Understanding the Impact of the COVID-19 Pandemic on Academic Performance and Effective Educational Policy for Recovery

Nathan Yang

Abstract

The COVID-19 pandemic has had a profound impact on academic performance across the United States. Although math achievement has generally declined since the pandemic, recent years show signs of recovery at a rate behind pre-pandemic levels. My thesis explores the variation in math scores across school districts and over time, offering insights into the demographic, socioeconomic, and educational factors shaping academic performance. By leveraging a comprehensive dataset that integrates academic performance data with county-level socioeconomic and demographic indicators as well as school district characteristics, I use hierarchical modeling to account for the nested structure of the data. I find that school districts with higher median income, higher social capital, and higher percentages of white and Asian students tend to have higher math scores.

School districts that received more emergency funding per student and relied heavily on virtual learning tended to have lower math scores, emphasizing the need for targeted interventions to support low-income students and mitigate challenges associated with virtual learning.

Project Background

The goal of this project is to model academic performance in school districts across the United States through various demographic and socioeconomic factors. The data sources include the American Community Survey (ACS), the Educational Opportunity Project, the Longitudinal School Demographic Dataset (LSDD), the Common Core of Data (CCD), and the Census. The data was joined by school district and county to create a comprehensive dataset for analysis.

I created some preliminary simple linear regression models within an interactive dashboard to explore the relationship between academic performance and various socioeconomic factors. I used the shiny (Chang et al. 2023) package to create the dashboard and the ggplot2 (Wickham 2016) package to create visualizations for the data. I used the rsconnect (Atkins et al. 2024) package to deploy the dashboard onto shinyapps.com where it can be publicly accessible.

Due to a lack of datasets aggregated to the school district level, this project focused on expanding my earlier work by incorporating data sources with county level statistics and an improved method for joining datasets to preserve more records and develop a more comprehensive dataset for analysis. Furthermore, I incorporated hierarchical modeling to account for the nested structure of the data and explored the impact of various predictors on academic performance.

Literature Review

The COVID-19 pandemic has had a significant impact on education, with many students experiencing disruptions in learning due to school closures and shifts to remote learning. Several studies have shown that academic performance declined during the pandemic and students from lower income areas were disproportionately affected. For example, Irwin et al. (2022) examined the disruption in postsecondary education plans due to the pandemic and found that lower-income families were more likely to experience disruptions in learning from canceled classes. The Educational Opportunity Project (EOP) by Stanford University created a scale for measuring academic performance across all school districts in the US and found that disadvantaged students suffered larger learning loss (E. M. Fahle et al. 2023). A followup study by the EOP also found that test scores recovered from 2022 to 2023 but that nonpoor students had greater gains than poor students, further widening the achievement gap between the two groups (E. Fahle et al. 2024). The dataset aggregated and curated by the EOP serves as an excellent base to model academic performance across school districts and years.

Other research has also shown that long periods of school closures and extensive remote learning have had a significant impact on academic performance. Kane and Reardon (2023) found that test scores declined more in school districts that were closed for longer, they found districts that were closed for 90% or more of the year during the 2021-2022 school year experienced a significant drop in math scores. Even when schools are experienced a recovery following the pandemic, the rate of recovery varies greatly across school districts and even across racial identities within a district (Miller, Mervosh, and Paris 2024). Research has further shown that the cohort graduation rate varies greatly across different student subgroups with Black, American Indian/Alaska Natives, and economically disadvantaged students all having lower graduation rates than their peers (McFarland, n.d.).

Extensive research in the field of educational policy has demonstrated that socioeconomic factors, school closures, and cultural factors are highly related to academic performance through the pandemic and moving forward. I wish to more extensively explore these relationships through my modeling and analysis. Additionally, I aim to identify the most important predictors of academic performance and how they interact with each other to affect student outcomes. One key aspect of this project is to understand how emergency funding from the Coronavirus Aid, Relief, and Economic Security (CARES) act has impacted academic performance and how it has been distributed across school districts.

Methodology

The first step of my project was to identify the datasets that I would be using for this project. I started with only looking at datasets aggregated by school district as that would be the most granular level relevant to my research question. I started with the American Community Survey (ACS), the Educational Opportunity Project, the Common Core of Data (CCD), and the Census. I reviewed the data dictionaries for each dataset to understand the variable encoding and the extent of the data. I also reviewed the data sources to understand the data collection process and the limitations of the data.

Data

Data Resources

The core dataset is from the Educational Opportunity Project (EOP) and contains academic performance data across school districts (Reardon 2024). The academic performance variables represent standardized state testing scores that were linked to the National Assessment of Educational Progress scale to make them comparable across schools. The value of this variable represents the grade level difference from the 2019 national average and is available for both math and reading although not comparable across subjects. My research focuses on the math scores and this metric will be referred to as the SEDA math score in this paper. Additionally, standardized testing was halted across the US in 2020 and only some states issued them in 2021. As such, the years of interest in this dataset are 2019, 2022, and 2023. This dataset contains academic performance variables aggregated across different student subgroups and subjects within 7390 school districts.

I then identified a dataset from the Census that contains mappings from school district to county (“School Districts and Associated Counties” 2021). About 30% of counties in this dataset contained only one school district with the average being six. This dataset contains the mappings for 18998 school districts to 3133 counties. This dataset was used to join all of the county-level datasets to the academic performance dataset.

The next core datasets were the Data Profiles (DP) from the American Community Survey (ACS). These DPs contain a selection of features from various ACS datasets that are curated to provide a consistent set of features across counties. These datasets contain demographic and socioeconomic features such as income, poverty, housing, education, and employment. These datasets are aggregated at the county level and represent the 2018 to 2022 5-Year estimated statistics for 3222 counties.

I attempted to use ACS datasets that were aggregated by school district but found that the data was very sparse and did not provide enough information for analysis.

The Common Core of Data (CCD) dataset contains information on membership, salaries, and revenue from local, state, and federal sources ("Local Education Agency (School District) Finance Survey (f-33) Data, (v.1a)" 2022). Additionally, this dataset contains COVID emergency relief funding from the Elementary and Secondary School Emergency Relief Fund (ESSER), which was allotted funding through the CARES Act for the purposes of education stabilization during the pandemic. This dataset also contains funding statistics from the Governor's Emergency Education Relief Fund (GEER), which was also allotted from the CARES act and was intended to provide emergency support to schools and higher education institutions. Both of these emergency funds also had extensions (ESSER II and GEER II respectively) that were provided through additional legislation. Both funds were also allocated to schools in alignment with Title 1 funding levels which are intended to provide additional funding to schools with a high percentage of students from low-income families. This dataset is aggregated at the school district level and contains data for 19572 school districts for the 2022 fiscal year.

The Covid School Data Hub (CSDH) dataset contains self-reported data from state education agencies on learning modality and enrollment ("Percentage of School Year Spent in-Person, Hybrid, or Virtual" 2023). The data used was the proportion every school district operated fully virtual, hybrid, or inperson during the 2020-2021 school year. This dataset is aggregated at the school district level and contains data for 14967 school districts.

The Social Capital Project (SCP) dataset contains information on social capital indicators such as family structure, religious attendance, and social trust ("Social Capital Project" 2023). This project used survey data from the ACS to construct these subindices with variables such as births per married woman, religious congregations, voting turnout, and violent crimes per population. A general "Social Capital Index" was then compiled from a linear combination of this subindices. This dataset is aggregated at the county level and contains data for 3142 counties with the indicators generated in 2017.

While these datasets come from a 5-year timeframe, these are variables that are relatively stable over time and are not expected to change significantly. The academic performance data is the most dynamic and will be the focus of the analysis. The other datasets will be used to provide context and additional features for the analysis. Additionally, many of these datasets will have fields that are highly correlated with each other and will need to be pruned down to a more manageable set of features.

Data Curation

I first joined the datasets solely by school district name. This was a simple join that matched the exact names of the school districts. However, this method had issues as many school districts had different names across different datasets due to no common naming convention. This would result in many records not being matched and a loss of data.

Next, I reviewed the data sources I initially picked out and identified additional data sources that could be useful for this project. These came primarily from reading various research papers and reports that studied similar topics. I used an excel spreadsheet to track all of the data sources and variables of interest.

After identifying an exhaustive list of datasets and variables, I began the process of downloading and cleaning the data. I used the tidyverse (Wickham et al. 2019) package to clean and manipulate the data to prepare the datasets to be joined. For my joining process, I used fuzzy matching techniques to join records that had similar school district names but not exact matches. The metrics I used for fuzzy matching were string distance and Jaccard difference.

String distance is a metric that calculates the number of character changes needed to transform one string into another while Jaccard difference is a metric that compares how many 2-letter pairs are shared between two strings. I used the stringdist (Loo 2014) package to calculate both of these metrics and determined thresholds from examining the distributions and matching strength for each metric. I then joined the datasets purely by matching state and calculated the metrics for every pair of school district names within a state. Once I had this dataset with all the potential matches, I developed an extensive filtering process to ensure I got the most accurate matches possible.

1. Filter for matches that both begin with the same letter: This prevents matches names containing North/South and East/West at the beginning are not accidentally mapped together due to the characters in these cardinal directions being similar
2. Filter for matches that end with the same three letters: This prevents matches such as "Abcdefgh county" and "Abcdefgh city" where the school districts may have the same name but are clearly different entities. This also resolves matching names that have numbers at the end such as "Abcdefgh 231" and "Abcdefgh 562" that clearly represent different school districts
3. For each school district in the academic performance dataset, I find its best match based on string distance with ties broken by Jaccard difference (and ties at this stage decided randomly).

This is an example of a dataset joined between my academic performance data and a dataset from the CCD. Using these string comparison metrics, I was able to preserve many records that would have been unmatched if I performed a direct name join. It is especially noticeable with abbreviated words that these metrics help to identify matches like with “Heights” being reduced to “Hts.” or “Community” being abbreviated to “Com” as shown below. Additional common abbreviations found in the school district names are “Saint” written as “St.” and cardinal directions only represented by the first letter.

Table 1: **Example of School District Matching**. This table shows the similarity between `seda_district` and `ccd_district` using a distance measure and Jaccard index.

seda_district	ccd_district	dist	jaccard
Beaverton Rural Schools	Beaverton Schools	6	0.2727273
North Daviess Community Schools	North Daviess Com Schools	6	0.2580645
Southern Wells Community Schools	Southern Wells Com Schools	6	0.2580645
North Lawrence Community Schools	North Lawrence Com Schools	6	0.2500000
South Harrison Community Schools	South Harrison Com Schools	6	0.2500000
Greenfield-Central Community Schools	Greenfield-Central Com Schools	6	0.2285714
Minnetonka Public School District	Minneapolis Public School District	5	0.2857143
Morris Area Public Schools	Moorhead Area Public Schools	5	0.2758621
West St. Paul-Mendota Hts.-Eagan	West St. Paul-Mendota Heights-Eagan	5	0.2432432
Minnesota Public School District	Minneapolis Public School District	5	0.2424242
North Branch Public Schools	North Branch Area Public Schools	5	0.1724138
Ridgefield Park School District	Ridgefield School District	5	0.1666667
Hamilton County CUSD 10	Hamilton Co CUSD 10	4	0.2727273
West Washington County CUD 10	West Washington Co CUD 10	4	0.2142857
Rising Sun-Ohio County Com	Rising Sun-Ohio Co Com	4	0.1818182

By joining datasets by exact district name, I would have only had 4441 records with the CCD data. However, using the fuzzy matching techniques, I was able to match 4576 records. This is a 3% increase in the number of records that were matched.

Through district name joining I was only able to match about 250 school districts with ACS income data. However, using the fuzzy matching techniques, I was able to match 281 school districts. This is a 12% increase in the number of records that were matched.

Early on in my project, I only selected datasets that were aggregated by school district and it unfortunately did not prove fruitful as many of the ACS datasets I investigated had very limited data on school districts. This resulted in poor record retention for future dataset merging in addition to reduced modeling data as demonstrated by the ACS income dataset. I transitioned to identifying the counties for school districts in the academic performance dataset and then joining the datasets by county. This proved to be much more successful as I was able to use many ACS Data Profiles (DP) datasets which are a selection of curated features from various ACS datasets that have greater consistency in data. This allowed me to retain more records and have a more comprehensive dataset for analysis. Only two datasets had sufficient coverage at the school district level, the CCD and the CSDH datasets.

The new comprehensive dataset is much larger and contains more features than the previous dataset. However, the loss of granularity from school district to county may have an impact on the accuracy of the modeling. I keep this in mind throughout my modeling phase and weigh this in when interpreting the results.

The final step for this comprehensive dataset was to filter for records that did not have any missing data in the key variables. This was to ensure that the data was clean and ready for analysis.

Before filtering for missing data, the dataset contained 10842 records with academic performance data across 2019, 2022, and 2023. After applying all filters for missing data, the dataset contained 2208 records. This was a significant reduction in the number of records as many school districts did not have complete coverage across all datasets.

I analyzed the geographic distribution of the data before and after filtering for missing data and found that the geographic diversity is noticeably affected by this filtering process as the number of states represented goes down from 40 to just 9. Additionally, states have a significant reduction in the number of school districts. For example, California had 1165 records before filtering and 0 after filtering. This is due to the high number of missing values in the California data. This will be a limitation in the analysis as the data is not representative of all states. In particular, the ACS and SCP data were only available for counties in 21 states of which only 14 were in common for both datasets. Further filtering brought this down to the 10 states that were common across all datasets.

Examining the geographic distribution of the remaining states, it is evident that the data only represents states in the Central and Eastern US except for Washington. This is another limitation of the analysis as the data is not representative of region variety in the US.

Due to this non-random pattern, standard imputation techniques such as mean imputation by state were deemed unsuitable. An alternative approach considered converting revenue data into categorical ranges (e.g., “Not Reported,” “0-X,” etc.), though this approach risked reducing the informative value of the continuous revenue variable.

Modeling

Preparation

I first transformed the dataset into a long format using the pivot_longer() function. This allowed me to convert the wide-format data on yearly SEDA math scores into a format suitable for longitudinal modeling.

Next, I created new variables to facilitate trend analysis. I calculated the number of years since 2019 (`yearsince2019`) and its square (`yearsqrdsince2019`) to account for potential nonlinear trends over time as there is a general increase in test scores from 2022 to 2023. To standardize the scale of financial variables and put variables on a more similar scale, I converted all revenue and salary figures from raw values into millions of dollars. I also calculated per-student revenue and salaries by dividing total figures by student membership counts as school financial and membership variables had extremely high correlation with each other. Additionally, key socioeconomic metrics such as median income, mean income, and owner-occupied property values were scaled by dividing by 1,000. Percentages for instructional modes (in-person, hybrid, and virtual) were adjusted to range from 0 to 100 for better interpretability.

Lastly, I needed a better identifier for county than just the County Name as that could be duplicated across states. As such, I ceated a new variable `county_state` that combined the county name and state abbreviation to create a unique identifier for each county.

Curation of Predictors

When modeling, it is important to check for collinearity between predictors because highly correlated predictors can lead to unstable estimates and inflated standard errors. I determined general categories for all the predictors and I calculated the correlation matrix for each set to identify highly correlated variables.

For the school modality variables, I found that the percentage of time spent in person and hybrid were highly correlated. I decided to remove both of these variables and just keep the percentage of time spent fully virtual (`share_virtual`) in the model as that kept interpretations more straightforward since I could delineate between the effect of learning environments that were fully virtual or had some component of inperson learning.

For the revenue and salary variables, I found that the membership, total revenue, and total salary variables were very highly correlated. I decided to remove all the total revenue and total salary variables from the model except for aggregate revenue variable that combined funding sources on the federal, state, and local level (total_revenue). Additionally, total revenue per student (revenue_per_student), ESSER funding per student (esser_per_student), and GEER funding per student (geer_per_student) were kept in the model as they were not correlated with any other variables and were important for understanding the financial resources available to school districts.

For the SCP variables, as previously mentioned, the County_Level_index is a linear combination of the other social capital variables. As such, I removed all the other social capital variables from the model except for the composite variable (County_Level_Index) due to correlation issues.

For the ACS social characteristic and employment variables, I found that all the variables regarding marital status were highly correlated with each other. I decided to remove all of these variables except for the percentage of married couple households (married_household). Interestingly, I found that the variables regarding educational attainment were not highly correlated with each other. As such, I kept the percentage of people over 25 with at least a high school degree (over_25_highschool_degree) and the percentage of people over 25 with at least a bachelor's degree (over_25_bachelors_degree). I found that the percentage of families with no workers in the past month and the percentage of married couple households where the householder worked full time in the past 12 months were very highly correlated. Interestingly, unemployment rate was not correlated with the other employment variables. As such I moved forward with the percentage of families with no workers (no_workers), the percentage with one worker (one_worker), and the unemployment rate (unemployment) in the model.

For the ACS demographic variables, I found that variables on the percentage of people that were native born, spoke English at home, and didn't speak English at home were all highly correlated. I decided to keep the percentage of people that were native born (native_born) in the model as it was the most interpretable of the three variables. The race and ethnicity variables that had high correlation were the percentage of people that identified as white (white_percent) and the percentage that identified as black (black_percent). I keep all the racial variables including asian (asian_percent) and hispanic (hispanic_percent) groups in the model however as they are all important for understanding the demographic makeup of the school district.

Additionally, I found that access to a computer was highly correlated to access to the internet. As such, I elected to keep only the percentage with acces to a computer (with_computer) in the model.

For the ACS income variables, median income and mean income jumped out as being extremely correlated with each other. I decided to stick with the median (median_income) because it is a more robust measure of central tendency. These income variables were also very highly correlated with the property value variables so those were removed. The other variables in this category were the percentage of people with health insurance coverage (with_health_insurance), the percentage of people in poverty (poverty), and the percentage of housing units that were occupied (`occupancy`). These variables were not highly correlated with any other variables and were kept in the model.

Once I compiled my list of variables, I then analyzed the correlation matrix for the final set of predictors to ensure that there were no highly correlated variables that could lead to multicollinearity in the model.

Following this final check, `with_health_insurance`, `poverty`, `occupancy`, `over_25_bachelors_degree`, `unemployment`, `one_worker`, and `native_born` were removed due to high correlations with other variables.

Two Level Modeling

I started with fitting two unconditional models: an unconditional means model and an unconditional growth model. These models were helpful in understanding the variation in SEDA math scores across school districts and over time. Examining this variance would also determine the necessity for hierarchical modeling.

Unconditional Means Model

The unconditional means model evaluates the variation in SEDA math scores across school districts and over time without including any predictors. The model is specified as:

$$
\begin{aligned}
Y_{ij} &= a_0 + \mu_i + \epsilon_{ij} \\
\mu_i &\sim N(0, \sigma^2_\mu) \\
\epsilon_{ij} &\sim N(0, \sigma^2)
\end{aligned}
$$

where Y_{ij} is the SEDA math score for school district i in year j, α_0 is the overall mean SEDA math score, μ_i is the random effect for school district i, and ϵ_{ij} is the residual error.

To determine the necesity for hierarchical modeling, I calculated the intraclass correlation coefficient (ICC) which is the proportion of between-district variance to total variance:

For this model, 76.7% of the total variance in SEDA math scores is due to differences between school districts. As such, this supports the need for hierarchical modeling due to the nested structure of the data.

Unconditional Growth Model

The unconditional growth model (UGM) extends the unconditional means model by including a linear time component `yearsince2019` to estimate how SEDA math scores change over time. I also included a quadratic time component `yearsqrdsince2019` to account for potential nonlinear trends in SEDA math scores over time since I know from prior research and EDA that there was a general improvement in academic performance from 2022 to 2023. The model is specified as:

$$\begin{aligned}
&Y_{ij} = a_0 + b_i \times \text{yearsince2019} + c_i \times \text{yearsqrdsince2019} + \mu_i + \epsilon_{ij} \\
&\mu_i \sim N(0, \sigma_\mu^2) \\
&\epsilon_{ij} \sim N(0, \sigma^2)
\end{aligned}$$

where b_i is the fixed effect of time on SEDA math scores and c_i is the quadratic effect of years since 2019 on SEDA math scores. The random effect μ_i still represents the random intercept for school district. ϵ_{ij} is the residual error.

Three Level Modeling

I then moved on to fitting a three-level hierarchical model as the data had shown considerable variance at the school district and county level. Multilevel hierarchical modeling would be able to explain the variance at the school district and county level and provide more accurate estimates of the fixed effects.

Unconditional Means Model

Revising my unconditional means model to account for county-level variation:

$$\begin{aligned}
Y_{ijk} &= \alpha_0 + \mu_i + \tau_{ij} + \epsilon_{ijk} \\
\mu_i &\sim N(0, \sigma_\mu^2) \\
\tau_{ij} &\sim N(0, \sigma_\tau^2) \\
\epsilon_{ijk} &\sim N(0, \sigma^2)
\end{aligned}$$

where Y_{ijk} is the SEDA math score for county i in school district j in year k, α_0 is the overall mean SEDA math score, μ_i is the random effect for county i, τ_{ij} is the random effect for school district j in county i, and ϵ_{ijk} is the residual error.

Variable Selection

Given my previously curated list of predictors, I created a model that included all of these variables to determine their significance and effect on SEDA math scores.

The full model was specified as:

$$Y_{ijk} = B_1X_1 + B_2X_2 + ... + B_{19}X_{19} + \mu_i + \tau_{ij} + \epsilon_{ijk}$$

With $B_1, ..., B_{19}$ representing the fixed effects for the 19 predictors and $X_1, ..., X_{19}$ representing the 19 predictors.

After which I filtered the predictors to only include those with a t-value greater than 2. I then created a new reduced model with these 8 predictors. The model was compared to the previous model using a drop-in-deviance test to determine if the reduced set of predictors significantly reduced the model fit. A drop-in-deviance test revealed a p-value less than 0.0 when comparing the two models, revealing that the reduced model had significantly worse performance. As such, the full model was selected as the final model.

Random Effects

Once I had finalized the fixed effects for my model, I then explored the random effects to understand the variation at the school district and county levels.

For all random effects I tested, I fit a 95% parametric boostrap confidence interval to the model coefficients to determine the significance of the random effects. In addition to the random slopes for year, I also explored random intercepts for school district and county level variables. Variables such as `share_virtual` were tested as random slopes because it is reasonable to assume that school districts or counties may have been more or less affected by the shift to virtual learning.

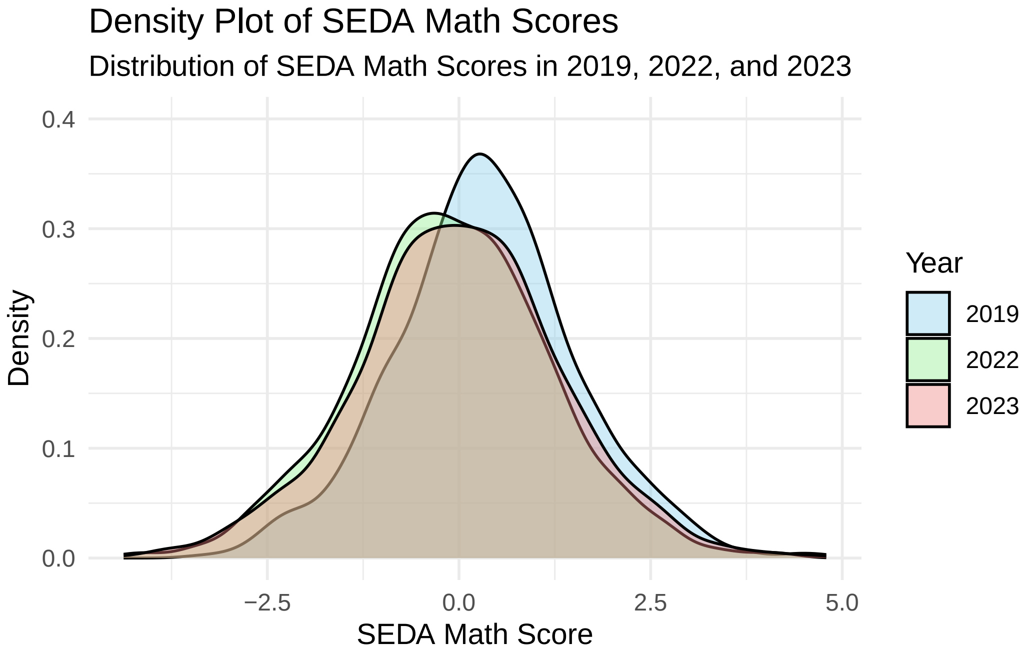
Results

Integrated Dataset Analysis

In addition to the hierachical modeling for predicting academic performance, I needed to understand the distributions and relationships between the features. I used the `ggplot2` (Wickham 2016) package to create histograms, scatter plots, and other visualizations to understand the data better. I also used the `dplyr` (Wickham et al. 2023) package for data manipulation and summarization.

First, I visualized SEDA math scores across school districts over time. I created several histograms that showed the distribution of SEDA math scores across school districts for each year contained in the data. Additionally, I examined relationships between SEDA math scores and my predictors such as school modality, socioeconomic variables, and social capital.

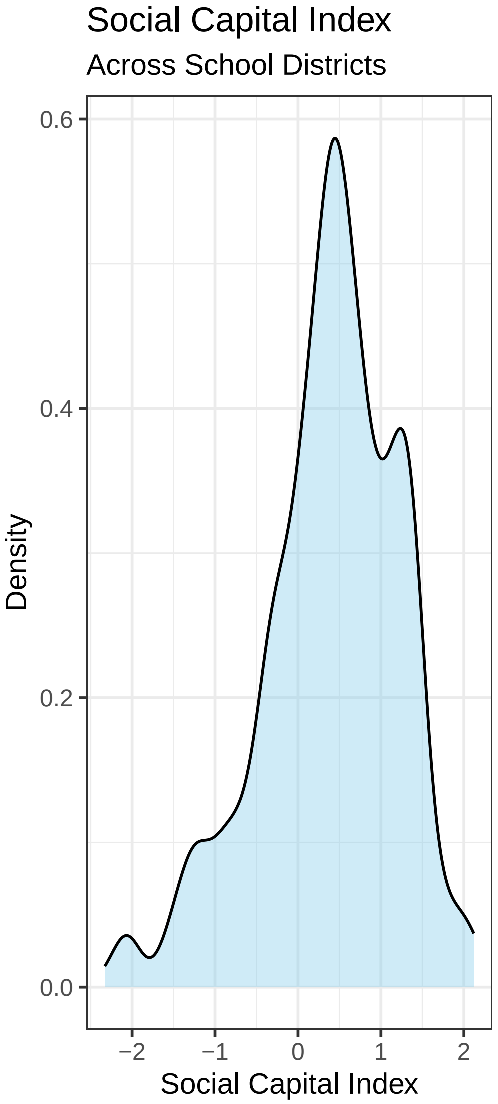
SEDA Math Score Distribution

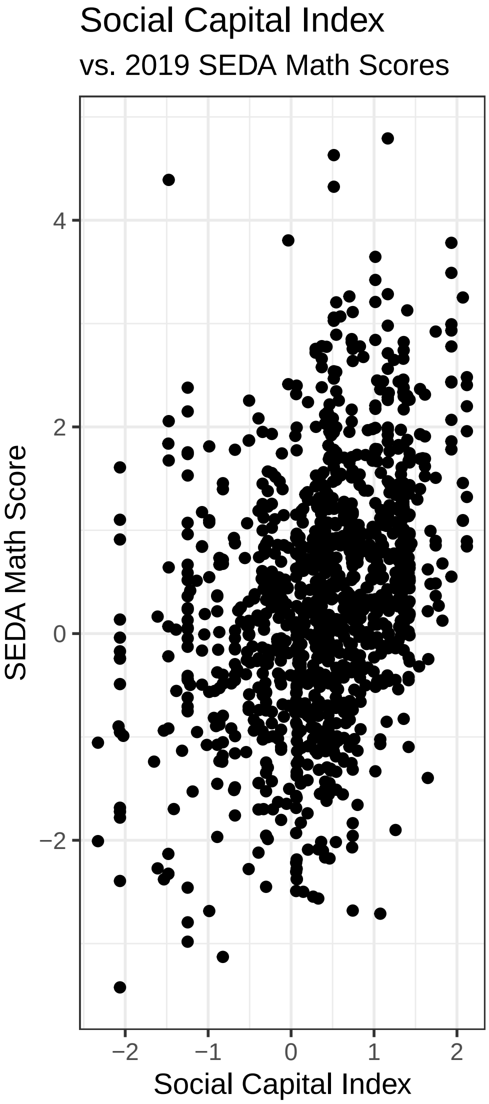


The distribution of SEDA math scores is noticeably shifted to the left in 2022 and 2023 compared to 2019. This indicates that the average SEDA math scores (in terms of standard deviations from the 2019 national average) across school districts decreased in this time period. This is consistent with the findings of other research that has shown a decline in academic performance during the COVID-19 pandemic.

Social Capital Index

This dataset also introduced new types of data of interest such as the Social Capital Index from the Social Capital Project dataset. This index is a composite measure of social capital that includes indicators such as family structure, religious attendance, social cohesion, and institutional trust compiled in 2017. I wanted to explore the relationship between this index and academic performance.



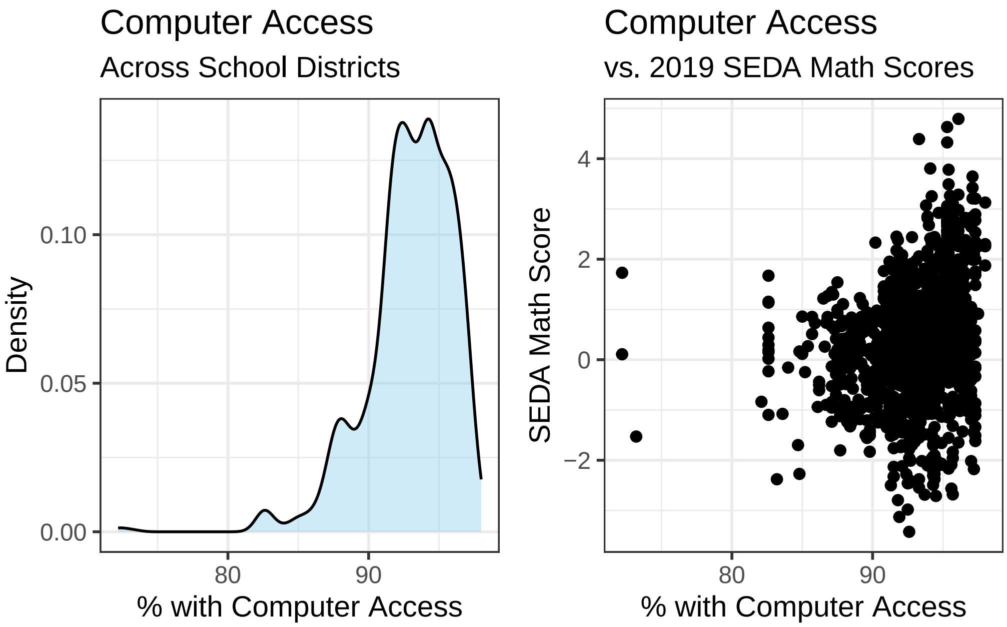


In the univariate distribution plot, the Social Capital Index appears to be approximately unimodal with a slight skew to the left

This scatter plot shows a weak positive relationship between the social capital index and SEDA math scores in 2019. This suggests that school districts with higher levels of community activity, trust in government, and family stability tend to have higher SEDA math scores. As such, this remained a key variable of interest for further analysis.

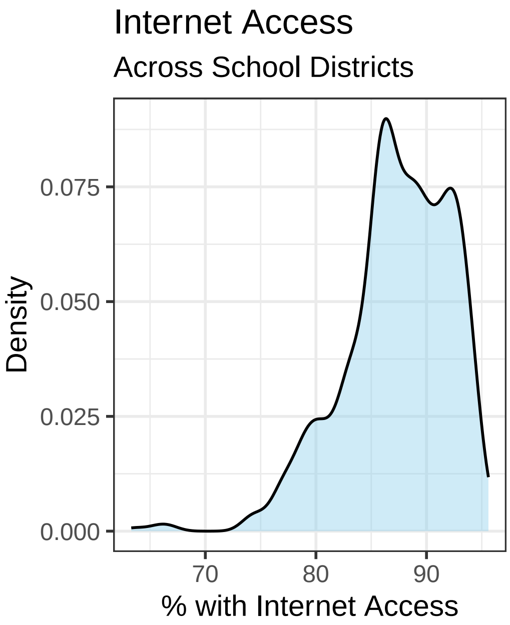
Technology Use

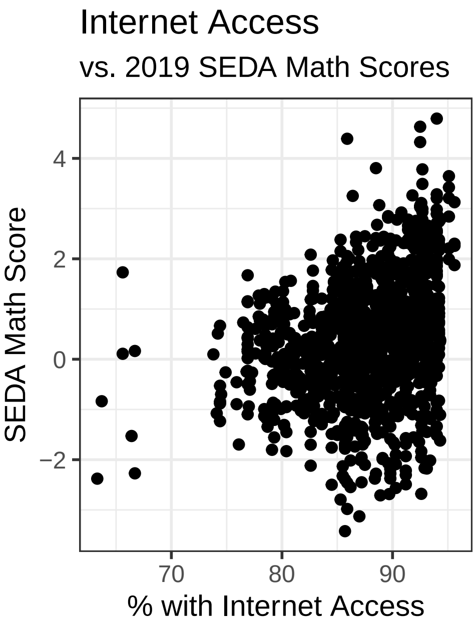
I also wanted to explore the relationship between technology use and academic performance. This data was from the ACS 2018-2022 5 year estimates. I wanted to examine how access to technology such as computers and the internet might impact SEDA math scores.



There does appear to be a weak positive relationship between computer access and SEDA math scores. This suggests that school districts with higher computer access tend to have higher SEDA math scores. This relationship is consistent with the idea that access to technology can positively impact academic performance.

Of the three school districts with the lowest percentages of reported computer access, two of them were from the same county (LaGrange County, Indiana) that had a high second language (`non_english`) speaking rate of 46.1% despite the native-born rate (`native_born`) being 98.0%. This county also has the lowest highschool completion rate (`over_25_highschool_degree`) in the dataset at 60.6%.



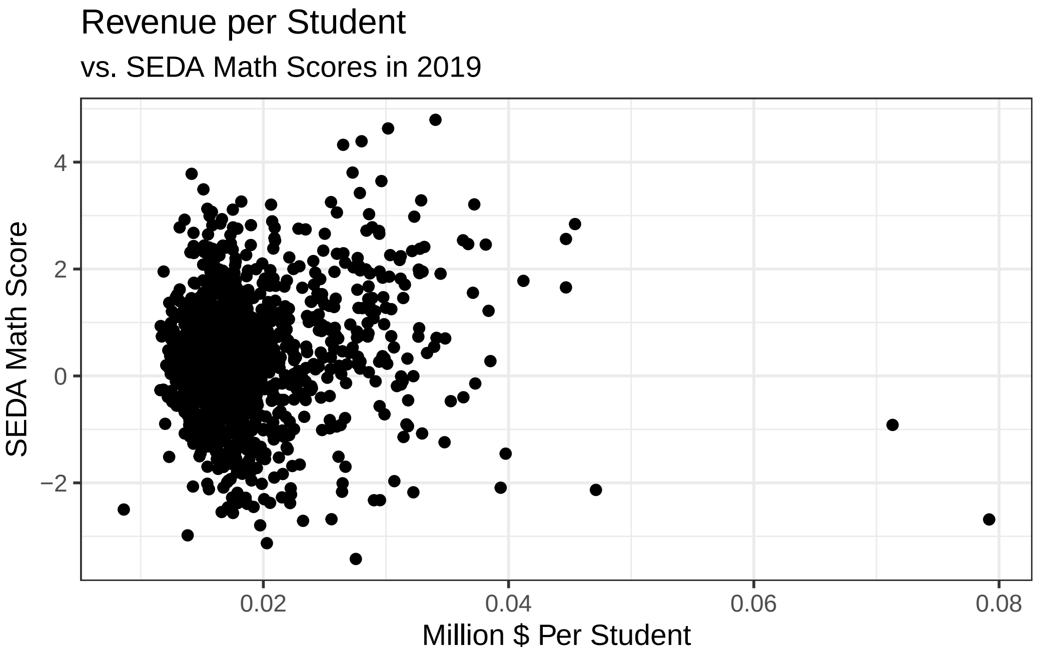


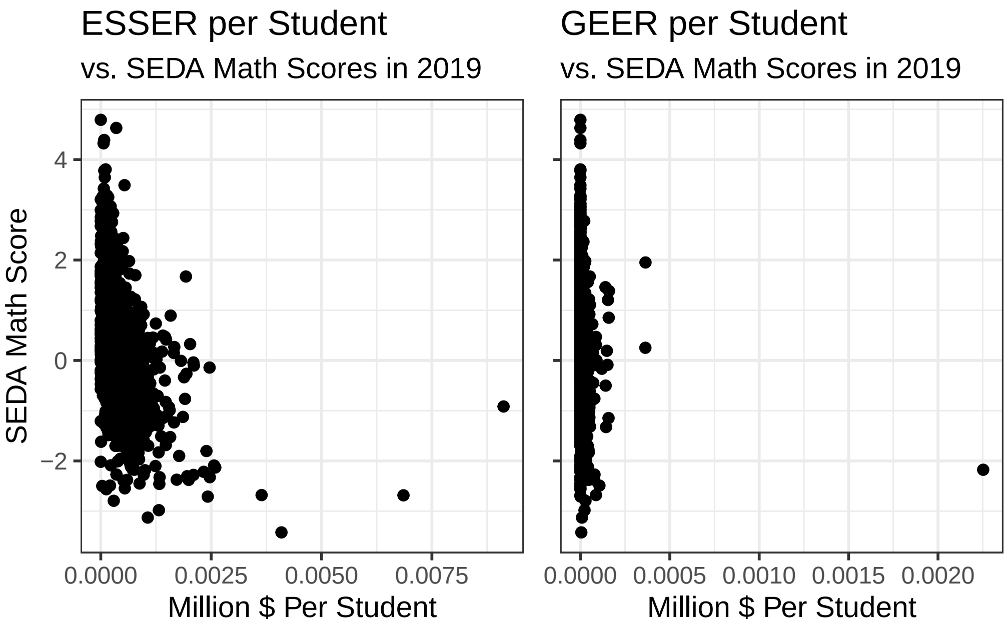
The scatterplot for internet usage also seems to suggest a weak positive relationship with SEDA math scores. This indicates that school districts with higher internet access tend to have higher SEDA math scores.

Revenue and Funding

I also wanted to explore the relationship between revenue and funding and academic performance. In particular, I was interested in the COVID-19 emergency relief funding (ESSER and GEER) that was provided to schools during the pandemic. I also wanted to examine the level of funding per student and how that might impact SEDA math scores.

revenue_per_student	esser_per_student	geer_per_student
Min. : 8621	Min. : 0.0	Min. : 0.00
1st Qu.:15434	1st Qu.: 170.9	1st Qu.: 0.00
Median :17510	Median : 344.4	Median : 0.00
Mean :19039	Mean : 454.0	Mean : 11.87
3rd Qu.:20650	3rd Qu.: 582.4	3rd Qu.: 12.22
Max. :79193	Max. :9123.4	Max. :2252.76





The summary statistics for each kind of funding variable show that the the vast majority of school districts fall into a similar range except for a small number of outliers. The scatter plots for revenue per student does not indicate any clear relationship with SEDA math scores. However, the scatter plots for ESSER and GEER funding per student show a distinct negative relationship with SEDA math scores. This suggests that school districts that received more emergency relief funding were associated with lower SEDA math scores.

The outliers on each of these plots do not always represent the same schools. For example, the school districts with the ten highest revenue per student statistics all received no GEER funding. The extreme outlier with the highest GEER funding per student received was Ypsilanti Community Schools district in Michigan with only 3078 students and had the second lowest SEDA math score in 2022 at -4.256 grade levels below the 2019 national average.

Two Level Modeling

Unconditional Means Model

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.0862806	0.0368079	2.344077
ran_pars	sedaadmin	sd___(Intercept)	1.2282861	NA	NA
ran_pars	Residual	sd___Observation	0.3740910	NA	NA

The two level unconditional means model produced an intercept of 0.086, representing the mean SEDA math score across all districts and years (in terms of standard deviations from the 2019 national average). The variance components were 1.23 and 0.374 for between-district and within-district variance, respectively.

$$ICC = \frac{\text{Between-district variance}}{\text{Between-district variance} + \text{Within-district variance}} = \frac{1.23}{1.23 + 0.374} = 0.767$$

From the ICC formula, it can be observed that 76.7% of the total variance in SEDA math scores was due to differences between school districts and thereby justified the need for hierarchical modeling.

Unconditional Growth Model

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.3493181	0.0374740	9.321601
fixed	NA	yearsince2019	-0.3426995	0.0141053	-24.295757
fixed	NA	yearsqrdsince2019	0.0643914	0.0036608	17.589419
ran_pars	sedaadmin	sd___(Intercept)	1.2356911	NA	NA
ran_pars	Residual	sd___Observation	0.2919050	NA	NA

The two level unconditional growth model yielded -0.343 as the coefficient for year (yearsince2019), showing that SEDA math scores have generally been declining over time. The coefficient for years since 2019 squared (yearsqrdsince2019) was also 0.064 indicating a recovery over longer periods of time. The variance components were similar to the UMM model with 1.24 and 0.292 for between-district and within-district variance, respectively. The ICC was 0.611, further supporting the need for hierarchical modeling.

Three Level Modeling

Unconditional Means Model

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.0093913	0.0521749	0.1799966
ran_pars	sedaadmin	sd___(Intercept)	1.0162011	NA	NA
ran_pars	county_state	sd___(Intercept)	0.6436244	NA	NA
ran_pars	Residual	sd___Observation	0.3740910	NA	NA

The three level unconditional means model produced an intercept of 0.0094, representing the mean SEDA math score across all districts, counties, and years. The variance components were 0.644, 1.016, and 0.374 for between-county, between-district, and within-district variance, respectively. This model shows that the variance across counties is smaller than the variance across districts, but is greater than the within-district variance. This supports the need for hierarchical modeling that includes both district and county level random effects.

Full Model

The full model with all curated variables produced many coefficients that were not signficant when examining the t-statistic. When trying a model with only the significant variables (t-statistics ≥ 2), the reduced model did had significantly reduced model fit as indicated by the p-value (0.0034) from the drop in deviance test.

Table 6: Full Model Output

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.0881	2.2667	0.0389
fixed	NA	yearsince2019	-0.3427	0.0141	-24.2958
fixed	NA	yearsqrdsince2019	0.0644	0.0037	17.5894
fixed	NA	share_virtual	-0.0169	0.0016	-10.3812
fixed	NA	revenue_per_student	0.0011	0.0008	1.4608
fixed	NA	esser_per_student	-0.0813	0.0063	-12.8045

effect	group	term	estimate	std.error	statistic
fixed	NA	geer_per_student	-0.1989	0.0386	-5.1524
fixed	NA	County_Level_Index	0.3161	0.0845	3.7431
fixed	NA	married_household	0.0087	0.0102	0.8514
fixed	NA	over_25_highschool_degree	0.0127	0.0183	0.6944
fixed	NA	with_computer	-0.0207	0.0212	-0.9785
fixed	NA	population	0.0000	0.0002	0.0082
fixed	NA	sex_ratio	-0.0093	0.0068	-1.3687
fixed	NA	hispanic_percent	0.0026	0.0114	0.2279
fixed	NA	white_percent	0.0104	0.0126	0.8247
fixed	NA	black_percent	0.0274	0.0132	2.0795
fixed	NA	asian_percent	0.0516	0.0204	2.5253
fixed	NA	no_workers	-0.0077	0.0096	-0.7948
fixed	NA	median_income	0.0087	0.0044	1.9830
ran_pars	sedaadmin	sd__(Intercept)	0.8646	NA	NA
ran_pars	county_state	sd__(Intercept)	0.3150	NA	NA
ran_pars	Residual	sd__Observation	0.2919	NA	NA

Table 7: Reduced Model Output

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.4374	0.0710	6.1647
fixed	NA	yearsince2019	-0.3427	0.0141	-24.2958
fixed	NA	yearsqrdsince2019	0.0644	0.0037	17.5894
fixed	NA	share_virtual	-0.0180	0.0015	-11.6354
fixed	NA	esser_per_student	-0.0780	0.0054	-14.4420
fixed	NA	geer_per_student	-0.2009	0.0387	-5.1876
fixed	NA	County_Level_Index	0.5336	0.0631	8.4629
fixed	NA	black_percent	0.0226	0.0052	4.3207
fixed	NA	asian_percent	0.0802	0.0107	7.5289
ran_pars	sedaadmin	sd__(Intercept)	0.8676	NA	NA
ran_pars	county_state	sd__(Intercept)	0.3401	NA	NA
ran_pars	Residual	sd__Observation	0.2919	NA	NA

Table 8: Drop in Deviance Test

term	npar	AIC	BIC	logLik	deviance	statistic	df	p.value
reduced_model	12	5232.66	5306.39	-2604.33	5208.66	NA	NA	NA
full_model	22	5221.84	5357.02	-2588.92	5177.84	30.82	10	6e-04

The full model had many insignificant predictors which will not be covered in the results section.

One interesting observation was that while sex_ratio was not a significant predictor in the full model, removing black_percent from the model actually made this variable significant. This suggests that there may be some correlation between these two variables that was not clearly captured in the EDA.

The negative coefficients for yearsince2019 and share_virtual reinforce how SEDA math scores have been declining over time and that schools with greater proportion of time spent fully virtual experienced greater learning loss. In contrast, the coefficient for yearsqrdsince2019 was positive indicating a recovery in SEDA math scores over longer periods of time.

Furthermore, the negative coefficients for esser_per_student and geer_per_student suggest that school districts with higher emergency funding per student had lower SEDA math scores. This is an important finding as it highlights a trend that may have been overlooked in the literature. The coefficients are very large due to the scaling of the variables and should be interpeted as the change in SEDA math scores for a one million dollar increase in funding per student.

The positive coefficient for County_Level_Index suggests that school districts in counties with less crime, greater family unity, social support, and institutional health tended to have higher SEDA math scores. Additionally, black_percent and asian_percent both have positive coefficients which indicates that school districts in counties with higher percentages of Black and Asian populations had higher SEDA math scores. The positive coefficient for median_income also strongly indicates a link between school districts in areas with higher median incomes and with higher SEDA math scores.

Random Effects

The confidence interval for the random effect intercept by district was 0.8194 to 0.9016 indicating that there was some variability across school districts. Whereas the random intercept by county was 0.2096 to 0.3893 which suggests that there was less variability across counties. This is consistent with the variance components from the unconditional means model which showed that the variance across counties was smaller than the variance across districts.

Table 9: Parametric Bootstrap Confidence Intervals for Full Model

	2.5 %	97.5 %
sd_(Intercept)|sedaadmin	0.8192	0.9026
sd_(Intercept)|county_state	0.2125	0.3928
sigma	0.2833	0.3002
(Intercept)	-4.2254	4.2699
yearsince2019	-0.3703	-0.3155

	2.5 %	97.5 %
yearsqrdsince2019	0.0574	0.0717
share_virtual	-0.0204	-0.0133
revenue_per_student	-0.0004	0.0026
esser_per_student	-0.0945	-0.0684
geer_per_student	-0.2775	-0.1204
County_Level_Index	0.1659	0.4883
married_household	-0.0097	0.0285
over_25_highschool_degree	-0.0239	0.0510
with_computer	-0.0609	0.0222
population	-0.0003	0.0003
sex_ratio	-0.0230	0.0043
hispanic_percent	-0.0209	0.0252
white_percent	-0.0147	0.0355
black_percent	0.0017	0.0546
asian_percent	0.0082	0.0853
no_workers	-0.0269	0.0132
median_income	0.0010	0.0166

The confidence interval for the random effect for slope of `yearsince2019` was 0 to 0.0820 which includes 0 and suggests that the random effect for this variable was not significant.

Table 10: Parametric Bootstrap Confidence Intervals for Full Model with Random Effects

	2.5 %	97.5 %
sd_(Intercept)|sedaadmin	0.8246	0.9081
sd_(Intercept)|county_state	0.2251	0.3908
sd_(Intercept)|yearsince2019	0.0000	0.3685
sigma	0.2839	0.3001
(Intercept)	-4.8395	4.2106
yearsince2019	-2.0423	1.5148
yearsqrdsince2019	-0.4201	0.4928
share_virtual	-0.0199	-0.0134
revenue_per_student	-0.0005	0.0026
esser_per_student	-0.0940	-0.0682
geer_per_student	-0.2779	-0.1280
County_Level_Index	0.1563	0.4772
married_household	-0.0120	0.0295
over_25_highschool_degree	-0.0218	0.0462
with_computer	-0.0603	0.0187
population	-0.0003	0.0003

	2.5 %	97.5 %
sex_ratio	-0.0224	0.0052
hispanic_percent	-0.0182	0.0240
white_percent	-0.0121	0.0369
black_percent	0.0052	0.0573
asian_percent	0.0083	0.0894
no_workers	-0.0269	0.0106
median_income	-0.0004	0.0172

The confidence interval for the random intercept for `share_virtual` was 0.0000 to 0.2426 which also contains 0 and suggests that the random effect for this variable was not significant.

Table 11: Parametric Bootstrap Confidence Intervals for Full Model with Share Virtual as a Random Effect

	2.5 %	97.5 %
sd_(Intercept)|sedaadmin	0.8193	0.9023
sd_(Intercept)|county_state	0.2080	0.3888
sd_(Intercept)|share_virtual	0.0000	0.2398
sigma	0.2832	0.3002
(Intercept)	-4.7624	4.1527
yearsince2019	-0.3711	-0.3150
yearsqrdsince2019	0.0573	0.0721
share_virtual	-0.0225	-0.0144
revenue_per_student	-0.0005	0.0026
esser_per_student	-0.0946	-0.0693
geer_per_student	-0.2781	-0.1200
County_Level_Index	0.1370	0.4657
married_household	-0.0128	0.0279
over_25_highschool_degree	-0.0225	0.0458
with_computer	-0.0581	0.0181
population	-0.0003	0.0004
sex_ratio	-0.0217	0.0060
hispanic_percent	-0.0208	0.0229
white_percent	-0.0161	0.0333
black_percent	0.0008	0.0514
asian_percent	0.0077	0.0912
no_workers	-0.0254	0.0127
median_income	0.0012	0.0170

Other random effects were tested and the models frequently failed to converge or produced confidence intervals that included zero. This suggests that the fixed effects captured most of the variance in SEDA math scores and the random effects for the predictors did not provide much additional information. As such, the final model included only the random intercepts across school districts and counties.

Discussion

Overall, the hierarchical modeling approach was successful in explaining the variance in SEDA math scores across school districts and over time. Multilevel model assumptions were met for residuals and all levels of predictors. The model provided valuable insights into the demographic, socioeconomic, and educational factors that influence academic performance.

The preliminary unconditional means models demonstrated the need for hierarchical modeling due to the high proportion of variance between school districts and counties. The unconditional growth models showed that SEDA math scores have been declining over time but with a slight recovery over longer periods which is consistent with current literature on the impact of the COVID-19 pandemic on academic performance. This general recovery of academic achievement in math has even been observed in 2024 school year although still failing to reach pre-pandemic growth rates (Lewis and Kuhfeld, n.d.).

From my full model, I found that emergency funding per student (both from ESSER and GEER) was negatively associated with SEDA math scores. While seemingly counterintuitive, this can be explained by how these funds were allocated in accordance with Title I funding which is based on the number of low-income students in a district. This suggests that school districts with more low-income students received more emergency funding but also had lower SEDA math scores. This is an important finding as it highlights the need for targeted interventions to support low-income students in the wake of the pandemic. The effect of virtual learning is also an important finding as it was negatively associated with SEDA math scores. This suggests that school districts that relied more on virtual learning during the 2020-2021 school year had lower SEDA math scores. This finding aligns with other research that has shown the negative impact of virtual learning on academic performance (Kane and Reardon 2023).

Percentage of black students as well as Asian students were both positively associated with SEDA math scores which is an interesting finding as it seems schools with more diverse student populations could experience higher academic performance.

The County Level Index was a significant predictor of SEDA math scores. This is an important finding as it highlights the role of community factors in academic performance, not something that has been extensively covered by previous research. Higher social and familial engagement, trust in government, and community support play an important role in shaping academic performance and should be considered in future research and policy decisions.

Through exploring random effects, I found that the effect of year on SEDA math scores was more variable across counties. This suggests that county characteristics may have a significant impact on how school districts within that county are affected by changes in academic performance over time. This is an area for future research to explore further.

An important limitation of this study is the lack of data from all states. The data was only available for 10 states which may not be representative of the entire country. Additionally, many of the data points were aggregated at the county level which may have reduced the granularity of the analysis. Future research could benefit from more comprehensive data that includes better coverage of states and school districts. This could be achieved through careful mean imputation and more lenient filtering criteria to include more school districts in the analysis. Additionally, the jaccard and stringdist thresholds I used in this project were quite strict and could be relaxed to include more school districts. Datasets from the CCD and Census both contained a unique identifier per school district which could have been used to match datasets more accurately as well as evaluate different matching methods.

Future research could conduct similar analysis with reading scores from the same dataset and researching the impact of the pandemic on other academic subjects. There is also the potential to investigate interaction terms between variables to better understand the relationship between these factors and academic performance. Other potential areas of research could involve school staffing vacancies. The NAEP found that 37% of public schools were operating with at least one teaching vacancy as of October 2023 of which 21% are dealing with multiple vacancies (De La Rosa, n.d.). This could be a potential area of research to understand the impact of staffing shortages on academic performance as it was affected by the pandemic as well as post-pandemic recovery.

Conclusion

The COVID-19 pandemic has undoubtedly had a significant impact on academic performance across the United States. While academic achievement in math has generally been declining over time since the pandemic, there are signs of recovery in recent years. School districts with higher emergency funding per student and reliance on virtual learning were negatively associated with SEDA math scores, highlighting the need for targeted interventions to support low-income students and address the challenges of virtual learning. My thesis aims to explain the variance in SEDA math scores across school districts and over time, providing valuable insights into the demographic, socioeconomic, and educational factors that influence academic performance.

References

Atkins, Aron, Toph Allen, Hadley Wickham, Jonathan McPherson, and JJ Allaire. 2024. “Rsconnect: Deploy Docs, Apps, and APIs to ’Posit Connect’, ’Shinyapps.io’, and ’RPubs’.”

https://CRAN.R-project.org/package=rsconnect.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. "Shiny: Web Application Framework for r." https://CRAN.R-project.org/package=shiny.

De La Rosa, Josh. n.d. "Press Release - Forty-Four Percent of Public School Students Began 2023-24 Year Behind Grade Level in at Least One Academic Subject, Principals Say - December 14, 2023." https://nces.ed.gov/whatsnew/press_releases/12_14_2023.asp.

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. 2022. "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.

Lewis, Karyn, and Megan Kuhfeld. n.d. "Recovery Still Elusive: 2023-24 Student Achievement Highlights Persistent Achievement Gaps and a Long Road Ahead."

"Local Education Agency (School District) Finance Survey (f-33) Data, (v.1a)." 2022. https://nces.ed.gov/ccd/Data/zip/sdf22_1a_sas7bdat.zip.

Loo, M. P. J. van der. 2014. "The Stringdist Package for Approximate String Matching" 6: 111–22. https://CRAN.R-project.org/package=stringdist.

McFarland, Joel. n.d. "Trends in High School Dropout and Completion Rates in the United States: 2019."

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.

"Percentage of School Year Spent in-Person, Hybrid, or Virtual." 2023. https://assets.ctfassets.net/9fbw4onh0qc1/XfBEuMLMOBgHrhmjBdVpc/8e555b362876da16ba52c85be5b2effe/District_Overall_Shares_03.08.23.csv.

Reardon, Ho, S. F. 2024. "Stanford Education Data Archive (Version 5.0)." https://purl.stanford.edu/cs829jn7849.

"School Districts and Associated Counties." 2021. https://www2.census.gov/programs-surveys/saipe/guidance-geographies/districts-counties/sdlist-21.xls.

"Social Capital Project." 2023. https://www.jec.senate.gov/public/index.cfm/republicans/socialcapitalproject.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

Appendix

Geographic Distribution of School Districts

Table 12: Geographic Distribution of Data by State Before Filtering

stateabb	n
CA	596
IL	576
TX	493
PA	471
MI	408
OH	345
NJ	338
WI	325
IN	278
MN	270
MA	221
MO	220
AR	206
GA	173
KS	172
WA	170
KY	156
AL	137
MS	135
TN	130
CT	119
NC	112
NE	104
AZ	96
OK	94
VA	91
NH	80
SC	74
SD	71
ID	69
FL	66

stateabb	n
LA	62
WV	55
WY	40
UT	34
RI	30
MD	24
NV	12
ND	7

Table 13: Geographic Distribution of Data by State After Filtering

stateabb	n
MI	332
WI	262
WA	162
NJ	128
NC	97
IN	86
NH	76
IL	2
MA	2
TN	1

Data Dictionary

Integrated Dataset Variables

Table 14: Data Dictionary

Variable	Description
sedaadmin	School district identifier developed by the Stanford Education Data Archive
seda_district	School district name
stateabb	State abbreviation
subject	Subject of the test (math or reading language arts)
gys_mn_2019_ol	SEDA math score in 2019
gys_mn_2022_ol	SEDA math score in 2022
gys_mn_2023_ol	SEDA math score in 2023

Variable	Description
County Names	County name
share_inperson	Percentage of time a school spent in person
share_hybrid	Percentage of time a school spent in a hybrid model
share_virtual	Percentage of time a school spent fully virtual
membership	Total student enrollment
total_revenue	Total revenue from all sources (federal, state, local)
total_state_revenue	Total revenue from state sources
total_fed_revenue	Total revenue from federal sources
total_local_revenue	Total revenue from local sources
total_salaries	Total salaries paid by the school district
total_instructional_salaries	Total salaries paid to instructional staff
total_esser1	Total funding received from ESSER I
total_esser2	Total funding received from ESSER II
total_arp_esser	Total funding received from ARP ESSER
total_geer1	Total funding received from GEER I
total_geer2	Total funding received from GEER II
County_Level_Index	Composite index measuring county-level characteristics
Family_Unity	Measure of family stability within the county
Community_Health	Measure of overall health outcomes within the county
Institutional_Health	Measure of institutional strength and effectiveness in the county
Collective_Efficacy	Measure of community trust and cooperation
married_household	Percentage of households that are married
married_household_children	Percentage of married households with children
male_married	Percentage of males who are married
female_married	Percentage of females who are married
male_never_married	Percentage of males who have never been married
female_never_married	Percentage of females who have never been married
male_divorced	Percentage of males who are divorced
female_divorced	Percentage of females who are divorced
over_25_highschool_degree	Percentage of individuals over 25 with a high school diploma
over_25_bachelors_degree	Percentage of individuals over 25 with a bachelor's degree
native_born	Percentage of the population that is native-born
only_english	Percentage of the population that speaks only English at home
non_english	Percentage of the population that speaks a language other than English at home
with_computer	Percentage of households with a computer
with_internet	Percentage of households with internet access
unemployment	Unemployment rate in the county
median_income	Median household income in the county
mean_income	Mean household income in the county
with_health_insurance	Percentage of the population with health insurance coverage

Variable	Description
poverty	Percentage of the population living in poverty
owner_occupied_value	Median value of owner-occupied housing units
occupancy	Percentage of housing units that are occupied
SMOC	Selected Monthly Owner Costs
rent	Median rent paid by renters
mortgage_percentage	Percentage of households with a mortgage
population	Total population of the county
sex_ratio	Females per 100 males
hispanic_percent	Percentage of the population that is Hispanic
white_percent	Percentage of the population that is White
black_percent	Percentage of the population that is Black
asian_percent	Percentage of the population that is Asian
no_workers	Percentage of households with no workers
one_worker	Percentage of households with one worker
employment_past_year	Percentage of the working-age population employed in the past year

Derived Variables

Table 15: Derived Variables Data Dictionary

Variable	Description
yearsince2019	Number of years since 2019
yearsqrdsince2019	Square of years since 2019
revenue_per_student	Total revenue per student
inst_salaries_per_student	Total instructional salaries per student
esser_per_student	Total ESSER funds per student
geer_per_student	Total GEER funds per student

Model Assumptions

Model Specification

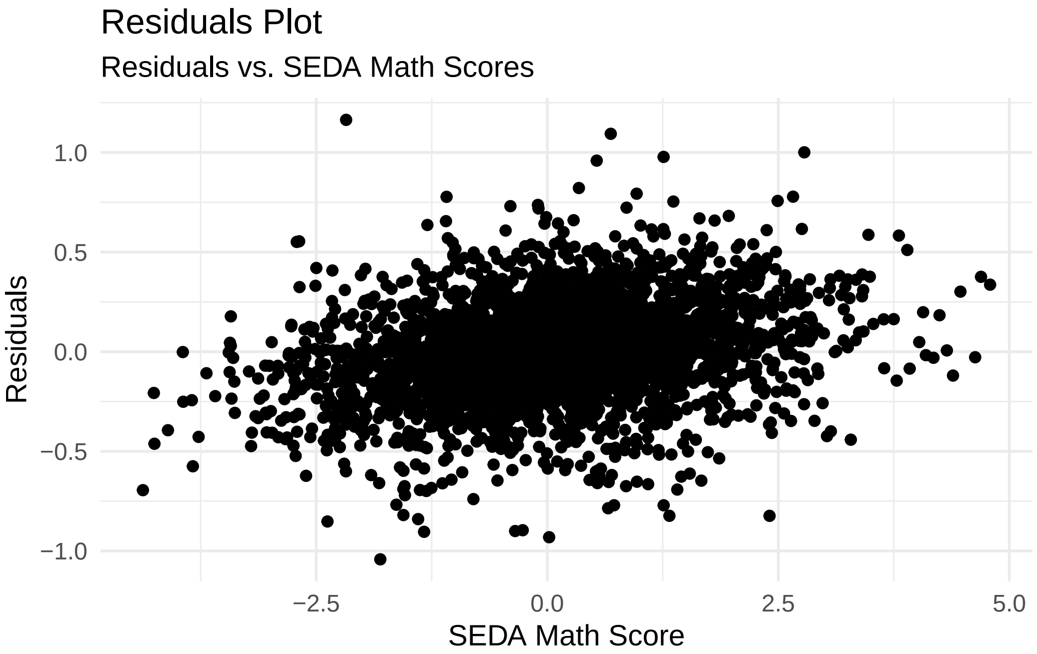
Final model went through several iterations to determine a strong set of predictors given the data. This included trying a full model with all non highly correlated variables to eventually being trimmed down to a reduced model with only significant predictors.

Functional Form

As seen in the Integrated Dataset EDA section, many of the predictors seem to have weak positive or negative linear relationships with the SEDA math scores. This was taken into account when specifying the model.

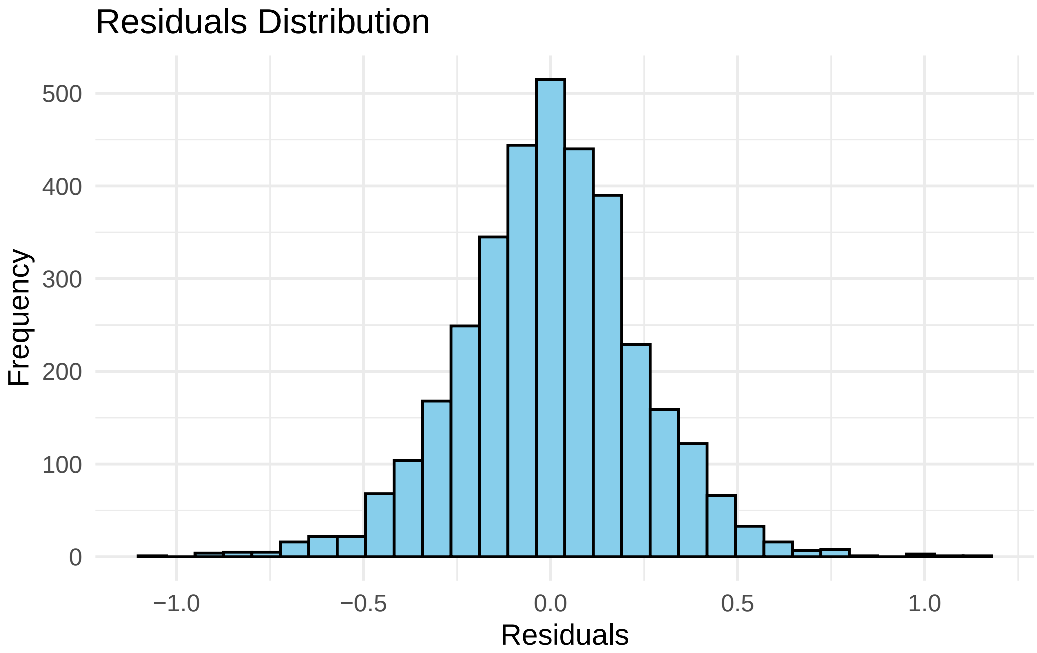
Residuals are Independent and Normally Distributed

The residuals were plotted against the SEDA math score with no discernible pattern. The correlation tests further show that the residuals are not significantly correlated with the predictor and are therefore independent.


Residuals Plot

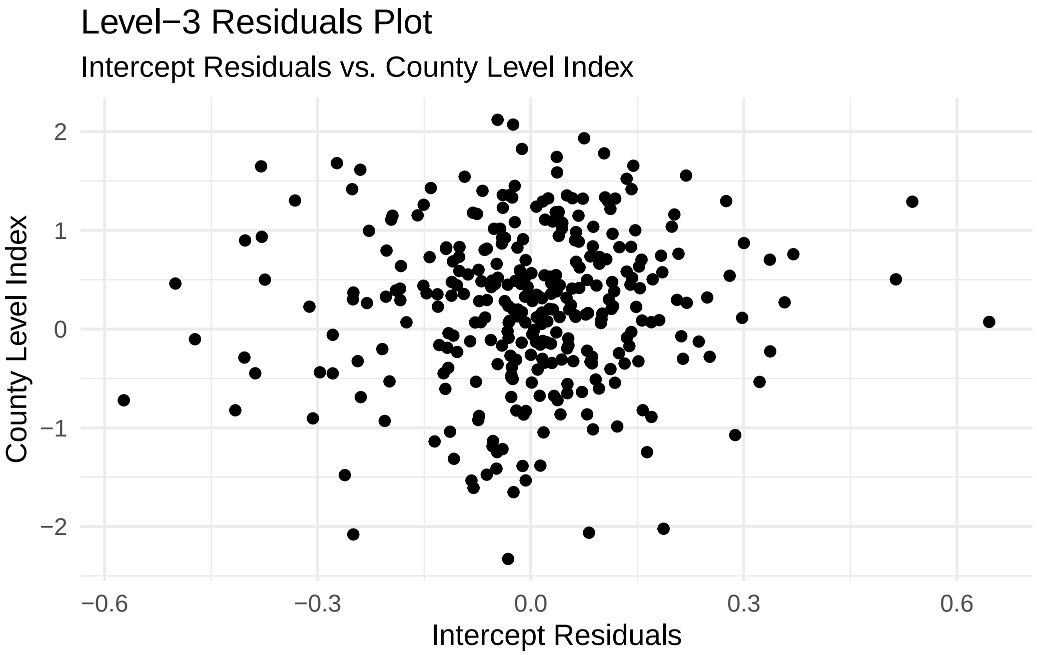
Residuals vs. SEDA Math Scores

The residuals plot also shows a general normal distribution.



Residuals vs Predictors

The residuals were plotted against the County Level Index and revealed little to no correlation. The correlation tests further show that the residuals are not significantly correlated with the predictor and are therefore independent.

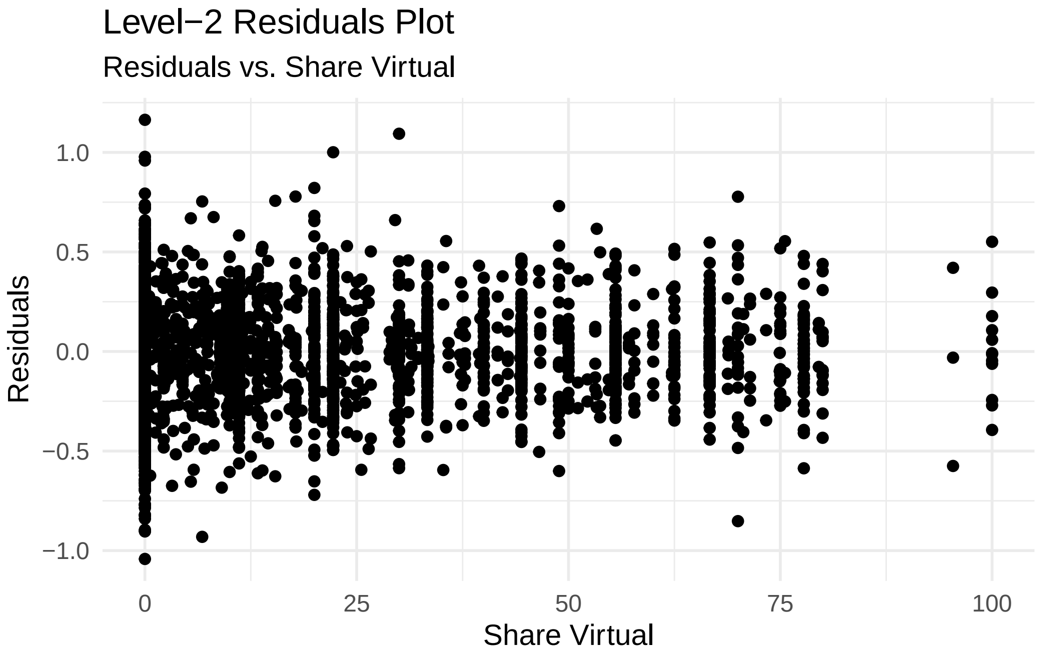


```
    Pearson's product-moment correlation

data:  data$intercept_resid and data$County_Level_Index
t = 1.1897, df = 322, p-value = 0.2351
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04311885  0.17385922
sample estimates:
       cor
0.06615216
```

Residuals vs District Predictors

The residuals were plotted against the share of time spent fully virtual and revealed little to no correlation. The correlation tests further show that the residuals have practically no correlation with `share_virtual`.



```
    Pearson's product-moment correlation

data:  data$resid and data$share_virtual
t = -1.344e-12, df = 3442, p-value = 1
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03339982  0.03339982
sample estimates:
          cor
-2.290854e-14
```