Statistics in Educational Policy

Project Report

Nathan Yang

Introduction

The Covid-19 pandemic has had a profound effect on schools in the US. For many students, the transformation from live personal engagement to remote learning or worse, schools shutting down, has been difficult to adjust to. This struggle is evident when comparing academic performance metrics pre-pandemic to post-pandemic.

My project aims to explore these primary questions:

- What are the factors that influence the academic performance of school districts?
- What factors are associated with discrepancies between math and reading academic performance per school district?

Research on the pandemic's effect on education has largely demonstrated that lower-income areas are heavily associated with lower academic performance and greater learning loss (Irwin et al., n.d.). The Educational Opportunity Project (EOP) by Stanford University measured the change in test scores between Spring 2019 to Spring 2022 as well as Spring 2022 to Spring 2023 along with comparing by district poverty rates (E. Fahle et al. 2024). Researchers have also found greater extent of remote learning was a strong indicator for increased learning loss but were not able to identify further strongly related factors. (E. M. Fahle et al. 2023). Additionally, there is not extensive research on discrepancies between math and reading scores when it comes to performance as well as recovery. As such, this is one of the knowledge gaps that this project seeks to resolve.

For this project, the academic performance data by school district was from the EOP. Additionally, the 2021 Census supplement on Internet/Computer usage was examined as a potential influence on academic performance and association with discrepancy. This project also used many datasets from the American Community Survey that covered population demographics, socioeconomic variables, child population statistics, and more.

Using this data, I created an interactive Shiny dashboard to visualize the factors that could be associated with academic performance and search for potential factors related to math vs reading score discrepancies.

Background

New York Times

There have been many articles published by the New York Times discuss data on academic performance over time, across school districts, and compared to numerous socioeconomic factors. Kane and Reardon (2023) summarizes the findings from the EOP along with other papers and found an association between lower test scores and areas that were poorer, had lower voting rates, had lower institutional trust, and longer school closures. Mervosh, Miller, and Paris (2024) aggregates data from various sources to compare percentage remote/hybrid, family income, and school closure with academic performance.

The trend of discussion appears to have also changed following the end of 2023, in particular, more articles are being published are post-pandemic recovery and strategies that have aided in this process. Miller, Mervosh, and Paris (2024) synthesized data from the EOP to visualize learning loss recovery across states and note the solutions and strategies that are being employed.

Education Reporting

The National Center for Education Statistics issues the biyearly National Assessment of Educational Progress (NAEP) in math, reading, writing, and science to analyze academic performance of students in the US ("National Assessment of Educational Progress (NAEP)," n.d.). This organization also developed specially weighted scales to compare the data across grades.

The EOP uses state-administered exams in math and reading across students from grades 3-8 across thousands of school districts over the course of several years (E. Fahle et al. 2024). This project also linked these test scores to the NAEP scale for comparability.

I drew heavy inspiration by the plots on the website of the Educational Opportunity website where they compared math/reading scores with different socioeconomic factors in a nice comprehensible scatter plot. I noted how effective this method was at mitigating the geographic spread issue so I adapted and extended this idea with my Shiny dashboard. Their dashboard offers filters and divisions by subgroup, subject, and year, but I wanted to extensively look into variables that could be compared to academic performance.

Government Reporting

The American Community Survey (ACS) is a yearly survey administired by the Census Bureau to random household owners to collect housing and population statistics. Altogether, the ACS datasets contain aggregate statistics across hundreds of variables and cover a diverse range of population characteristics on a school district level basis.

Data

The Eductional Opportunity Project was developed by researchers at Stanford University to study disparities in academic performance across school districts. This project uses state-administered standardized tests across grade 3-8 students in both math and reading with the primary years of interest being 2019, 2022, and 2023. The values are scaled based off a 2019 national standard making scores comparable across school districts. The grade year standard (GYS) dataset specifically indicates each unit represents a full grade level of proficiency and is comparable across years.

The Current Population Survey (CPS) is a national monthly survey designed to get information about labor statistics in the US. The 2021 CPS Computer/Internet Use supplement was of particular interest because I had a hypothesis that internet/technology usage could one of the primary factors that would differentiate performance in math and reading. As such, I was hoping to compare academic performance and technology usage across matching geographical areas. This did not end up in the final deliverable due to incompatible geographic matching with the academic performance dataset.

The ACS became my primary source for school district data in order to draw comparisons across school districts and academic performance. Many of the datasets were challenging to work with and required a lot of data preparation and cleaning to be machine readable. I ended up using ACS datasets that covered topics in housing, economics, and demographics.

Methodology

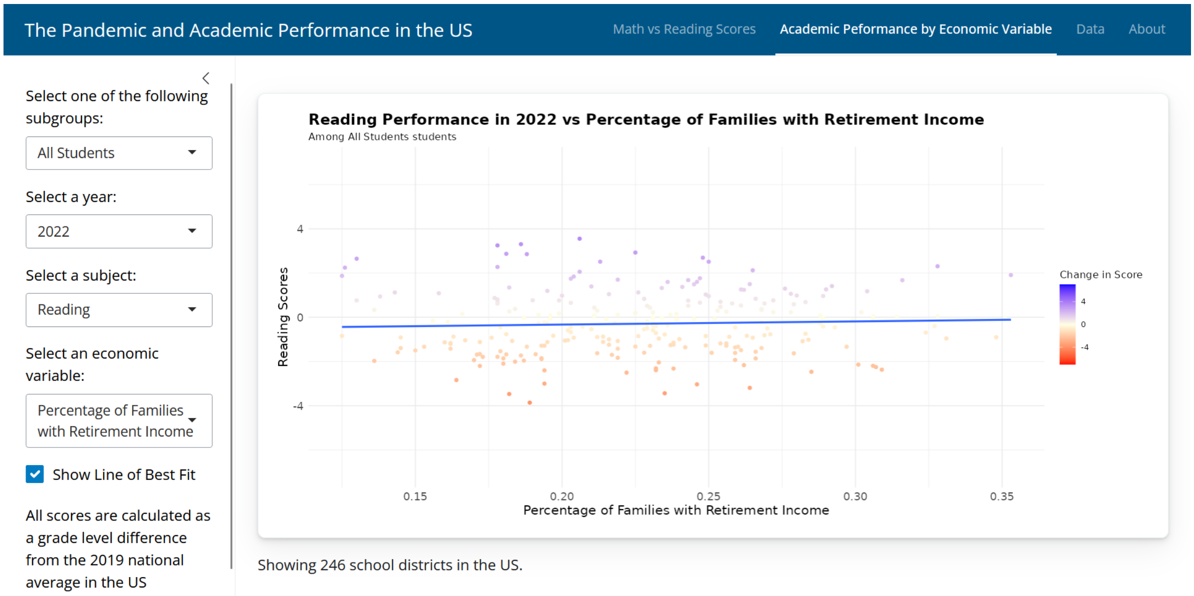
The first step required was to determine what datasets I needed to collect and perform any necessary data cleaning or preparation. The ACS datasets in particular were challenging to work with since they did not have a consistent format pattern and contained difficult to read variable names. Once that preliminary cleaning was completed, the ACS datasets were joined into the EOP dataset by school district name to create a large cumulative dataset. Additionally, I had to pivot longer several fields in order to build interactive elements on my map that relied upon field condensation.

My initial goal was to create a geospatial visualization where school districts across the US could be represented as points and be interacted with to gleam more information. I experimented with various R packages like the maps (Becker, Minka, and Deckmyn. 2023), usmap (Di Lorenzo 2024), and sf (Pebesma and Bivand 2023) packages to determine which had the capabilities, customization, and functionality I was looking for. After experimenting with these maps packages, I began to notice flaws in trying to create this visualization as a central piece of my project. In particular, the housing/income ACS dataset only represented 719 school districts across 34 states which would have not translated well on a map. Additionally, the CPS supplement was distributed by county instead of school district, which made it much more difficult to justify trying to use a map since it couldn't be joined with the EOP data without losing location specificity. As such, I concluded that a map should not be my ultimate goal for the project due to restrictions on the geographic spread of my datasets and that I should instead expand on the scatterplots featured on the EOP website.

I used R Shiny to create interactive properties for my plots, flesh out the functionality, and make additions to my dashboard to make it stand out more compared its inspiration. I added features to better help identify trends like fixed plot scaling, lines of best fit, and a more distinct color gradient. I also wanted to explore more variables and factors that could explain academic performance and specifically discrepancies between math and reading scores. A filter by year was also added to visualize the effect of the pandemic.

Results

The dashboard I created offers filters and divisions similar to the existing dashboard, but I analyzed a greater number of socioeconomic and demographic factors that could affect academic performance.


The Pandemic and Academic Performance in the US
Math vs Reading Scores
Academic Peformance by Economic Variable
Data
About

Select one of the following subgroups: All Students

Select a year: 2022

Select a subject: Reading

Select an economic variable: Percentage of Families with Retirement Income

☑ Show Line of Best Fit

All scores are calculated as a grade level difference from the 2019 national average in the US

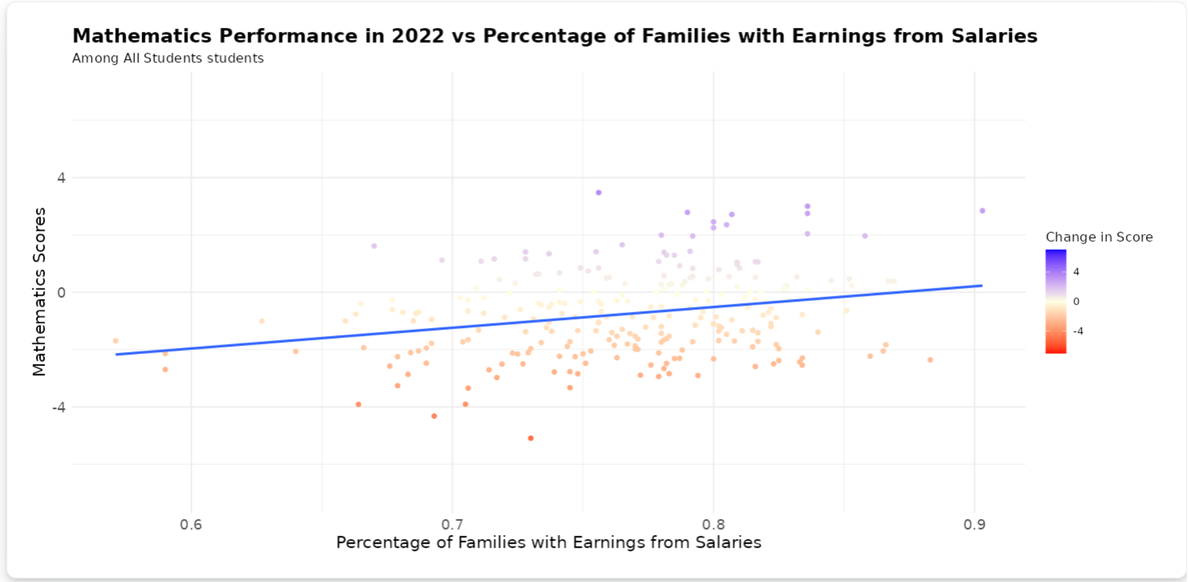
Reading Performance in 2022 vs Percentage of Families with Retirement Income

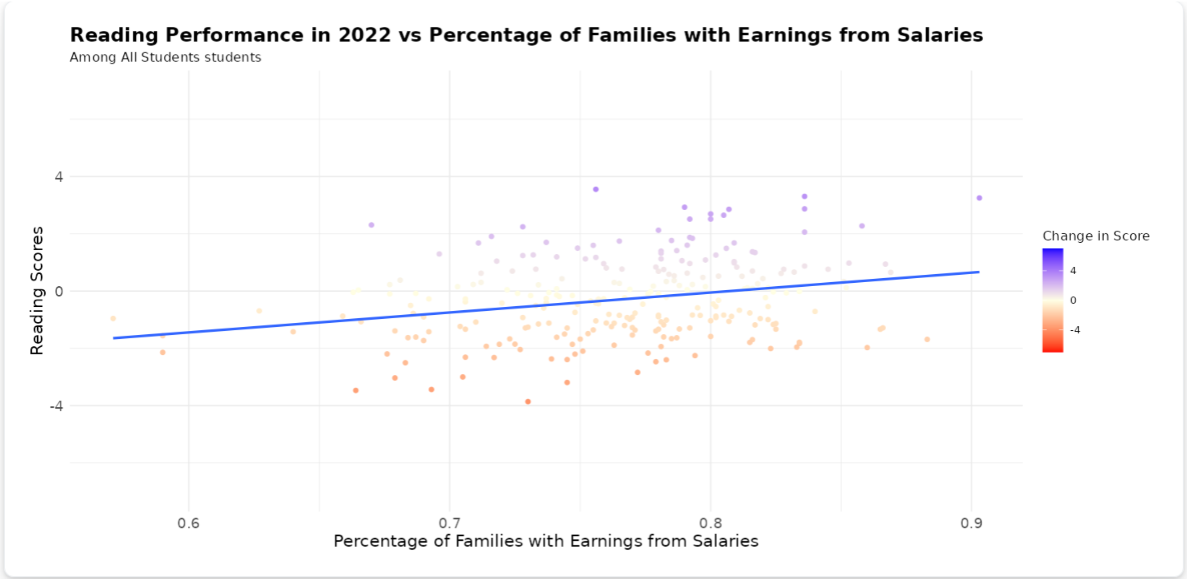
Among All Students students

Showing 246 school districts in the US.

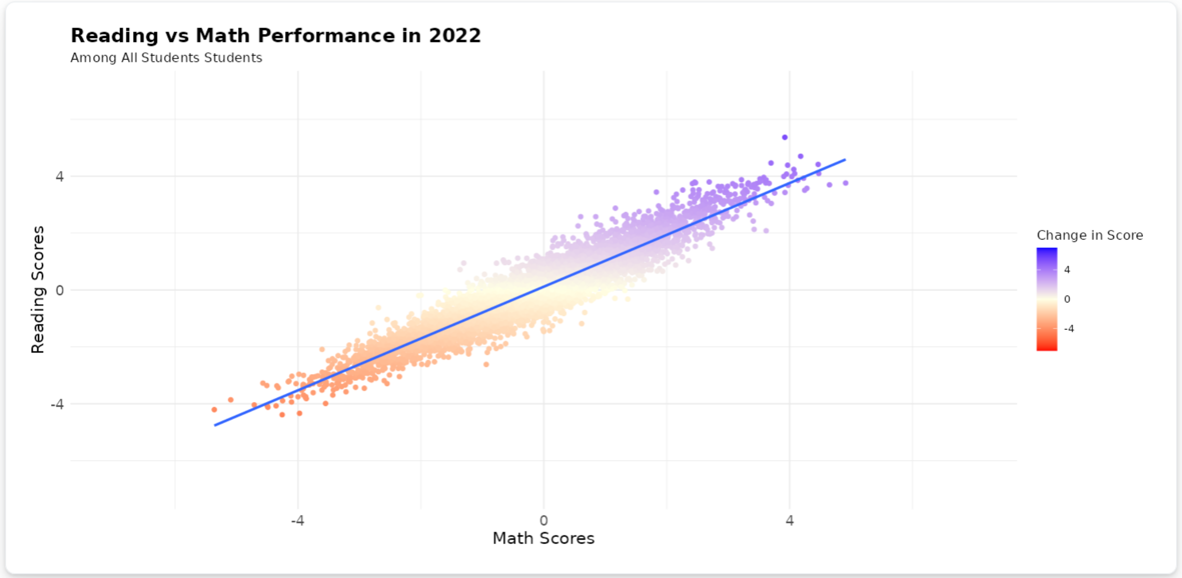
Additionally, all the plots used a fixed y-axis scale and include a toggleable line of best fit for the purposes of better observing trends and making comparisons while changing parameters.

From plotting socioeconomic factors, it was clear that indicators of areas being lower-income or lower-earning tended to be associated with lower academic performance in both math and reading.





These two plots examine percentage of families with salary earnings compared to math and reading academic performance in 2022. The line of best fit indicates a slightly positive relationship between greater percentage of earnings and increased performance.



This plot looks at math vs reading scores in 2023 and demonstrates that performance in one subject was indicative of performance in the other across school districts.

Furthermore, this app allows distinguishing between different student subgroups: All, White, Black, Hispanic, and Economically Disadvantaged. However, not all school districts have data specific per subgroup. Additionally from comparing any plots from 2022 to 2023, the recovery effect is very noticeable. Similarly, changing any plot from 2019 to 2021 results in notable decrease in overall performance.

Discussion/Conclusion

My original research questions cannot be conclusively answered from the work of this project; however, I was able to discover and confirm the findings of many other researchers that lower-income/poorer school districts generally had worse academic performance than high-income/wealthier school districts. Additionally, I did not identify any socioeconomic variables that resulted in noticeable discrepancies between reading and math scores. Similarly, the plot of reading scores vs math scores showed that how school districts were performing in one subject reflected how they performed in the other. As such, it remains difficult to answer the second question.

Comparing academic performance in 2019 to 2022, it becomes very evident that the pandemic had a considerable negative effect on academic performance. Comparing academic performance from 2022 to 2023 shows that schools districts are largely recovering from learning loss resulting from the pandemic. Additionally, comparing 2019 to 2023 academic performance shows that the recovery has still not yet brought schools back yet to 2019 profieciency levels. My conclusions matches the findings of other projects that looked into the effect of the pandemic on academic performance and the learning loss recovery in 2023.

My dashboard looked into the difference between reading and math performance as an interactive tool. In particular, it expands upon the scatterplot on the EOP website with additional socioeconomic variables, fixed axis, and streamlined filtering that result in easier comparisons across years, subjects, subgroups, and district level variables.

The existing EOP dashboard allows interactivity with each of the school districts represented as points and provides detailed information about the school district when selected. Additionally, their points are sized based on school population. While this project also uses socioecnomic data from the ACS, their plot appears to have superiopr coverage of school districts.

To take my project further, I would like to continue exploring different factors and variables per school district to draw more insights into what seems to affect academic performance. I would also like to include more interactivity with the scatterplot and other kinds of plots to add further engagement. Additionally, I would also want to look into more COVID-related school district statistics to better demonstrate the effect of the pandemic on academic performance.

Appendix

Data Dictionary:

Variable Name	Description
state	The state abbreviation of the School District
school_district	The name of the School District
subgroup	The racial/economic subgroup
year	The year of the data
Mathematics	The mathematics score
Reading	The reading score

References

Becker, Original S code by Richard A., Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2023. "Maps: Draw Geographical Maps." https://CRAN.R-project.org/package=maps.

Di Lorenzo, Paolo. 2024. "Usmap: US Maps Including Alaska and Hawaii." https://CRAN.R-project.org/package=usmap.

Fahle, Erin M, Thomas J Kane, Tyler Patterson, Sean F Reardon, Douglas O Staiger, and Elizabeth A Stuart. 2023. "School District and Community Factors Associated With Learning Loss During the COVID-19 Pandemic," May.

Fahle, Erin, Thomas J Kane, Sean F Reardon, and Douglas O Staiger. 2024. "The First Year of Pandemic Recovery: A District-Level Analysis," January.

Irwin, Véronique, Ke Wang, Sarah Hein, Jijun Zhang, Riley Burr, Ashley Roberts, Amy Barmer, et al. n.d. "Report on the Condition of Education 2022."

Kane, Thomas, and Sean Reardon. 2023. "Opinion | Parents Don't Understand How Far Behind Their Kids Are in School." *The New York Times*, May. https://www.nytimes.com/interactive/2023/05/11/opinion/pandemic-learning-losses-steep-but-not-permanent.html.

Mervosh, Sarah, Claire Cain Miller, and Francesca Paris. 2024. "What the Data Says about Pandemic School Closures, Four Years Later." *The New York Times*, March. https://www.nytimes.com/2024/03/18/upshot/pandemic-school-closures-data.html.

Miller, Claire Cain, Sarah Mervosh, and Francesca Paris. 2024. "Students Are Making a 'Surprising' Rebound from Pandemic Closures. But Some May Never Catch Up." *The New York Times*, January. https://www.nytimes.com/interactive/2024/01/31/us/pandemic-learning-loss-recovery.html.

"National Assessment of Educational Progress (NAEP)." n.d.

Pebesma, Edzer, and Roger Bivand. 2023. "{Spatial Data Science: With Applications in r}." https://doi.org/10.1201/9780429459016.