Causal Effect of Online Learning during COVID Pandemic on Student Performance

Ayumu Yamagishi
ECON427 Introduction to Econometrics II: Prediction and Causal Inference
GitHub Repo page: https://github.com/000Aym612/CausalInferenceForOnlineLearning

COVID-19 pandemic had a significantly huge impact on human life that changed our daily lifestyle. The graph(Image) indicated that the number of COVID-19 patients and dead people exponentially increased from the beginning of 2020 and gradually continued its power until the middle of 2022(CDC, COVID Data Tracker). Through this pandemic, we have experienced lots of changes in some fields. Educational fields were not the exception. Starting from the end of 2019 and beginning of 2020, a large number of schools/districts replaced their traditional in person class style with online learning or hybrid learning style. Even though the pandemic ended, the online learning style remains due to its accessibility, reducing instructors loadworks or efficiency. However, the effectiveness of those online learning methods are still unclear. Some researchers suggested that students tended to be lazy not to do homework or cheat in the online assignments and exams. Widiasih (2022) concluded that online learning during the COVID-19 pandemic had significantly bad influence on the student learning outcomes. Also, the study (Bernard F,D, 2021) aimed to compare the effectiveness of classroom learning at the University of Cape Coast during the pandemic with that before the pandemic, and it concluded that classroom learning was more effective than online learning.

In this study, two primary datasets were used to estimate the causal effect of online learning on student performance during the COVID-19 pandemic. The first dataset, derived from the Stanford Education Data Archive (SEDA 2022), provides standardized academic performance scores for school districts across the United States for the years 2019 and 2022. Each observation in this dataset corresponds to a unique combination of school district, subject (mathematics or reading/language arts), and year. The key outcome variable used from this dataset is the mean_ys, which represents a standardized test score that has been normalized to the national average in 2019 for a given grade and subject. This variable allows for meaningful comparisons across years, subjects, and states. Accompanying this are the standard errors of the scores and student sample sizes, which help determine the statistical reliability of each observation. The analysis primarily focuses on the “all students” subgroup to capture general trends in academic outcomes.

The second dataset, from the COVID School Datahub, records the monthly percentage of students receiving in-person, hybrid, or fully online instruction in each school district. Each row identifies a district and a specific month, along with the respective share of students enrolled in each instructional mode. A treatment variable D was constructed from this dataset, where a value of 1 indicates that the district’s share of hybrid and virtual instruction exceeded 50% in a given month, signaling a shift toward online learning. By aggregating this monthly variable across all months, the study generated a measure of online_months, indicating the total number of months a district spent primarily in an online learning environment. Districts with at least one month marked as D = 1 were categorized into the treatment group, while others formed the control group.

Because the two datasets are not perfectly aligned, the analysis was limited to the set of school districts and states that were present in both sources. This ensured consistency in measuring the change in academic performance and exposure to online learning. The key dependent variable for the analysis was the change in standardized test scores (delta_score), calculated as the difference in mean_ys scores between 2022 and 2019. This outcome was then compared between the treatment and control groups using the Difference-in-Differences (DiD) method. DiD allowed the study to control for time-invariant district-level characteristics and capture the average effect of online learning, using the pre-pandemic year as a baseline. In addition, the Synthetic Control Method was employed to create more precise counterfactual estimates by matching treated states with a weighted combination of control states that exhibited similar pre-pandemic performance. Together, these methods enabled the study to draw a more accurate causal inference on the impact of online learning during the COVID-19 pandemic.

In our analysis, we use the SEDA 2022 dataset to compare students' exam scores from the years 2019 and 2022. This comparison is made based on subject area, racial subgroup, and a standardized measurement indicating how far each school district's average score deviates from the national standard deviation for that year. Several variables in the dataset are particularly relevant to our research. The variables **"sedaadmin"** and **"sedaadminname"** are used to identify each school district and its corresponding state. The variable **"subject"** distinguishes between the two subject areas: **mth**, which stands for mathematics, and **rla**, which refers to reading/language arts or comprehension. The variables "**ys_mn19_ol"** and **"ys_mn22_ol"** represent the mean standardized test scores in 2019 and 2022, respectively, on the Year Standardized (YS) scale. The accompanying variables **"ys_mn19_ol_se"** and **"ys_mn22_ol_se"** provide the standard errors associated with those means. These values indicate the level of uncertainty around the average scores. The variable **"ys_chg_ol"** captures the change in standardized scores between 2019 and 2022, serving as a direct measure of the change in student performance over time. The standard error of this change, **"ys_chg_ol_se"**, is also provided to assess the statistical reliability of the observed difference. Together, these features allow us to evaluate the causal effect of online learning on academic outcomes by examining score differences across districts and instructional modes before and after the pandemic.

The "District-Monthly Percentage In-Person, Hybrid, or Virtual" dataset, provided by the COVID-19 School Data Hub, offers detailed monthly information on the instructional modes adopted by school districts across 48 U.S. states during the COVID-19 pandemic. This dataset is designed to capture how the mode of instruction—fully in-person, hybrid, or fully virtual—varied across districts and over time, enabling researchers to analyze patterns of educational delivery throughout the pandemic period. Each observation in the dataset corresponds to a specific school district and month, and contains data on the proportion of students who received instruction in each mode. The key variables include a district identifier (`leaid`), the district's name, the state in which it is located, and a `month` field representing the time of observation in YYYY-MM format. Three primary numeric fields are used to indicate instructional modality: `share_inperson`, `share_hybrid`, and `share_virtual`. These represent the percentage of students in a given district and month who were taught fully in person, through a hybrid model, or fully virtually, respectively. For each observation, the three percentages are designed to sum to approximately one, accounting for the full distribution of instructional formats. In research, this dataset is particularly valuable for constructing treatment variables that distinguish between districts based on their level of exposure to online learning. For instance, researchers may define a district as "treated" if more than 50 percent of instruction in a given month was delivered either virtually or through a hybrid format. By aggregating such treatment definitions across time, one can also compute the total number of months a district spent primarily in an online learning environment—an indicator often used to study the long-term impact of online learning.

Data Overviews and Preprocessing

To begin the analysis, we first processed the District_Monthly_Shares_03.08.23.csv dataset, which provides monthly records of instructional delivery modes—specifically, the proportions of students receiving in-person, hybrid, or fully virtual instruction across school districts. We identified months in which the combined share of hybrid and virtual instruction exceeded 50% as months of "online learning." For each district, we then calculated the total number of such months, creating a new variable, Online_months, which serves as an indicator of a district's overall exposure to online instruction during the pandemic.

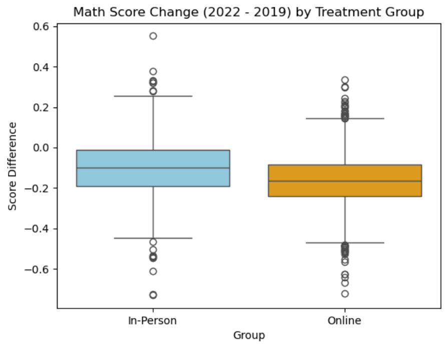
Next, we incorporated academic performance data from the seda2022_admindist_poolsub_ys_beta.csv dataset. This dataset contains standardized test scores from 2019 and 2022, disaggregated by district and subject (mathematics or reading/language arts). It includes estimates of means and standard errors, allowing us to track learning progress over time. To reduce noise, we limited our analysis to records where the student subgroup is marked as "all." After standardizing district names to uppercase, we merged the two datasets using district IDs and state abbreviations as matching keys. This merge enabled us to directly compare each district's online learning exposure with corresponding changes in academic performance.

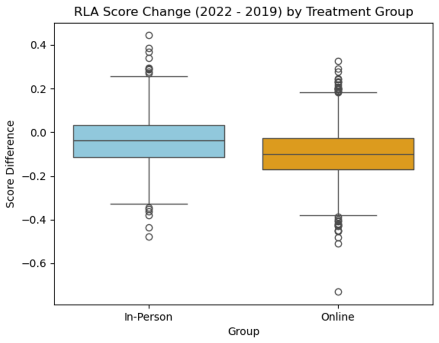
Treatment Effect of Online Learning on Students' Test Score Changes

To estimate the Average Treatment Effect (ATE), we defined the treatment group as districts with at least one month of predominantly online instruction (Online_months ≥ 1) and the control group as districts that remained fully in-person. For each district, we computed the difference in standardized test scores between 2019 and 2022, and then compared the average of these changes across the two groups. The results revealed that mathematics scores declined by approximately 6.03% and RLA scores by 5.85% in districts that adopted online learning—suggesting a potentially adverse effect of remote instruction on student achievement.

Difference in Difference on Students' Test Score Changes for 2019 and 2022

To strengthen the causal interpretation of these findings, we applied a Difference-in-Differences (DiD) approach. This method estimates the effect of online learning by comparing changes in test scores over time between treated and control districts while accounting for common shocks experienced by all districts during the pandemic. Our regression model included indicators for treatment status, post-pandemic year, and their interaction, with the interaction term capturing the DiD estimate. The results for both subjects mirrored the ATE findings, further supporting the conclusion that online instruction negatively affected student outcomes.





Synthetic Control Method

Finally, we conducted a complementary analysis using the Synthetic Control Method (SCM) to evaluate the impact of online learning at the state level. Unlike the prior district-level analyses, this approach provides a macro-level perspective by comparing actual outcomes in treated states—such as California—with those of a weighted combination of control states that did not implement widespread remote instruction. By constructing a "synthetic California" based on pre-pandemic academic performance trends, we were able to approximate what the state's scores might have looked like in the absence of online learning. The analysis showed that California's actual scores in 2022 were substantially lower than those of the synthetic control, reinforcing the evidence that online instruction had a measurable and negative effect even at the state level.

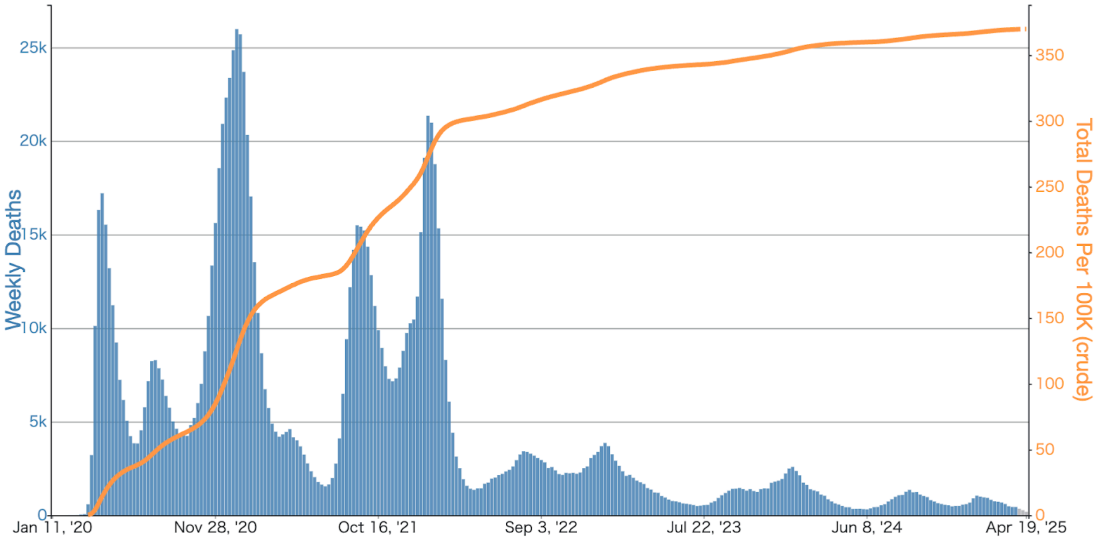
Conclusion

This study set out to investigate the causal impact of online learning during the COVID-19 pandemic on student academic performance in the United States. Motivated by the sudden and widespread shift from traditional in-person instruction to remote or hybrid models, we aimed to provide empirical evidence on whether this transformation affected educational outcomes—and if so, to what extent. To do this, we drew on two large-scale datasets: the Stanford Education Data Archive (SEDA 2022), which offered standardized test scores from 2019 and 2022, and the COVID School Data Hub, which provided district-level information on instructional modality throughout the pandemic. Our research design combined multiple causal inference techniques to ensure robustness. We began by calculating the Average Treatment Effect (ATE), comparing test score changes between school districts that experienced significant exposure to online learning and those that did not. The results showed a measurable decline in academic performance: mathematics scores dropped by approximately 6.03%, and reading/language arts scores by 5.85% among districts with greater online instruction exposure. To account for unobserved time-invariant factors, we implemented a Difference-in-Differences (DiD) analysis using the pre-pandemic year (2019) as a baseline. The DiD estimates closely mirrored the ATE results, reinforcing the conclusion that online learning had a statistically and substantively negative effect on student outcomes. Recognizing that educational policies are often made at the state level, we further applied the Synthetic Control Method (SCM) to assess effects on a broader scale. By comparing California—one of the states with prolonged remote instruction—to a weighted combination of control states, we constructed a counterfactual estimate of what test scores might have looked like without widespread online learning. The gap between actual and synthetic outcomes confirmed that even at the state level, online learning was associated with significant academic losses. Taken together, these findings suggest that while online learning offered critical continuity during a global crisis, it came at an academic cost—particularly in terms of standardized test performance. As school systems consider the continued integration of online instruction in a post-pandemic era, it will be essential to pair such approaches with targeted interventions that mitigate learning loss and ensure equity. Our results emphasize the importance of designing online learning environments that are not only accessible and efficient but also pedagogically effective and engaging.

Visualizations

Trend of COVID-19 in the U.S. from 2019 to 2025

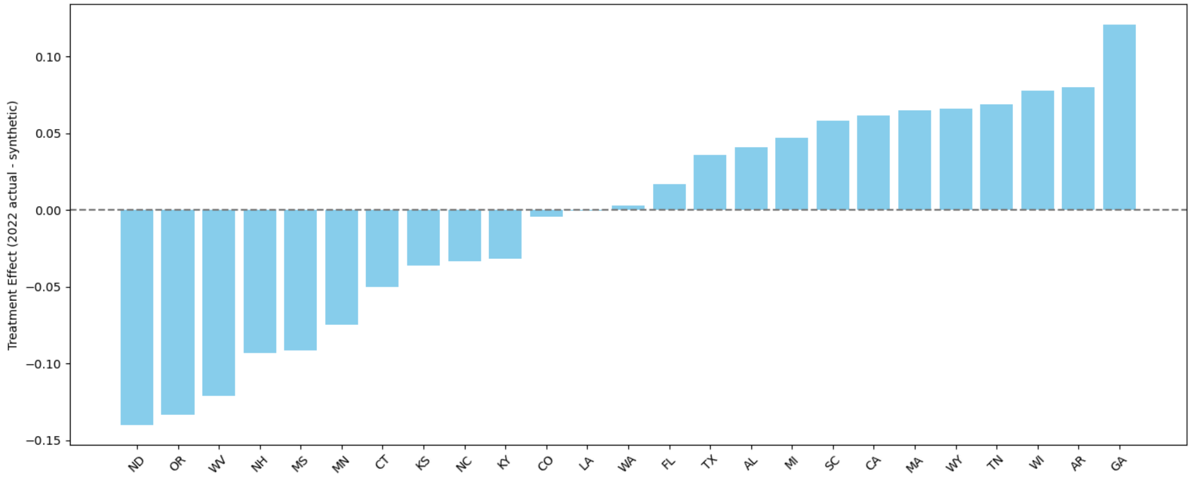
Provisional COVID-19 Deaths and Total COVID-19 Death Rate per 100,000 Population (Crude), by Week, in The United States, Reported to CDC



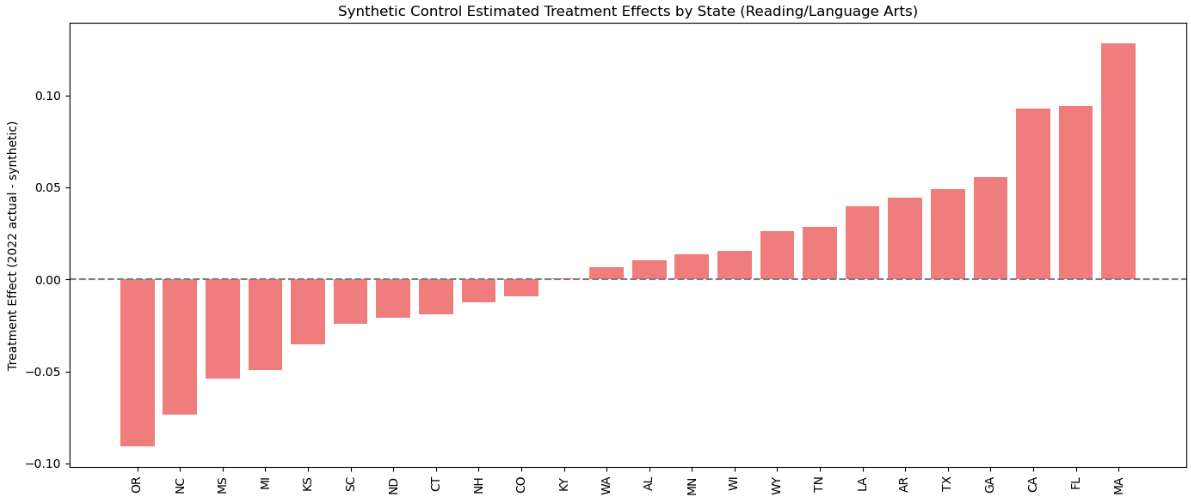
Centers for Disease Control and Prevention. COVID Data Tracker. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2025, April 30. https://covid.cdc.gov/covid-data-tracker

Synthetic Control Model for Math by Treatment Effects

Synthetic Control Estimated Treatment Effects by State (Math)

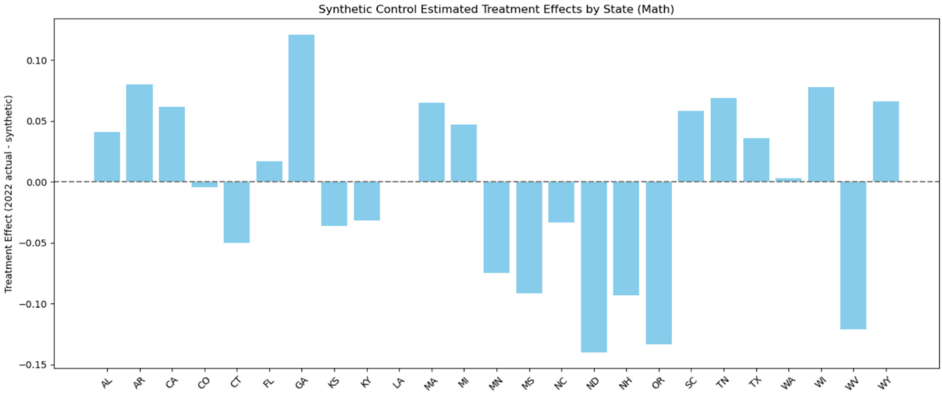


Synthetic Control Model for RLA by Treatment Effects

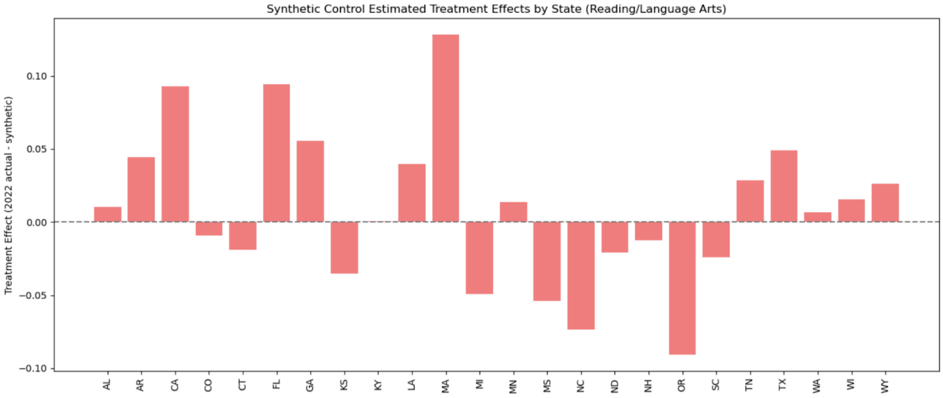


Synthetic Control Models for RLA and Math by State

Synthetic Control Estimated Treatment Effects by State (Math)



Synthetic Control Estimated Treatment Effects by State (Reading/Language Arts)



References

- Centers for Disease Control and Prevention(CDC), Trends in United States COVID-19 Deaths, Emergency Department (ED) Visits, and Test Positivity by Geographic Area：https://covid.cdc.gov/COVID-DATA-TRACKER/#trends_weeklydeaths_totaldeathratecrude_00
- The Stanford Education Data Archive: https://edopportunity.org/
- Covidshooldatahub: https://www.covidschooldatahub.com/data-resources
- Widiasih, Restuning1,; Suryani, Suryani2; Rakhmawati, Windy3; Arifin, Hidayat4. The Impact of Online Learning among Adolescents during the COVID-19 Pandemic: A Qualitative Study of Mothers' Perspectives. Iranian Journal of Nursing and Midwifery Research 27(5):p 385-391, Sep–Oct 2022. | DOI: 10.4103/ijnmr.ijnmr_91_21
- Bernard Fentim Darkwa, Samuel Antwi, published by Open Access Library Journal, Vol.8 No.7, 2021, "From Classroom to Online: Comparing the Effectiveness and Student Academic Performance of Classroom Learning and Online Learning"