Chicago Public Schools Mathematics and Literacy Skills

Austen Lowitz, Annie DeForge, William Teng

Contents

Introduction

In today's educational landscape, data-driven decision-making has emerged as an integral tool for enhancing the student experience. Leveraging the power of data allows educational stakeholders to identify critical factors that contribute to a student's academic success, allowing for effective interventions and informed policy adjustments[1].

The Chicago Public School District is amongst the largest and most diverse urban school districts in the US. With that comes unique challenges and opportunities in applying data to drive academic improvement. Currently, only 21% of students had proficient ratings on the statewide reading exam, and only 17% had proficient ratings in math[2]. Recognizing this potential, our analysis seeks to uncover evidence-based actionable insights to enhance student performance for Chicago's elementary school students.

Through this analysis, we aim to provide Chicago elementary school superintendents with a robust framework for developing multi-year roadmaps to enhance academic success. This can include changes from staffing and leadership decisions, school security code changes, and teaching staff rostering to disciplinary policies. The end result will be a more responsive, equitable, and effective educational system that empowers students to thrive academically. This work contributes to the broader movement towards data-informed education, embodying the shared vision of educators, policymakers, and researchers alike to harness the transformative potential of data for the betterment of our schools and communities.

This study estimates the average proportion of elementary school students who are meeting progress and proficiency targets in Chicago. The data shows details on the status of the school and the quality of the classroom environment. We propose a set of regression models to show the learning environment variables that are significant to predicting student success.

Data and Methodology

The data for this study is from the 2012 School Progress Report Card for the elementary schools in the Chicago Public School System. It was made public by the city of Chicago. Each row represents an elementary school in the school district. The dataset had 460 rows, however, there were schools that had missing values or had a "not enough data record" in our variables of interest. The average amount of students at a school that were meeting the national average on standardized tests or meeting the expected growth rate was 50% for schools with no missing values and 46% for schools with missing values. The spread was similar for schools with and without missing values as well, although the range was slightly greater for schools with no missing values and thinner tails. We decided that there wasn't any obvious pattern in the data, so we would exclude the schools that had missing values, which brought our total amount of observations to 281.

Average daily student attendance was one of the main variables of interest provided in the dataset. Studies have shown that student attendance is one of the most important factors in student success in the classroom[3]. Motivation in general has been shown to be very important for learning, and attendance is a key measure of a student's academic motivation[4]. Additionally, this dataset also contains a grade for several learning environment variables. Schools were graded on a 5 point scale from "very weak" to "very strong" on how well the school involved families, how supportive the environment was, how safe the school was, if the leadership was focused, if the instruction was focused and challenging, and how well the teachers worked together. Additionally, there was an average daily attendance variable for teachers. These student motivation and learning environment variables were the ones that we were primarily interested in to understand which ones were significantly associated with academic success. We are interested in using a regression model to evaluate which of these classroom experience variables have significant coefficients to explain the variability in academic success.

We also wanted to control for other factors that impact the school that were not directly related to what was in the classroom. To do so, we incorporated an indicator variable showing if the school was on a regular schedule or a shortened summer schedule to reduce learning loss over summer break, the number of years the school had been on probation, disciplinary policy (the number of misconducts resulting in suspension and average days of suspension), and location (longitude and latitude) as variables to consider in our model.

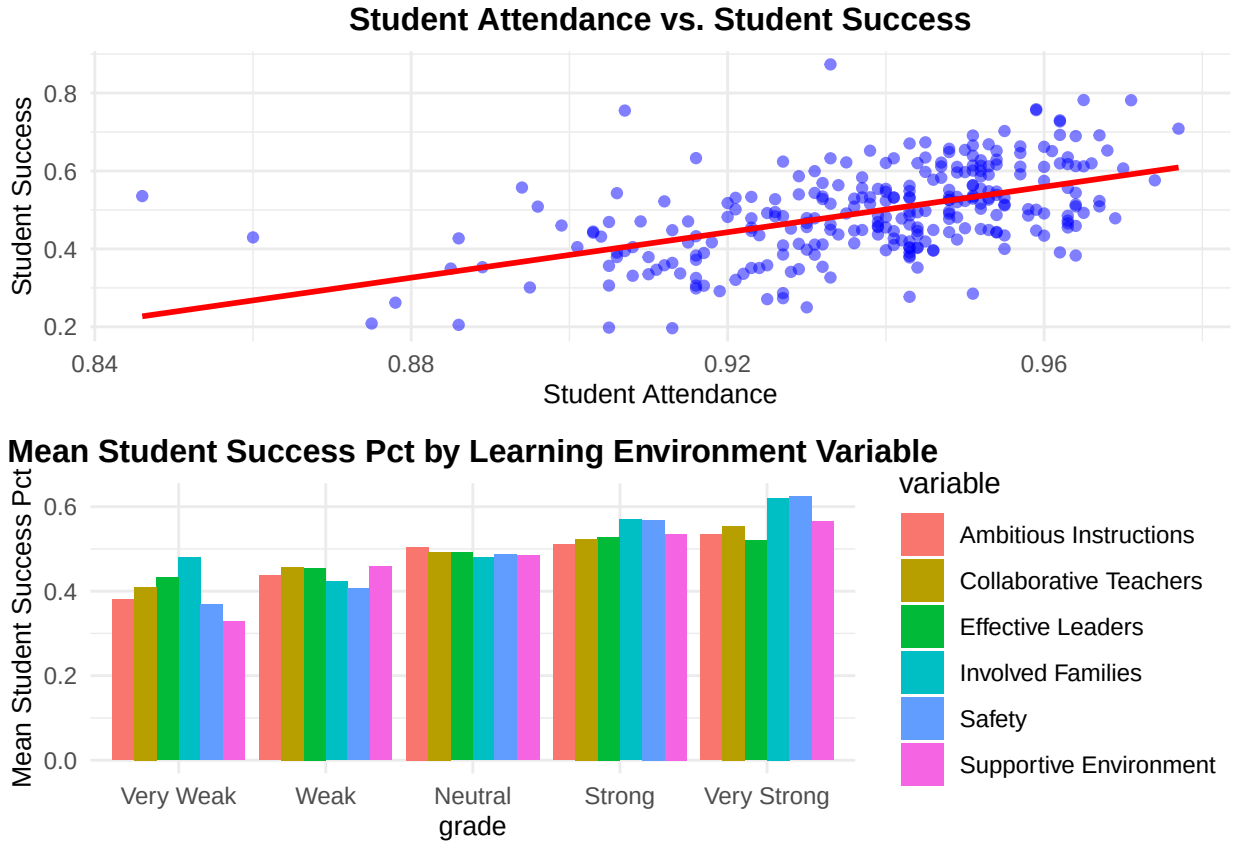
[1]Mandinach, E. B., & Gummer, E. (n.d.). Data-driven decision making: Components of the ... - sage journals. https://journals.sagepub.com/doi/abs/10.1177/016146811511700402

[2]Senior. (2023, June 29). Explore chicago public schools. Niche. https://www.niche.com/k12/d/chicago-public-schools-il/

[3]Every School Day Counts: The Forum Guide to collecting and using attendance data. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/pubs2009/attendancedata/chapter1a.asp.

[4]Davis, L. (2015). STUDENT ATTENDANCE: A SCHOOL'S INTERVENTION TO INCREASE ATTENDANCE. [online] Available at: https://www.nwmissouri.edu/library/researchpapers/2015/Davis,%20Luke.pdf [Accessed 7 Aug. 2023].

We operationalized student success to incorporate two metrics: the proportion of students at the school who met or exceeded the national average on reading and math test, and the proportion of students who improved the expected amount between the fall and spring tests. We averaged these two proportions to create a combined metric on student success, which we used as our outcome variable in our models.



Our exploratory plots showed a roughly linear relationship between our variables of interest and student success. The structure of our model was the following:

$$\hat{\text{Student Success}} = \beta_0 + \beta_1 \cdot \text{Student Attendance} + \text{Y}\gamma + \text{Z}\delta$$

Y is a row vector of learning environment variables and γ is a column vector of the coefficients. Z is a row vector of the additional control covariates and δ is a column vector of the coefficients.

Because of the importance of student attendance, we wanted to run a naive model with just this variable. We ran a wald test for nested models, testing one variable at a time starting with the learning environment variables as they appeared in the dataset to identify the significant learning environment variables and did the same with the other covariates.

Results

Table 1 below compares four models: our "Naive" model with just one predictor variable, Student Attendance, a model containing our learning environment variables, a model with significant learning environment variables plus significant controls, like School Track and Years on Probation. Last but not least, we have a "Complex" model which includes all our predictor variables.

Across all models, the beta coefficient on Student Attendance was highly significant ($p<0.01$), indicating a strong positive relationship with student success. The coefficient point estimate decreases as more features are added to the model, ranging from 2.92 in the Naive model to 1.08 in the Complex model. This shows how the beta coefficient is pushed down by the inclusion of other environment and control variables.

Table 1: Model Comparison

	Regression Table			
	Naive	Learning Environment (LE)	LE w/ controls	Complex
	(1)	(2)	(3)	(4)
Student Attendance	2.92^{***} (0.29)	1.81^{***} (0.29)	1.24^{***} (0.30)	1.08^{***} (0.31)
Suspension Percent				0.02 (0.02)
Suspension Days				−0.01 (0.01)
Teacher Attendance		0.89 (0.58)	0.83 (0.55)	0.99^{*} (0.57)
Longitude				−0.02 (0.07)
Latitude				−0.10 (0.11)
School Schedule (Track E)			-0.03^{**} (0.01)	-0.03^{***} (0.01)
Years on Probation			-0.01^{***} (0.002)	-0.01^{***} (0.002)
Constant	-2.24^{***} (0.27)	-2.08^{***} (0.54)	-1.40^{***} (0.52)	−9.50 (8.27)
Involved Families		✓	✓	✓
Supportive Environment		✓	✓	✓
Safety		✓	✓	✓
Effective.Leaders				✓
Ambitious.Instruction				✓
Collaborative.Teachers				✓
Observations	281	281	281	281
R^2	0.27	0.49	0.55	0.58
Adjusted R^2	0.27	0.47	0.52	0.52
Residual Std. Error	0.10 (df = 279)	0.08 (df = 266)	0.08 (df = 264)	0.08 (df = 248)
F Statistic	103.96^{***} (df = 1; 279)	18.57^{***} (df = 14; 266)	20.25^{***} (df = 16; 264)	10.64^{***} (df = 32; 248)

Note: School Schedule (Track E) is an indicator variable for having a shortened summer schedule. The categorical covariates each have 5 levels, a grade on the category from 'very weak' to 'very strong'.

To illustrate the scale of these effects, consider a hypothetical school with a Student Attendance rate of 90%. Applying the learning environment model with controls, a school that experienced a increase in student attendance by 10 percentage points would be predicted to experience a 12 percentage point increase in the average proportion of students who are meeting or exceeding the national average or meeting expected growth rate in a school year. Although Teacher Attendance was significant in the nested model test when we added it as a second variable, once other significant variables were added in, it lost it's significance. This was due to the standard error being fairly high.

Two variables, School Schedule (Track E) and Years on Probation, appear to have a significant negative relationship on student success in the Controls and Complex models. Years on Probation had a small effect size, but with a very small standard error and it was very robust to different model specifications. A school being on a reduced summer schedule (Track E) was also negatively associated with Student Success. There may be an element of reverse causality here in that a school where students are struggling more is more likely to try a different summer schedule to reduce school break learning loss.

Several variables, including Suspensions Percent, Suspension Days, Longitude, and Latitude, were included in the Complex model but did not show statistical significance. This is due to some extent on our operationalization of Student Success. For instance, there is a significant North-South divide on socio-economic status in Chicago, so longitude had a strong linear relationship with national performance on math and reading tests, but an insignificant relationship with growth on math and reading tests.

From examining the Complex model, we can see some of the negative effects of including insignificant variables. In the complex model, the confidence on Teacher Attendance is likely very overestimated since it increases dramatically from the learning environment with controls model. Additionally, in the complex model, the coefficients of the dummy variables from improving the grade Supportive Environment flips from a positive association with Student Success to a negative one.

Considering the different model specifications, the Learning Environment with controls is the best model because it can give estimates for the classroom related variables that teachers and superintendents will be interested to learn about, while controlling for other school factors, without creating the interpretability problems that would arise from using the less parsimonious complex model.

Limitations

Given our large sample size (218 observations), we can utilize a multiple regression model to assess the key drivers on student success, assuming that our data meets the following statistical limitations: The data must be Independent and Identically Distributed (IID), and there must exist no perfect collinearity between features.

Each row in our dataframe represents a unique elementary school in Chicago. Because the data includes the entire population of elementary schools in Chicago rather than a specific sample, we have reduced concerns related to sampling bias since every school is represented in the dataset. Since we are analyzing the entire population of elementary schools in Chicago, we assume that the schools operate under similar conditions and follow similar educational standards and policies. This supports the notion that our observations are drawn from the same distribution, therefore meeting the IID assumption. That said, our scope is limited towards only being able to make generalizations about Chicago public schools.

We examined the relationships between independent variables through a correlation matrix and found no correlations near 1, indicating no perfect collinearity. We did not observe any correlations that were high enough to be a concern. Additionally, we examined the VIF of all of the possible variables together (the complex model) and saw that all of the variables were below 5, which shows that multicollinearity is not a concern for this data.

Since our data is IID and has no perfect collinearity, we pass all of our large linear model assumptions and therefore rule out any concerns for using a multiple regression model for this analysis.

Regarding structural limitations, one key omitted variable is school funding, which could have significant implications for our model's accuracy. School funding is expected to be positively correlated with our target variable, Student Success. Moreover, more school funding would likely enhance some of our learning environment variables like safety and student attendance. By having school funding as an omitted variable, we may be inadvertently attributing more significance to our existing predictor variables, thus overestimating their impact. Including data on school funding would therefore drive the main effect towards zero, reducing the likelihood of a type I error.

Conclusion

Through our analysis, we can conclude that for elementary students in Chicago public schools, student's success is significantly associated with their school attendance, school status, and learning environment, including variables like Involved Families, Supportive Environment, and Safety.

Since student success can benefit society in many ways, from increasing tax revenue to creating a more vibrant society[5], it is crucial for superintendents to use this information to align priorities when developing future road maps to drive student success. As a starting point, superintendents can focus on increasing student attendance, which had the largest coefficient amongst our significant predictor variables. Crucially, by identifying significant relationships, further randomized experimental research can be done on the association that were found to be signficant through out our investigation in order to establish what the causal relationship is between the learning environment and student's academic success.

That said, addressing the challenge of student attendance requires a multifaceted approach; one that goes beyond simply boosting attendance numbers. It calls for initiatives that foster a safer school environment, cultivate family engagement, and address underlying issues such as safety concerns that may be hindering attendance in the first place. These efforts can resonate beyond attendance, potentially reducing the number of years a school is on probation—another significant predictor of student success in our study.

In conclusion, our study serves as a testament to the power of data-driven decision-making in education. By delving into the complex interplay of factors that contribute to student success, we have provided actionable insights that can be harnessed to foster a more enriching educational experience for elementary school students in Chicago.

[5]How do college graduates benefit society at large?. APLU. (2023, March 1). https://www.aplu.org/our-work/4-policy-and-advocacy/publicuvalues/societal-benefits/