher losses for female students), migration background, socio-economic status, and need for special education. Finally, in South Africa, Ardington et al., 2021 found that learning loss was between 57 to 81 percent of a school year, with girls being disproportionately negatively affected in word reading compared to boys.

Domain-specific Variations

Apart from geographic and demographic influences, learning is manifold and different domains revealed differential effects. Most studies either tackled learning losses in mathematics and/or reading ability. The general picture seems to be that learning losses were slightly more pronounced in mathematics compared to reading. For example, Donnelly and Patrinos (2022) reviewed studies across seven countries and revealed negative effects for reading, ranging from -0.06 to -0.10 SD, while those for mathematics were slightly stronger on average, ranging from -0.05 to -0.19 SD. In line with this, Betthäuser et al. (2023) showed in a review of 42 studies across 15 countries that the average learning loss was more pronounced in mathematics (-0.18 SD) than in reading (-0.09 SD). Higher losses in reading compared to mathematics are in accordance with findings from further original studies, such as Borgonovi and Ferrara, 2023 (Italy), Engzell et al., 2021 (Netherlands), Tomasik et al., 2021 (Switzerland), Haelermans et al., 2022, and Oostdam et al., 2024 (both Netherlands). In contrast, a review by Hammerstein et al., 2021 with 11 studies across eight countries found almost identical effects for reading (-0.09 SD) and mathematics (-0.10 SD). All in all, effects on reading seemed to be more stable compared to mathematics (Haelermans et al., 2022; Oostdam et al., 2024). This was also supported by findings from Miller et al., 2023 as well as Depping et al., 2021 who even reported positive effects for mathematics. One possible reason for higher effect stability of reading could be the homogeneity of the construct. While mathematics is composed of various sub-disciplines such as calculus, geometry, analysis, or probability theory, reading is a more uniform construct.

Short-term vs. Long-term Effects

Moreover, the temporal distance to the school closures influenced the measurable learning loss. In the short-term, learning losses were in general evident and quantifiable. Studies across various contexts consistently found that students experienced setbacks in their academic progress, immediately after school closures. On average, these short-term learning losses accounted to approximately -0.1 standard deviations. However, the situation appeared less severe when researchers investigated long-term effects. Skar et al., 2023 provided evidence of a positive trend in Norway for the domain of writing. Their research suggested that over time, the negative impacts of school closures diminished, and in some cases, students showed signs of recovery or even improvement. Several other studies noted a rebound in academic achievement as well. Oostdam et al. (2024) found that the negative impact on reading skills observed during the first lockdown was substantially mitigated by positive recovery effects following the second lockdown in the Netherlands. Schult et al., 2022b corroborated these results for Germany, indicating that the pattern of initial decline followed by recovery was observed in other contexts as well.

However, the overall picture remains complex and requires further exploration. Variations in recovery patterns across different countries and contexts suggest that multiple factors, including the duration and quality of remote learning, access to resources, and socio-economic disparities, play critical roles in shaping the long-term effects of school closures (Halloran et al., 2021; Learning, 2021; Schult et al., 2022b). This complexity is reflected in the diverse outcomes reported in various studies, underscoring the need for continued research to comprehensively understand the full extent and nature of the impact of school closures on learning losses.

Mitigation of Variations

So far, we have established that the geographic origin of the population studied, their socio-demographic characteristics, the domain studied, and the temporal distance to the closures moderate the impact that school closures had on learning losses. In order to confine the effects of school closures as much as possible and to obtain high internal validity, these influences need to be mitigated. This can be achieved on various levels, such as restricting the study design or through statistical control for possible confounders. In the following we will present our study and show how we approached the mitigation of the possible confounding influences discussed above.

The Current Study

Our current study was part of the larger project XXX. This project comprised two cohorts of elementary school students between grades two and four (first cohort) and between grades three and four (second cohort). Of these students, one sub-sample participated in systematic reading interventions while the other sub-sample had regular reading lessons. The reading interventions were teacher-led and designed to train reading fluency during the second grade, reading comprehension in third grade, and meta-cognitive competencies in the context of reading comprehension in fourth grade. Reading progress was monitored in terms of reading fluency, reading comprehension, and orthography (see, Birkel, 2007; Kraus et al., accepted; Wimmer & Mayringer, 2016). As an additional part of the study, some teachers regularly provided written feedback on different aspects of their reading instruction (Kraus et al., 2020). The study was conducted in Bavaria, Germany, between 2018 and 2023.

By restricting the study sample to a geographically and age-related homogeneous sample from Southern Germany, we minimized the influence of different school closure durations and coping abilities of different countries. With respect to the time perspective, we consider short-term and long-term effects separately. Finally, in terms of the domain, we restricted ourselves to reading fluency, a readily measurable and well-defined construct.

Reading Fluency

Reading fluency is one of the three sub-processes of reading: word reading (decoding), reading fluency, and reading comprehension (Cadime et al., 2017). Decoding can be simplified to the accuracy of reading and refers to the ability to translate (groups of) letters into sounds (Lai et al., 2014; Silverman et al., 2013). Reading fluency is commonly defined through the automated reading of letters, words and sentences (Silverman et al., 2013). Finally, reading comprehension simplified means building a mental representation of the text upon a readers knowledge (Kintsch, 1998). The three sub-processes or sub-facets build on each other and are therefore considered to partly be prerequisites to each other (Lai et al., 2014; Psyridou et al., 2023). While learning to decode is usually completed after the first years of reading lessons, reading fluency continues to develop until late secondary school (van de Ven et al., 2017). Indeed, reading fluency develops quite homogeneously across a large period of time with an almost linear increase between grades two and six, and shows a slow flattening out towards the final level between grades six and nine (Gärtner, 2010; Wimmer & Mayringer, 2016). While some authors divide reading fluency into reading accuracy, reading speed and prosody (e.g. Kuhn et al., 2010), we stick to the more narrow definition of reading fluency as reading speed, as proposed by Silverman et al. (2013) or Fuchs et al. (2001). Therefore, studying the effects of school closures on reading fluency yields the advantage of having a theoretically well-founded, narrowly defined, and homogeneously developing construct.

School Closure Effects on Reading Fluency

The fact that reading fluency is developing during elementary school and therefore susceptible to disruptions in the learning process, was also apparent in former analyses on the effects of school closures on reading fluency. Various European studies observed a significant learning loss or reduced learning progress in reading fluency, especially in younger students (Baschenis et al., 2021; Förster et al., 2023; Nandi et al., 2023, e.g.). Also, Starling-Alves et al., 2023 found a significant decline in fluency scores with the highest loss in grade two. Alves et al., 2022 reported a significant loss of reading fluency in second grade as well, but not in higher grades. Overall, Molnár and Hermann, 2023 reported an average learning loss of 0.25 SD, with lower primary school students particularly affected. Only few studies found that either older students suffered from greater loss than younger students Kuhfeld, Soland, and Lewis, 2022; Lerkkanen et al., 2023 or reported non-significant or positive effects on reading fluency (Hallin et al., 2022; Richter et al., 2022). Finally, a small sample study in Portugal by Rosendo et al., 2023 found no statistically significant differences in accuracy, speed, and reading fluency performance, but significant differences in perceived fluency quality. To investigate whether these finding could be backed up by a causal analysis, we employed the Rubin causal model to analyze the effects of school closures in Southern German third graders on reading fluency.

The Rubin Causal Model

The Rubin model (Rubin, 2005) is a formal mathematical framework for causal inference and relies on potential outcomes under treatment and control, defining a causal effect as the difference between observed and counterfactual outcomes. We aim to limit the range of these counterfactual outcomes by considering an extensive set of confounding variables such as gender, educational background, migration status, participation in a systematic intervention, and diverging pre-trends that could influence reading fluency development. At the same time, our findings are expected to have adequate external validity because our sample is large, demographically representative, and collected in an applied setting with regular teacher-led instruction (the reading intervention originally studied is now highly recommended by the Bavarian ministry of education as a curricular element in reading instruction).

Summary, Rationale, and Research Questions

The body of research reviewed indicates that the COVID-19 pandemic led to significant learning losses, particularly in mathematics and reading, with younger students and those from disadvantaged backgrounds being the most affected. However, there was also evidence that certain interventions, such as reading aloud at home, could mitigate potential losses. While reading was adversely affected by the COVID-19 pandemic, the specific impacts varied widely based on factors such as grade level, SES, and geographical location. Reading fluency tended to be affected for younger students at the elementary school level. Most studies so far looked at short-term effects of school closures, for that it is intriguing to additionally gauge the long-term effects of COVID-related school closures.

Therefore, our study aims to compare the reading fluency development of a cohort of third graders affected by school closures with an expected development in the absence of school closures. Due to the design of our study, this expected development can be derived from a trend analysis of pre-COVID cohorts, conducted with the same population and the same measurement instrument, refined through in-depth comparison with a large matched post-COVID control cohort from the same study. This approach is seldom applied, because it requires large sample sizes and high data quality (Furquim et al., 2019). Both requirements may be assumed to be met by our sample as well as the pre-COVID cohorts. Our key research questions investigate (1) the short-term (after one month) and (2) long-term (after one year) effects of third-grade school closures on reading fluency.

Method

Sample

The sample of the main analysis of this study consisted of $n = 9,083$ elementary students in Bavaria, Germany, divided into two cohorts. The affected cohort experienced school closures during third grade, the other during second grade, before entering the study (control cohort). Students underwent an average of about five measurements, totaling 38,088 data points. Table 1 gives an overview of the actual composition of the sample, conditional on the time points with respect to school grades.

Table 1

Sample sizes conditional on measurement time points

grade	begin_2	middle_2	end2_begin_3	middle_3	end_3	middle_4	end_4	total
COVID cohort	8009	6640	6334	2206	3722	6348	885	34,144
control cohort	-	-	847	755	788	778	776	3,944
total	8009	6640	7181	2961	4510	7126	1661	38,088

In terms of demographics, our sample consisted of students who predominantly had no migration status ($n_{no\ mig} = 5322$). Yet, more than a third of the students had a second generation migration status ($n_{mig\ parents} = 3499$), and very few children migrated to Germany themselves ($n_{mig\ child} = 82$, $n_{NA} = 180$). Most children spoke German at home ($n_{just\ german} = 5894$) about a third spoke German and another language at home ($n_{german\ \&\ other} = 2377$), and only few students did not speak German at home ($n_{no\ german} = 783$, $n_{NA} = 29$). The sample was about balanced in terms of gender ($n_{girls} = 4481$, $n_{boys} = 4466$, $n_{NA} = 136$). The educational background was expressed in the estimated amount of books at the students' homes (see, Sieben & Lechner, 2019). Most students reported to have three or more multi-level bookshelves at home ($n = 2670$). Fewer students reported to have either two multi-level bookshelves ($n = 1831$) or one multi-level bookshelf ($n = 2334$) and even fewer students reported to have but one single shelf of books ($n = 1440$). A minority reported to own no or very few books ($n = 653$).

Our sample was complemented with norm data from two versions of the same standardized reading fluency test Salzburger Lesescreening (SLS; English: Salzburg Reading Screening) SLS1-4, (Mayringer & Wimmer, 2003) and SLS 2-9 (Wimmer & Mayringer, 2016), and a control cohort of low-performers from Peters et al. 2021. The norm sample of the SLS 1-4 (Mayringer & Wimmer, 2003) comprised a total of $n_{norm1-4} = 1867$ students, with 215 up to 295 students per measurement time point. Their data were collected in Austria and Bavaria, Germany. The norm sample of the SLS 2-9 (Wimmer & Mayringer, 2016) comprised a total sample of $n_{norm2-9} = 4166$ students, with more than 1000 students per measurement time point (Gärtner, 2010). Their data were collected in Austria. The low-performers' samples stemmed from six individual studies (Förster & Souvignier, 2011, 2014, 2015; Förster et al., 2018; Hebbecker & Souvignier, 2018; Peters et al., 2022), and comprised a total of $n_{low} = 1346$ students, with the smallest sample comprising 122 and the largest sample comprising 287 low performers.

Design and Procedure

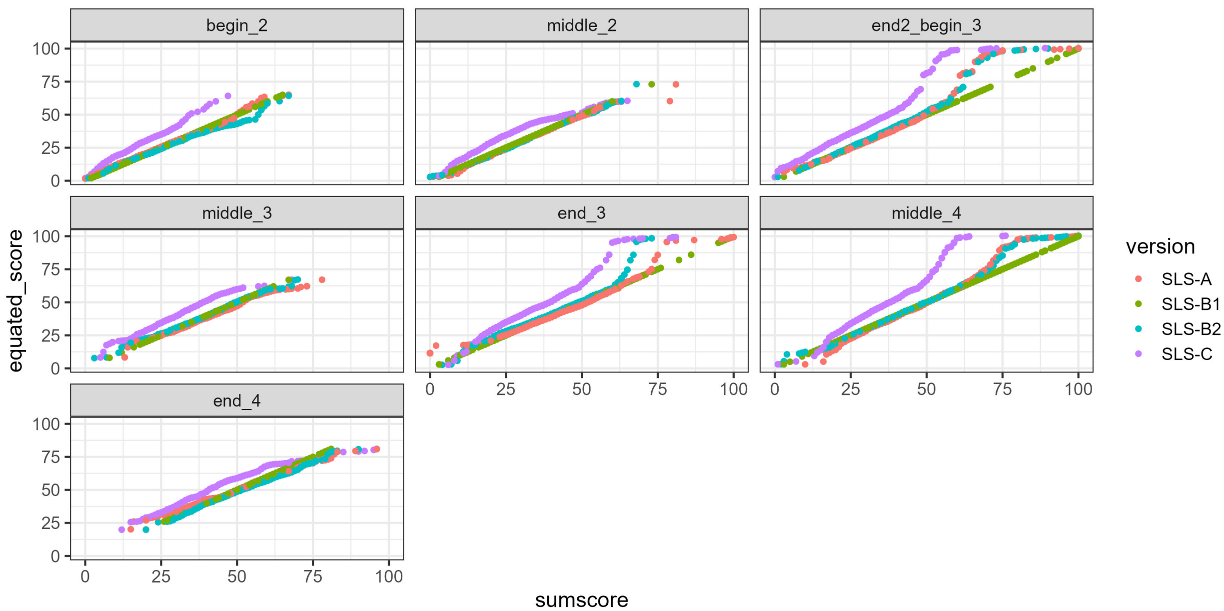
Our sample was divided into two cohorts: the COVID cohort, which was affected by the school closures during the study, and the control cohort, which was not. Measurements took place every six months. The affected cohort had seven measurement time points: four before school closures (starting in 2018) and three during the pandemic (March, 2020 - July, 2021). Assessments started in grade two. The control cohort had five measurement time points, starting after the pandemic related school closures in October 2021, with the first assessment in grade three. Only trained teachers or university staff carried out all assessments to ensure data quality. The norm sample data of the SLS 1-4 and SLS 2-9 were collected as part of the test development during the school year of 2007/2008. The data fom the low-performers from (Peters et al., 2021) were collected as parts of quasi-experimental pretest-posttest design intervention studies with a duration of at least one school year. Their data were collected between 2000 and 2018.

Measures

Demographic variables such as gender, family language(s), migration status, and number of books at home as a proxy for SES and educational background (Heppt et al., 2022) were available for our Bavarian sample and were assessed with an additional questionnaire administered when students entered the study. Reading fluency was assessed using the SLS 2-9, (Wimmer & Mayringer, 2016) . The SLS operationalizes reading fluency by having students indicate the substantive correctness of as many sentences as possible out of a maximum of 100 sentences, during a three-minute period (e.g. "Trees can talk."/ German: "Bäume können sprechen." vs. "Water is wet."/ German: "Wasser ist nass."). It is a validated four-version test with appropriate psychometric properties, such as a convergent validity with reading comprehension of $r = .43$ (Wild et al., 2022) and an internal consistency varying between $\alpha = .95$ (Kraus et al., n.d.) and $\alpha = .98, CI_\alpha = [.97, .98]$ (this study). Psychometric modeling indicated metric but not scalar measurement invariance for test versions. We tested metric measurement invariance in the item response theory (IRT) framework by fitting Rasch models (Rasch, 1960) to every test version-specific subpopulation. This ensures that the factor structure (one-dimensional) and the discrimination parameters (set to one) are the same for all subpopulations. We removed items that were not solved in more than 99% of the observed results to speed up calculations. This led to $j = 74$ items of the 100 items being part of the analysis and a scale ranging up to 20 points above the average score of 54 in the last measurement occasion. Rasch models for every test version fit the data, as indicated by non-significant Andersen likelihood ratio tests (Andersen, 1973, all $p_s > .920$). Furthermore, we tested for scalar measurment invariance by comparing multigroup IRT models with constraints on the item difficulty parameters to models without constraints in the difficulty parameters, in which the test versions represented the multiple groups. A subsequent model comparison indicated that unconstrained models fit the data significantly better ($\Delta logLik = -10821.9, \chi^2 = 21643.68, df = 219; p < .001$). We therefore equated the response data for test versions (Kolen & Brennan, 2013). We used equipercentile equating as described by Kolen and Brennan (2004). Figure 1 presents the results of the equating process, thus the mapping functions for our datasets. It shows that version SLS-C was consistently more difficult than the other test versions. This can be read of the equated scores presented on the *y*-axix for version C being higher than the original sum score represented on the *x*-axis. For all other versions we see that differences in difficulty emerge for higher raw scores hitting the amount of 60 sentences. However, the display shows no other consistent pattern across measurement time points.

Figure 1

Equating functions



Analyses

To assess causal effects, a Difference-in-Differences (DiD) approach (Furquim et al., 2019) was applied. The underlying idea of DiD is to compare the changes in outcomes over time between a group that is exposed to a treatment (here: the COVID cohort) and a group that is not (here: the non-COVID cohort or control cohort). In our study, we compare several measures of reading fluency. Whereas for the COVID cohort the school closures fall in between these measurements, there were no school closures for the control cohort in between measurements, as this cohort was assessed one year after the COVID cohort. We then compare the differences with respect to the two groups between the differences with respect to the measurement time points. For the DiDs to be meaningful, the study setup has to meet several assumptions.

One important assumption for DiD to be valid is the parallel assumption. It posits that, in the absence of the treatment, the difference in outcomes between the treatment group and the control group remained constant over time. In other words, any changes in the outcome for the control group over time should mirror the changes that would have occurred for the treatment group if the treatment had not been applied. Since this assumption tackles a counterfactual, it can never be observed. In consequence, there are several factual indicators that the parallel assumption is warranted. Specifically, we used pre-trend estimation (Wing et al., 2018) and propensity score matching (for an exhaustive description, see Austin, 2011) to validate the parallel assumption.

Pre-trend Estimation

For once, the trends in the outcomes for both the treatment and control groups should be similar before the treatment is implemented. In our example, we therefore also consider the reading fluency development over the course of four measurement time points before the COVID-related school closures. This amounts to a pre-trend surveillance of one and a half school years. We also estimated time trends in a separate analysis with data from the same measurement instrument, but different research groups. More specifically, we used the Peters et al. (2021) study data and data from the norm samples of the SLS. After descriptively displaying the six month average reading fluency development in the data, we estimated a hierarchical linear regression on the test raw scores, with time points as a linear predictor and random intercept effects for the data origin (see, Hilbert et al., 2019). Since the Peters et al. (2021) study comprised only students from the lowest 25 percentile of their original samples, this comparison gives us information about the uniformity of the reading fluency development across the performance range.

Propensity Score Estimation

Secondly, confounding influences on the treatment effect could be unevenly distributed between the treated and the control group. In our example socio-demographic variables, such as migration status and educational background, could influence the reading fluency development (Gärtner, 2010). Therefore, we used a matching algorithm to control for socio-demographic influences. We controlled for the fact, whether students received a systematic reading training vs. regular lessons, for gender, for migration status and family language, and for the number of books in the household. By matching the COVID-cohort and the control cohort, we cannot rule out the overall influence of the confounding variables, but we can minimize the differential effects between the COVID-cohort and the control cohort. Therefore, average differences between the groups cannot be attributed to the influence of the confounding variables used in the matching process. To achieve this, we applied propensity score matching. The propensity score itself is defined as the probability of receiving the treatment (here: being affected by school closures) as a function of covariates - in our case the confounding variables discussed above. The propensity scores were estimated using a greedy k-nearest neighbors approach with $k = 5$ and a generalized linear model to estimate the distance. Afterwards, the propensity distances were transformed into weights by inverse probability weighting in the subsequent regressions, as proposed by Stuart et al. (2014).

Causal Analysis

The actual DiD were implemented in the framework of linear regression with robust clustered standard errors which controlled for longitudinal design dependencies. Our regression equation was:

$$Y_{\mathrm{LQ}} = \beta_0 + \beta_1 X_{\mathrm{group}} + f(X_{\mathrm{time}}) + \beta_2 X_{\mathrm{group}} X_{\mathrm{lag0}} + \beta_3 X_{\mathrm{group}} X_{\mathrm{lag1}} + \beta_4 X_{\mathrm{group}} X_{\mathrm{lag2}} + \varepsilon \quad (1)$$

Where:

Y_{LQ} = scaled test value of SLS

β_0 = mean test value of control cohort at the beginning of the study

β_1 = mean difference of affected cohort at the beginning of the study

X_{group} = dummy variable for cohort membership (control vs. affected cohort)

$f(X_{\mathrm{time}})$ = linear time trend derived from trend analysis of pre-COVID data

X_{time} = continuous time variable

β_2 = lagged effect of school closure in the affected cohort

X_{lag0} = dummy variable for the time point one month before school closure

β_3 = direct effect of school closure in the affected cohort

X_{lag1} = dummy variable for the time point a month after school closure

β_4 = lagged effect of school closure in the affected cohort

X_{lag2} = dummy variable for the time point a year after school closure

ε = residuum

It was set up according to the following considerations: Firstly, we transformed our dataset into long format, with each observation reflecting a person at a particular time point. Y_{LQ} represents the values of the reading fluency scores observed for every person at the individual time points, resulting in multiple values of Y_{LQ} per person. A time-dependent trend for reading fluency development is included through the term $f(X_{\mathrm{time}})$. An additional regression weight β_1

models average group differences across all time points. β_2, the estimator of the interaction between the group and the time point before the school closures X_{lag0} indicates the pre-treatment development (i.e., the pre-school closure differences between the two groups). Finally, the parameters β_3 and β_4 represent the deviation from the otherwise expected development of the COVID cohort one month (β_3) and one year (β_4) after the school closures.

To derive the DiD estimates, we compared the estimates of β_3 and β_4 to the estimate of β_2. This difference expresses the difference in differences between the groups between the pre-school closure and the post-school closure time points. We tested the null hypotheses $\beta_3 = \beta_2$ and $\beta_4 = \beta_2$.

To approach the term $f(X_{time})$ we based the functional form (e.g. linear vs. non-linear) of the time trend on the time trend derived from the pre-COVID data Peters et al. (i.e., from the SLS norm samples and 2021). Further, we double-checked the derived time trend by descriptive statistics and displayed the group-specific reading fluency development of our sample graphically (see Figure 2). We then modeled $f(X_{time})$ in two different ways: In one model, we used individual time points as dummy-coded variables. In a second model we assumed a linear time trend and estimated β_{time} for a continuous variable representing the measurement time points in consecutive order (begin of 2^{nd} grade = 0, middle of 2^{nd} grade = 1, and so forth).

The interaction effects for the group variable with lag0 (before the school closures), with lag1 (right after the school closures) and with lag2 (one year after the school closures) were represented by additional dummy-coded variables for these specific time points. As described above, these allowed us to estimate possible derivations from the otherwise assumed constant time trend across groups. Finally, we allowed for correlated ε_i whenever reading fluency values belonged to the same person. This accounted for the nested data structure of measurements being clustered inside persons.

Software

All analyses were conducted using the statistical software R (R.4.3.3) (R Core Team et al., 2024). Propensity scores were estimated using the package *MatchIt*, (Stuart et al., 2011). We further used the packages: *mirt* (Chalmers, 2012), *eRm* (Mair et al., 2010), and *psych* (William Revelle, 2023) for psychometric analyses. The package *equate* (Albano, 2016), was used for the equipercentile test equating, and the packages , *lmerTest* (Kuznetsova et al., 2017) and *MuMIn* (Bartoń, 2023), were used for the regression analysis. We share our analysis code and a synthetic dataset derived from our data at https://osf.io/bu2ad/?view$_{o}nly = 946b100b31764018bf2fadf4fcd67772$.

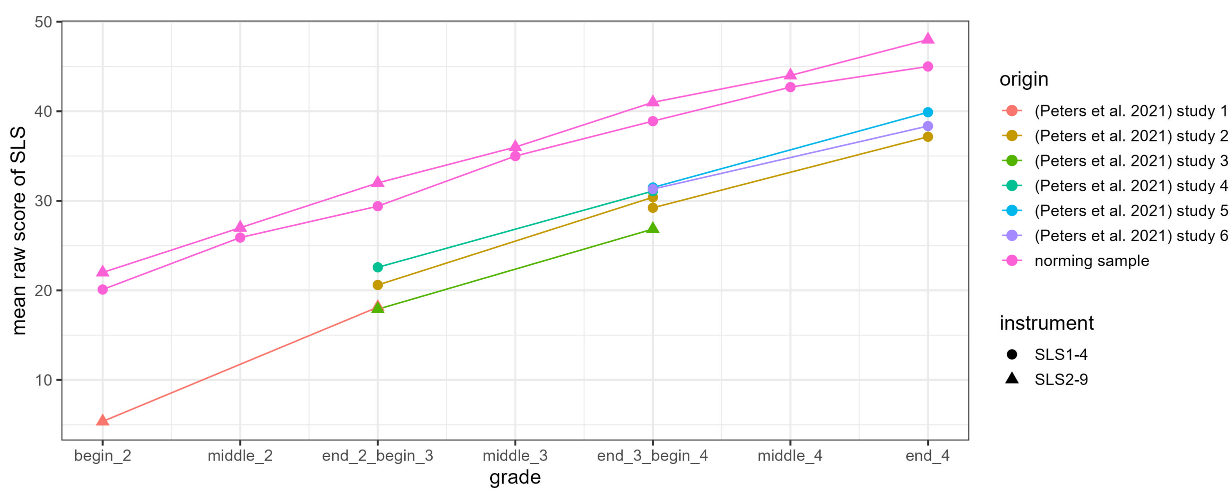
Results

Pre-trend Estimation

Pre-COVID data indicated a linear development of SLS raw scores with an average six month increase of 4.30 raw score units, as displayed in Figure 2. Students read an average of 20 sentences in the beginning of grade two, an average of about 33 at the middle of grade three, and increased their reading fluency to an average of about 46 sentences at the end of grade four.

Figure 2

Trend analysis



The six-month increase did not vary conditional on the initial reading fluency level, so that a uniform development across quantiles could be expected. Neither did the effect size vary with the SLS version (SLS1-4 vs. SLS2-9). A linear time effect and a random intercept effect for data origin explained 98.2% of the variance in mean raw scores (see Figure 2).

Propensity Score Estimation

The *knn* matching was based on the demographic variables gender, migration status, family language, and number of books at home. Additionally, we matched for the participation in our systematic reading intervention. Propensity scores with respect to the membership of the COVID-cohort were derived and used for inverse probability weighting (Austin, 2011). Baseline characteristics for all variables before and after the propensity score weighting are displayed in Table 2

Table 2

Baseline characteristics propensity score variables

	before propensitiy score weighting		after propensity score weighting	
Group	control	COVID	control	COVID
intervention (yes)	$\pi = .88$	$\pi = .77$	$\pi = .88$	$\pi = .88$
gender (girl)	$\pi = .51$	$\pi = .50$	$\pi = .51$	$\pi = .51$
migration status (*)	$\pi = (.47, .50, .01)$	$\pi = (.61, .36, .01)$	$\pi = (.50, .49, .01)$	$\pi = (.50, .49, .01)$
family language (**)	$\pi = (.52, .37, .10)$	$\pi = (.68, .24, .08)$	$\pi = (.54, .35, .10)$	$\pi = (.54, .36, .10)$
number of books (***)	$\pi = (.06, .14, .26, .19, .34)$	$\pi = (.07, .16, .26, .21, .29)$	$\pi = (.06, .14, .25, .13, .43)$	$\pi = (.06, .15, .27, .17, .35)$

Note. π = sample distribution; values do not add up to 100% due to missing values; * levels: none, parents, child; ** levels: German, German & other, other; *** levels: none or few, one single shelf, one multi-level shelf, two multi-level shelves, three or more multi-level shelves

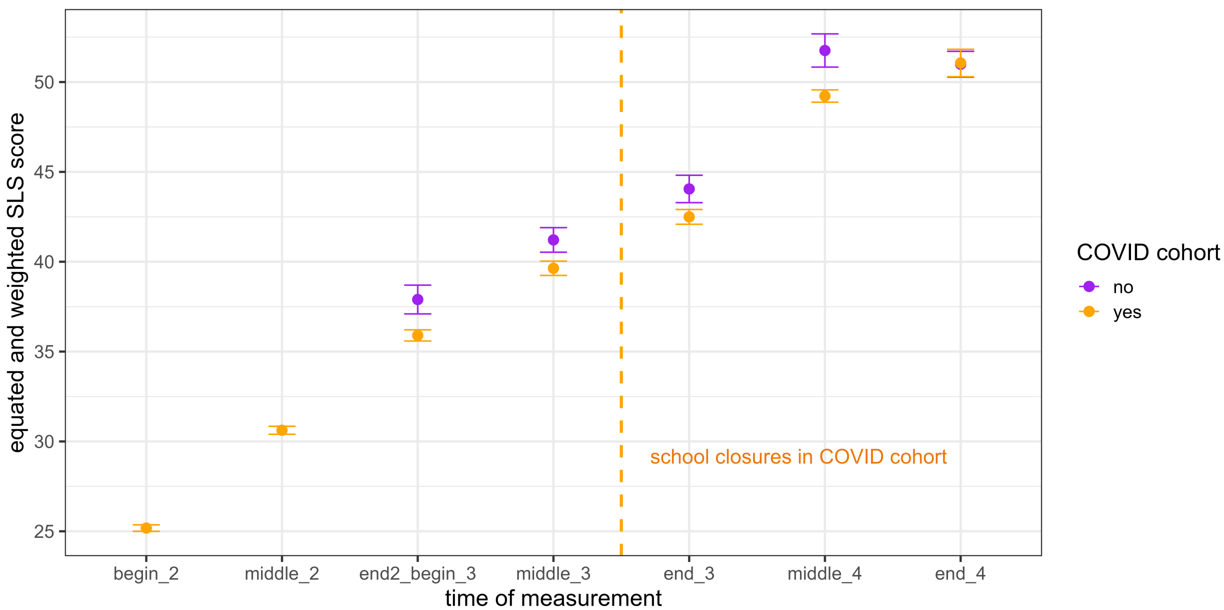
While pre-weighting distributions partly differed on up to 16% for single categories, the post-weighting distributions were very similar. There was only one exception for the variable “number of books” and the category “three or more multi-level shelves”, with a difference of 8% in the category frequency.

Causal Analysis

Visual inspection of the descriptive development in our data did not show any apparent loss in reading fluency development in the COVID cohort during school closures (see Figure 3).

Figure 3

Descriptive Development of COVID and control cohort



Note. Brackets indicate 95% confidence intervals of the displayed group means.

However, we observed some irregularities in reading fluency development between the end of third grade and the end of fourth grade. This led us to investigate, whether modeling time as a discrete rather than as a continuous variable would fit the data better. We therefore compared the fit of two regression models using a likelihood ratio test. The first regression model included a linear time trend, the second regression model a discrete time trend. The results displayed in Table 3 show that the discrete time model fitted the data significantly better, indicated by a significant χ^2-test and several fit indices, such as lower AIC and BIC for the discrete time trend model.

We therefore interpreted the model with a discrete time trend. The parameter estimates of the model are displayed in Table 4.

Our first finding was that our sample had a higher average in SLS scores at the beginning

Table 3

Model comparison between linear time model and discrete time model

Model	npar	AIC	BIC	logLik	deviance	Chisq	Df	p(>Chisq)
linear time trend	8	285162	285231	-142573	285146			
discrete time trend	13	284986	285097	-142480	284960	186.08	5	< .001

of the study with $\beta_0 = 27.66$, compared to the norm sample mean of ($\mu = 22$). The COVID cohort performed significantly lower across the whole study duration, expressed by the significant negative group effect (β_1). The six-month increase in reading fluency varied between a maximum of 7.21 equated scores (between the middle and the end of grade three, expressed by $TP[end3] - TP[middle3]$ and stagnation (between the middle and the end of grade four, expressed by $TP[end4] - TP[middle4]$).

Finally, the regression analysis supported the descriptive finding that school closures had no causal effect on reading fluency. Comparing β_4 and β_3 to β_2, yielded non-significant results ($\chi^2_{\beta_3\beta_2}(1) = 0.0045,\ p_{\beta_3\beta_2} = .947$; $\chi^2_{\beta_4\beta_2}(1) = 1.33,\ p_{\beta_4\beta_2} = .250$). This means that the time trend estimated by $f(X_{time})$ did not vary significantly after the school closures in comparison to the time point before the school closures. Yet, the effect for COVID cohort x lag3 (β_4) was positive and significant. This means, that the COVID cohort caught up to the control cohort one year after the pandemic (end_4). All time and group effects together accounted for 10% of the variance in reading fluency scores. The random effects of the model indicates that inter-individual variance in the intercepts due to the clustered data structure accounted for an additional 9% of the variance in reading fluency scores, as expressed by the difference between the marginal and the conditional R^2.

Table 4

Regression results for equated score

Predictors	Parameter	Estimates	*CI*	*p*-value
(Intercept)	β_0	27.66	26.40 – 28.93	<.001
COVID cohort	β_1	-2.71	-3.98 – -1.44	<.001
TP [middle_2]	$f(X_{time})$	5.58	5.35 – 5.81	<.001
TP [end2_begin_3]	$f(X_{time})$	10.68	10.44 – 10.91	<.001
TP [middle_3]	$f(X_{time})$	13.40	11.38 – 15.42	<.001
TP [end_3]	$f(X_{time})$	16.33	14.35 – 18.32	<.001
TP [middle_4]	$f(X_{time})$	23.54	23.30 – 23.79	<.001
TP [end_4]	$f(X_{time})$	23.22	21.23 – 25.22	<.001
COVID cohort × lag0	β_2	0.76	-1.28 – 2.81	.464
COVID cohort × lag1	β_3	0.85	-1.15 – 2.84	.407
COVID cohort × lag3	β_4	2.18	0.14 – 4.21	.036
Random Effects				
σ^2		614.06		
τ_{00}(student ID)		67.86		
ICC		0.10		
$N_{student\ ID}$		9083		
observations		38088		
marginal R^2 / conditional R^2		.10 / .19		

Note. CI = Confidence Interval; ICC = Intraclass Correlation Coefficient; TP = time point.

Discussion

In this study, we identified the effect of COVID-related school closures on reading fluency during third grade in Germany. We found that there was neither a significant short-term loss (one month after the school closures) nor a significant long-term loss (one year after the school closures). The absence of learning losses was evidenced by consistent time trends between a control cohort and a cohort affected by COVID-related school closures at the respective time points.

Interpretation of Results

Our findings partly contradict several international research findings, which indicated that COVID-related school closures had in fact detrimental effects on reading fluency e.g., Betthäuser et al., 2023 and align with other research results, especially for Germany, such as Depping et al. (2021) and Förster et al. (2023), and Schult et al. (2022b). However, previous research projects primarily employed cross-sectional, observational designs, making it difficult to attribute their reported effects causally to the school closures.

In our study, we adopted a longitudinal approach and applied causal analysis. To define causality in non-experimental settings, it is essential to closely confine the counterfactual development. We therefore had to confine our expected reading fluency development without school closures as closely as possible. First, we adhered to a simple and clear definition of reading fluency, operationalized by the number of sentences read in a standardized reading test within a set time period. Using a normed standardized test instrument allowed us to compare the reading fluency development of our sample to the average expected development from pre-COVID studies using the same test (Peters et al., 2021) and a sample from the test norming procedure. Additionally, reading fluency is assumed to develop rapidly during elementary school, making changes measurable, and to develop homogeneously, following a linear trend until about grade six (Gärtner, 2010). Our pre-trend analysis supported these assumptions and showed a linear half-annual increase in reading fluency, homogeneous across the performance range. The homogeneous linear development over time and the availability of a test covering a wide range of reading fluency allowed us to measure reading fluency consistently with the same instrument from the beginning of grade two to the end of grade four, enabling direct comparisons of test scores across these time points. Finally, the DiD analysis on our sample also indicated the expected linear half-annual increases in reading fluency. In summary, the choice of an ability with a homogeneous development, a narrow definition of the construct reading fluency, and the measurement with a standardized instrument allowed us to control time-dependent trends in our causal analysis. The inclusion of multiple measurement points made it also possible to differentiate between short-term and long-term effects, which were both non-significant and complement each other in the general interpretation of our results.

Another favorable aspect of our research design was the ability to control for factors influencing learning loss during the COVID-19 pandemic. We restricted our sample to one region in Germany (Bavaria), which ensures that the duration of school closures was identical for all students in the sample. We also confined our sample to third-graders, thereby minimizing differential effects for different age groups observed in other studies (e.g., Domingue et al., 2022; Molnár & Hermann, 2023; Starling-Alves et al., 2023).

Since multiple studies have shown that demographic characteristics such as personal student background influence the impact of school closures on learning loss, we used propensity score weighting to balance these influences between our affected and control cohorts. Additionally, as instruction quality affects the development of reading fluency (Schilcher et al., 2022) – and some of our sample had participated in a systematic reading fluency intervention during or before entering the study – we accounted for intervention participation in the propensity score estimation. This led to balanced distributions regarding demographic and instructional variables and limited confounding influences on the effect of school closures on reading fluency in our study.

Limitations Toward Causal Interpretations

Despite our efforts to rigorously control our research design, there are potential threats to a valid causal interpretation of our findings. While past research consistently indicated that breaks from instruction are detrimental to learning (e.g., Schult et al., 2022b), the extent to which COVID-related school closures led to instructional interruptions is unclear. Bavarian schools offered distance relatively early during the closures, which might have been as effective as regular in-person instruction, at least regarding reading fluency. While findings of Schult et al. (2022b) suggest otherwise, we still cannot rule out that teachers focused their reading instruction on reading fluency during COVID school closures because it is relatively easy to implement in a distance learning setting (compared to instruction on reading comprehension). Indeed, previous findings from our sample indicate that some teachers turned to the systematic reading fluency intervention from our study during second grade, as they felt unable to deliver the planned reading comprehension intervention scheduled for third grade (Wild et al., 2022).

Additionally, our control cohort consistently outperformed the COVID cohort even before the school closures. Notably, the control cohort experienced school closures during grade two. This possibly led to increased efforts by teachers and parents to compensate for perceived or actual learning losses from the recent school closures. Between the end of grade three and the middle of grade four, we observed an unusually high increase in reading fluency, followed by the COVID cohort catching up to the control cohort by the end of grade four. This catch-up coincided with stagnation or a slight decrease in the control cohort's reading fluency, which deviates from the expected linear development suggested by norm data and previous studies using the SLS. One possible explanation is that the last measurement in the control cohort occurred just before the summer break, potentially resulting in low motivation for the additional test. Another explanation could be selective dropout (see Depping et al., 2021): The COVID cohort's sample size decreased during the peak of the pandemic. It stabilized after schools closed but drastically dropped again by the end of grade four.

In summary, while we controlled for time trends, demographic variables, and some instructional factors, such as participation in a systematic reading intervention, extraordinary instructional efforts in the control cohort post school closures, and selective dropout pose a caveat to the causal interpretation of our findings. Given the observed deviations from expected development starting after grade three, caution is warranted – particularly regarding the causal interpretation of the long-term effects of school closures on reading fluency.

Implications for External Validity

There are certain particularities of our study that should be considered when generalizing our findings to other abilities, populations, or settings. Firstly, we investigated the development of a skill that was assumed to be internalized and mostly automated before the school closures. Learning losses due to breaks in instruction are particularly detrimental when the learning process is still in an early stage a stable level of proficiency has not yet been attained. If we had examined reading comprehension, which is still developing in third grade, the school closures might have shown a more severe impact on its development. This was also suggested by an earlier study on the same cohort (Wild et al., 2022). Secondly, reading fluency is a skill that can be maintained without extensive in-person instructional effort once a stable level has been reached (Kuhn & Stahl, 2003). The Bavarian State Ministry for education therefore reacted at an early stage and provided parents with instructions and materials for an easy-to-implement reading fluency training at home. Therefore, it is possible that teachers, parents, and even students themselves managed well to compensate for the lack of in-person teaching during the school closures. If so, the strong relationship between reading fluency and parental support in reading (Bus et al., 1995; Oostdam et al., 2024) as well as as the influence of the SES (Depping et al., 2021; Engzell et al., 2021; Hammerstein et al., 2021; Kuhfeld, Soland, & Lewis, 2022; Segers et al., 2023) may have shown even more clearly during the school closures and closed the expected gap.

Conclusion

Our study shows that the widely spread perception that school closures were catastrophic for learning (e.g., Bailey, Duncan, Murnane, & Yeung, 2021) needs to be reconsidered with greater nuance. It is important to keep in mind that we only examined reading fluency during the third grade – this is a specific skill during a specific period and cannot be generalized to the overall effect of school closures on learning. However, at least for skills that have been mostly internalized and automated, such as reading fluency in third graders, the effects of COVID-related school closures may not have been as severe as initially assumed by early predictive studies (e.g. Kuhfeld et al., 2020) or the general public (Howard, 2023). Therefore, future investigations should focus nuanced approaches to disentangle differential effects of the closures.

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1–36. https://doi.org/10.18637/jss.v074.i08

Alves, L. M., Carvalho, I. M., Santos, L. F. D., Ribeiro, G. d. L., Freire, L. d. S. C., Martins-Reis, V. d. O., Labanca, L., & Celeste, L. C. (2022). Reading fluency during the COVID-19 pandemic: A longitudinal and cross-sectional analysis. *Arq. Neuropsiquiatr.*, *80*(10), 994–1003. https://doi.org/10.1055/s-0042-1758446

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*, 123–140. https://doi.org/10.1007/BF02291180

Ardington, C., Wills, G., & Kotze, J. (2021). COVID-19 learning losses: Early grade reading in South Africa. *International Journal of Educational Development*, *86*, 102480. https://doi.org/10.1016/j.ijedudev.2021.102480

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Bailey, D. H., Duncan, G. J., Murnane, R. J., & Yeung, N. A. (2021). Achievement gaps in the wake of covid-19. *Educational Researcher*, *50*(5), 266–275. https://doi.org/10.3102/0013189X211011237

Bao, X., Qu, H., Zhang, R., & Hogan, T. P. (2020). Modeling Reading Ability Gain in Kindergarten Children during COVID-19 School Closures. *International Journal of Environmental Research and Public Health*, *17*(17), 6371. https://doi.org/10.3390/ijerph17176371

Bartoń, K. (2023). *Mumin: Multi-model inference* [R package version 1.47.5]. https://CRAN.R-project.org/package=MuMIn

Baschenis, I. M. C., Farinotti, L., Zavani, E., Grumi, S., Bernasconi, P., Rosso, E., Provenzi, L., Borgatti, R., Termine, C., & Chiappedi, M. (2021). Reading skills of children with dyslexia improved less than expected during the COVID-19 lockdown in italy. *Children (Basel)*, *8*(7), 560. https://doi.org/10.3390/children8070560

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behavior*, *7*(3), 375–385. https://doi.org/10.1038/s41562-022-01506-4

Birkel, P. (2007). *Weingartener grundwortschatz rechtschreib-test für zweite und dritte klassen:(wrt 2+)*. Hogrefe, Verlag für Psychologie.

Birkelund, J. F., & Karlson, K. B. (2023). No evidence of a major learning slide 14 months into the COVID-19 pandemic in Denmark. *European Societies*, *25*(3), 468–488. https://doi.org/10.1080/14616696.2022.2129085

Borgonovi, F., & Ferrara, A. (2023). COVID-19 and inequalities in educational achievement in Italy. *Research in Social Stratification and Mobility*, *83*, 100760. https://doi.org/10.1016/j.rssm.2023.100760

Bus, A. G., van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Rev. Educ. Res.*, *65*(1), 1–21. https://doi.org/10.3102/00346543065001001

Cadime, I., Rodrigues, B., Santos, S., Viana, F. L., Chaves-Sousa, S., do Céu Cosme, M., & Ribeiro, I. (2017). The role of word recognition, oral reading fluency and listening comprehension in the simple view of reading: A study in an intermediate depth orthography. *Reading and writing*, *30*, 591–611.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Contini, D., Di Tommaso, M. L., Muratori, C., Piazzalunga, D., & Schiavon, L. (2022). Who lost the most? mathematics achievement during the COVID-19 pandemic. *B. E. J. Econom. Anal. Policy*, *22*(2), 399–408.

Crosson, A. C., & Silverman, R. D. (2022). Impact of COVID-19 on early literacy instruction for emergent bilinguals. *Read. Res. Q.*, *57*(1), 5–14. https://doi.org/10.1002/rrq.456

Depping, D., Lücken, M., Musekamp, F., & Thonke, F. (2021). Kompetenzstände hamburger schüler*innen vor und während der Corona-Pandemie [competence levels of hamburgs students before and during the corona pandemic. In D. Fickermann & B. Edelstein (Eds.), *Die deutsche schule: Beiheft 17. schule während der Corona-Pandemie. neue ergebnisse und überblick über ein dynamisches forschungsfeld* (pp. 51–79).

Di Pietro, G. (2023). The impact of covid-19 on student achievement: Evidence from a recent meta-analysis. *Educ. Res. Rev.*, *39*(100530), 100530. https://doi.org/10.1016/j.edurev.2023.100530

Domingue, B. W., Dell, M., Lang, D., Silverman, R., Yeatman, J., & Hough, H. (2022). The effect of COVID on oral reading fluency during the 2020–2021 academic year. *AERA Open*, *8*, 233285842211202. https://doi.org/doi.org/10.1177/23328584221120254

Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *PROSPECTS*, *51*(4), 601–609. https://doi.org/10.1007/s11125-021-09582-6

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proc. Natl. Acad. Sci. U. S. A.*, *118*(17), e2022376118. https://doi.org/10.31235/osf.io/ve4z7

Everaert, P., Opdecam, E., & Maussen, S. (2017). The relationship between motivation, learning approaches, academic performance and time spent. *Accounting Education*, *26*(1), 78–107.

Förster, N., Forthmann, B., Back, M. D., & Souvignier, E. (2023). Effects of the COVID-19 pandemic on reading performance of second grade children in germany. *Read. Writ.*, *36*(2), 289–315. https://doi.org/10.17105/SPR44-1.60-75

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short-and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction*, *56*, 98–109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Förster, N., & Souvignier, E. (2011). Curriculum-based measurement: Developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators. *Learning Disabilities: A Contemporary Journal*, *9*(2), 65–88.

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction*, *32*, 91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review*, *44*(1), 60–75.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. In *The role of fluency in reading competence, assessment, and instruction* (pp. 239–256). Routledge.

Furquim, F., Corral, D., & Hillman, N. (2019). A primer for interpreting and designing difference-in-differences studies in higher education research. *Higher education: Handbook of theory and research: Volume 35*, 1–58. https://doi.org/10.1007/978-3-030-31365-4_5

Gärtner, K. (2010). Normierungsstudie zum salzburger lesescreening. *Statistische Nachrichten*, *5*, 383–389.

Gore, J., Fray, L., Miller, A., Harris, J., & Taggart, W. (2021). The impact of COVID-19 on student learning in New South Wales primary schools: An empirical study. *The Australian Educational Researcher*, *48*(4), 605–637. https://doi.org/10.1007/s13384-021-00436-w

Haelermans, C., Korthals, R., Jacobs, M., de Leeuw, S., Vermeulen, S., van Vugt, L., Aarts, B., Prokic-Breuer, T., van der Velden, R., van Wetten, S., & de Wolf, I. (2022). Sharp increase in inequality in education in times of the COVID-19-pandemic. *PLoS One*, *17*(2), e0261114. https://doi.org/10.1371/journal.pone.0261114

Hallin, A. E., Danielsson, H., Nordström, T., & Fälth, L. (2022). No learning loss in Sweden during the pandemic: Evidence from primary school reading assessments. *International Journal of Educational Research*, *114*, 102011. https://doi.org/10.1016/j.ijer.2022.102011

Halloran, C., Jack, R., Okun, J. C., & Oster, E. (2021). *Pandemic schooling mode and student test scores: Evidence from us states* (tech. rep.). National Bureau of Economic Research.

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

Hebbecker, K., & Souvignier, E. (2018). Formatives assessment im leseunterricht der grundschule–implementation und wirksamkeit eines modularen, materialgestützten konzepts. *Zeitschrift für Erziehungswissenschaft*, *4*(21), 735–765. https://doi.org/10.1007/s11618-018-0834-y

Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education*, *25*(4), 903–928. https://doi.org/10.1007/s11218-022-09704-8

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *26*(1).

Howard, J. (2023). Children lost about 35 percent of a normal school year's worth of learning during the pandemic, study suggests [Available at: https://edition.cnn.com/2023/01/30/health/covid-learning-loss-study-wellness/index.html (Accessed: June 27th, 2024)]. *CNN*.

Kennedy, A. I., & Strietholt, R. (2023). School closure policies and student reading achievement: Evidence across countries. *Educ. Assess. Eval. Acc.*, *35*(4), 475–501. https://doi.org/10.31219/osf.io/93rgz

Kintsch, W. (1998). *Comprehension : A paradigm for cognition*. Cambridge University Press. https://worldcat.org/title/1148595541

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices. https://doi.org/10.1007/978-1-4939-0317-7

Kolen, M. J., & Brennan, R. L. (2013). *Test equating: Methods and practices*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4939-0317-7

König, C., & Frey, A. (2022). The impact of COVID-19-related school closures on student achievement - a meta-analysis. *Educational Measurement: Issues and Practice*, *41*(1). https://doi.org/10.1111/emip.12495

König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during covid-19 school closure: Teacher education and teacher competence effects among early career teachers in germany. *European journal of teacher education*, *43*(4), 608–622. https://doi.org/10.1080/02619768.2020.1809650

Kraus, E., Wild, J., Bühner, M., Schilcher, A., & Hilbert, S. (accepted). Applying cross-validated psychometric models to the bavarian reading test (BYLET). *European Journal of Psychological Assessment*.

Kraus, E. B., Pargent, F., Hilbert, S., & Augustin, T. (n.d.). Introducing the treatment decision framework (treadef)–a decision theoretic approach to using evaluation study data to inform individual treatment decisions. *Available at SSRN 4846774*.

Kraus, E. B., Wild, J., Schilcher, A., & Hilbert, S. (2020). Reading related activities in second grade and their effect on fluency-an evaluation of teachers' classroom practices. https://psyarxiv.com/g2f5x/

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test score patterns across three COVID-19-impacted school years. https://doi.org/10.26300/ga82-6v47

Kuhfeld, M., Soland, J., Lewis, K., Ruzek, E., & Johnson, A. (2022). The COVID-19 school year: Learning and recovery across 2020-2021. *AERA Open*, *8*, 233285842210993. https://doi.org/10.1177/23328584221099306

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. *Educ. Res.*, *49*(8), 549–565. https://doi.org/10.3102/0013189X20965918

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Read. Res. Q.*, *45*(2), 230–251. https://doi.org/10.1598/RRQ.45.2.4

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *J. Educ. Psychol.*, *95*(1), 3–21. https://doi.org/10.1037/0022-0663.95.1.3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, S. A., George Benjamin, R., Schwanenflugel, P. J., & Kuhn, M. R. (2014). The longitudinal relationship between reading fluency and reading comprehension skills in second-grade children. *Read. Writ. Q.*, *30*(2), 116–138. https://doi.org/10.1080/10573569.2013.789785

Learning, R. (2021). How kids are performing: Tracking the midyear impact of covid-19 on reading and mathematics achievement (special report series, winter 2020–2021 edition).

Lerkkanen, M.-K., Pakarinen, E., Salminen, J., & Torppa, M. (2023). Reading and math skills development among finnish primary school children before and after COVID-19 school closure. *Read. Writ.*, *36*(2), 263–288. https://doi.org/10.1007/s11145-022-10358-3

Lockl, K., Attig, M., Nusser, L., & Wolter, I. (2021). Cognitive and affective-motivational factors as predictors of students' home learning during the school lockdown. *Front. Psychol.*, *12*, 751120. https://doi.org/10.3389/fpsyg.2021.751120

Ludewig, U., Kleinkorres, R., Schaufelberger, R., Schlitter, T., Lorenz, R., König, C., Frey, A., & McElvany, N. (2022). COVID-19 pandemic and student reading achievement: Findings from a school panel study. *Front. Psychol.*, *13*, 876485. https://doi.org/10.3389/fpsyg.2022.876485

Mair, P., Hatzinger, R., & Maier, M. (2010). Erm: Extended rasch modeling. https://doi.org/10.18637/jss.v020.i09

Mayringer, H., & Wimmer, H. (2003). Salzburger lese-screening für die klassenstufen 1-4 (sls 1-4).

Miller, A., Fray, L., & Gore, J. (2023). Correction to: Was COVID-19 an unexpected catalyst for more equitable learning outcomes? A comparative analysis after two years of disrupted schooling in Australian primary schools. *The Australian Educational Researcher*. https://doi.org/10.1007/s13384-023-00624-w

Molnár, G., & Hermann, Z. (2023). Short- and long-term effects of COVID-related kindergarten and school closures on first- to eighth-grade students' school readiness skills and mathematics, reading and science learning. *Learn. Instr.*, *83*(101706), 101706. https://doi.org/10.1016/j.learninstruc.2022.101706

Nandi, A., Haberland, N., Kozak, M., & Ngô, T. D. (2023). The gendered effects of the COVID-19 pandemic on adolescent literacy and schooling outcomes in india. *NPJ Sci. Learn.*, *8*(1), 42. https://doi.org/10.1038/s41539-023-00193-8

Oecd. (2020). The impact of COVID-19 on education: Insights from education at a glance 2020. *Organisation for Economic Co-operation and Development*.

Oostdam, R., van Diepen, M., Zijlstra, B., & Fukkink, R. (2024). Effects of the COVID-19 school lockdowns on language and math performance of students in elementary schools: Implications for educational practice and reducing inequality. *Eur. J. Psychol. Educ.*, *39*(1), 129–149. https://doi.org/10.1007/s10212-023-00679-4

Peters, M. T., Förster, N., Hebbecker, K., Forthmann, B., & Souvignier, E. (2021). Effects of data-based decision-making on low-performing readers in general education classrooms: Cumulative evidence from six intervention studies. *Journal of Learning Disabilities*, *54*(5), 334–348. https://doi.org/10.1177/0022219421101158

Peters, M. T., Hebbecker, K., & Souvignier, E. (2022). Effects of providing teachers with tools for implementing assessment-based differentiated reading instruction in second grade. *Assessment for Effective Intervention*, *47*(3), 157–169. https://doi.org/10.17105/SPR44-1.60-75

Pilonieta, P., Whittingham, C. E., & Washburn, E. K. (2023). "changing our teaching": First grade reading instruction and before and during COVID-19. *Read. Writ.*, *37*(3), 1–30.

Psyridou, M., Tolvanen, A., Niemi, P., Lerkkanen, M.-K., Poikkeus, A.-M., & Torppa, M. (2023). Development of silent reading fluency and reading comprehension across grades 1 to 9: Unidirectional or bidirectional effects between the two skills? *Reading and Writing*, *36*(8), 1969–1996. https://doi.org/10.1007/s11145-022-10371-6

R Core Team, R., et al. (2024). R: A language and environment for statistical computing.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Richter, C. G., Siegelman, N., Mahaffy, K., Sabatini, J., & Hoeft, F. (2022). The impact of computer–assisted technology on literacy acquisition during covid-19-related school closures: Group–level effects and predictors of individual–level outcomes. *Frontiers in Psychology*, *13*, 1001555. https://doi.org/10.3389/fpsyg.2022.1001555

Rosendo, D., Pereira, A., Moreira, T., Núñez, J. C., Martins, J., Fróis, S., Paupério, C., & Rosário, P. (2023). Reading in COVID-19 Pandemic Times: A Snapshot of Reading Fluency of Portuguese Elementary School Students. *Children*, *10*(1), 143. https://doi.org/10.3390/children10010143

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. http://www.jstor.org/stable/27590541

Schilcher, A., Wild, J., Kraus, E., & Hilbert, S. (2022). Filby-2–ein leseflüssigkeitstraining für alle schülerinnen und schüler? differenzielle effekte. *Zeitschrift für Sprachlich-Literarisches Lernen und Deutschdidaktik*, *2*. https://doi.org/10.46586/SLLD.Z.2022.9606

Schult, J., Mahler, N., Fauth, B., & Lindner, M. A. (2022b). Long-term consequences of repeated school closures during the COVID-19 pandemic for reading and mathematics competencies. *Front. Educ.*, *13*. https://doi.org/10.1080/09243453.2022.2061014

Segers, E., In 't Zandt, M., Stoep, J., Daniels, L., Roelofs, J., & Gubbels, J. (2023). Differential effects and success stories of distance education in covid-19 lockdowns on the development of reading comprehension in primary schools. *Read. Writ.*, *36*(2), 377–400. https://doi.org/10.1007/s11145-022-10369-0

Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences*, *1*(1), 1–6. https://doi.org/10.1186/s42409-018-0006-0

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Sci. Stud. Read.*, *17*(2), 108–133. https://doi.org/10.1080/10888438.2011.618153

Skar, G. B., Graham, S., & Huebner, A. (2023). The long-term effects of the COVID-19 pandemic on children's writing: A follow-up replication study. *Educ. Psychol. Rev.*, *35*(1), 15. https://doi.org/10.1007/s10648-023-09729-1

Starling-Alves, I., Hirata, G., & Oliveira, J. B. A. (2023). Covid-19 school closures negatively impacted elementary-school students' reading comprehension and reading fluency skills. *International Journal of Educational Development*, *99*, 102753. https://doi.org/10.1016/j.ijedudev.2023.102753

Storey, N., & Zhang, Q. (2021, September). *A meta-analysis of the impact of COVID-19 on student achievement*. https://doi.org/10.35542/osf.io/qekw2

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology*, *14*, 166–182. https://doi.org/10.1007/s10742-014-0123-z

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of statistical software*. https://doi.org/10.18637/jss.v042.i08

Stuart Mill, J. (1843). A system of logic, ratiocinative and inductive. *Collected Works of John Stuart Mill*, *7*.

Tomasik, M. J., Helbling, L. A., & Moser, U. (2021). Educational gains of in-person vs. distance learning in primary and secondary schools: A natural experiment during the COVID-19 pandemic school closures in switzerland. *Int. J. Psychol.*, *56*(4), 566–576. https://doi.org/10.1002/ijop.12728

UNESCO. (2022). Unesco map on school closures. https://covid19.uis.unesco.org/global-monitoring-school-closures-covid19/country-dashboard/,(Accessed:%20%7BJune%2028th,%202024%7D)

van de Ven, M., Voeten, M., Steenbeek-Planting, E. G., & Verhoeven, L. (2017). Post-primary reading fluency development: A latent change approach. *Learning and Individual Differences*, *55*, 1–12. https://doi.org/10.1016/j.lindif.2017.02.001

Wild, J., Kraus, E., Steinert, M., Hilbert, S., & Schilcher, A. (2022). Ein Lesestrategietraining im Distanzunterricht? Wie sich das veränderte Lernumfeld während der Corona-Pandemie auf die Entwicklung des Leseverstehens von Schülerinnen und Schülern auswirkt. In M. Philipp & S. Jambor-Fahlen (Eds.), *Lesen: Prozess-und Produktperspektiven von der Wortebene bis zu multiplen Texten*. Beltz Juventa.

William Revelle. (2023). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.3.6]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Wimmer, H., & Mayringer, H. (2016). *Salzburger lese-screening für die schulstufen 2-9: Sls 2-9*. Huber.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual review of public health*, *39*, 453–469. https://doi.org/10.1146/annurev-publhealth-040617-013507

Zalewska, A., Gałczyk, M., Sobolewski, M., & Fernandes, H. (2023). A pilot cross-sectional study on the level of depression and physical activity among students in poland and portugal in the second year of the COVID-19 pandemic. *J. Clin. Med.*, *12*(7). https://doi.org/10.3390/jcm12072541

Zierer, K. (2021). Effects of pandemic-related school closures on pupils' performance and learning in selected countries: A rapid review. *Educ. Sci. (Basel)*, *11*(6), 252. https://doi.org/10.3390/educsci11060252