Methodology

In order to compare the correlations between familial and school factors in student outcome, we create models for all of our outcomes of interest so that confounding variables are accounted for. We use models that include predictors from both groups so that the effect sizes of variables of interest are adjusted for by confounding variables in not only the group it belongs in, but also by predictors in the other group. This will allow us to more accurately interpret effect sizes of predictors.

Then, to investigate differences between the two groups of predicts, we will compare the magnitudes of coefficients of the significant predictors. Since most of the variables we are using are categorical, coefficient magnitudes will be compared directly. In cases where variables are continuous, we will take into account an average value of the variable to fully quantify the effect size for comparison. For each model, we will categorize familial and educational predictors into three axes for easier comparison.

The first axis compares the financial situation between the student's home and school. To do this, we compare the student's family's SES status composite against the type of school he/she attends, whether a majority of the students there receive free lunch, and the lowest teacher salary. Although the type of school does determine how much money is spent on the students and school facilities, it also represents the wealth of the student's family, since only wealthy families are able to pay for private schools. However, since family SES is already adjusted for, we can take this to be largely the effect of the school related effects. In addition, we are using the indicator of whether a majority of the students receive free lunch as a proxy for the financial environment the school resides in. It may include more information about whether the school is able to give free lunch to those who need it, which is a limitation that can be seen in the results.

The second axis we use is the motivational factors between home and school. We look at the highest number of years of education the student's parents have received and also how many years of education they hope for the student. On the school side, we analyze the number of years of education the student's math and English teachers want the student to complete. In addition, we take into consideration the percentage of sophomores who are in a college preparation program at school.

The third axis is the difference in correlations between the conditions at home and at school. We compare access to technology by looking at whether the student has access to a computer and Internet at home and whether or not student learning is hindered at school by a lack of technology. In addition, we investigate home conditions, such as family composition, and school conditions, such as learning hindrances.

Finally, once we have analyzed predictor effect sizes across axes within each model, we will perform a comparison of each of these axes among the models. This will allow us to more thorougly investigate whether familial or educational factors are a larger determinant, as we analyze various outcomes of student success through different groupings of factors.

Standardized Test Composite Score

Our first measure of student success is standardized test composite score, which combines the math and reading scores. Since test composite is a continuous variable, we decided to an ordinary least squares model specified below:

$$
\begin{aligned}
\text{test composite}_i =& \beta_0 + \beta_1 1(\text{race=API})_i + \beta_2 1(\text{race=Black})_i + \beta_3 1(\text{race=Hispanic})_i + \beta_4 1(\text{race=White})_i \\
&+\beta_5 1(\text{family composition=two parents})_i + \beta_6 \text{SES}_i + \beta_7(\text{parents' \# years ed.})_i \\
&+\beta_8(\text{\# years ed. parents desire for student})_i + \beta_9 1(\text{has computer/internet at home})_i \\
&+\beta_{10} 1(\text{school type=public})_1 + \beta_{11}(\text{\# years ed. math teacher desires for student})_i \\
&+\beta_{12}(\text{\# years ed. English teacher desires for student})_i + \beta_{13}(\text{\% sophomores in college prep program})_i \\
&+\beta_{14}(\text{lowest teacher salary (thousands)})_i + \beta_{15} 1(\text{percentage students with free lunch} > 50\%) \\
&+\beta_{16} 1(\text{learning hindered by lack of space})_i + \beta_{17} 1(\text{learning hindered by poor building conditions})_i \\
&+\beta_{18}(\text{learning hindered by poor heating/air/light})_i + \beta_{19}(\text{learning hindered by lack of text/supplies})_i \\
&+\beta_{20}(\text{learning hindered by poor facilities})_i + \beta_{21}(\text{learning hindered by poor technology})_i \\
&+\beta_{22} 1(\text{percentage students with free lunch} > 50\%)_i 1(\text{learning hindered by lack of space})_i + \epsilon_i, \text{ where } \epsilon_i \overset{i.i.d.}{\sim}
\end{aligned}
$$

OLS models have model assumptions of independence, normality, homoscedasticity (constant variance), and linearity between predictors and the response. Although observations are sampled from individual students, there may be problems with independence, since students may come from the same school. Therefore, friend groups, school conditions, and other similarities between experiences of students in the same school may affect independence. This is a limitation of our model, as the actual school of each student is not available in the data, so it cannot be accounted for. For future analysis, this data can be requested from the NCES to create a mixed effects model. To assess the other assumptions and multicollinearity, model diagnostics are analyzed in the Appendix.

Socioeconomic Status

Another measure of student success is socioeconomic standing after leaving school. We analyze the student's SES quantile 9 years after the base year survey when the student was in 10th grade. The data includes a measure of SES, and we performed sensitivity analysis using linear regression of this measure, but the model performed poorly. In addition, it is easier to interpret the outcome as quantiles, since the data source does not make it clear how the SES quantifications are calculated. Therefore, since the outcome is an ordinal categorical variable, we use ordinal logistic regression as follows:

$$
\begin{aligned}
&\textit{logit}(P(\text{SES quantile} \leq x)) = \beta_0 - (\beta_1 1(\text{race=API})_i + \beta_2 1(\text{race=Black})_i + \beta_3 1(\text{race=Hispanic})_i + \beta_4 1(\text{race=White})_i \\
&\quad +\beta_5 1(\text{family composition=two parents})_i + \beta_6 \text{SES}_i + \beta_7(\text{parents' \# years ed.})_i \\
&\quad +\beta_8(\text{\# years ed. parents desire for student})_i + \beta_9 1(\text{has computer/internet at home})_i \\
&\quad +\beta_{10} 1(\text{school type=public})_1 + \beta_{11}(\text{\# years ed. math teacher desires for student})_i \\
&\quad +\beta_{12}(\text{\# years ed. English teacher desires for student})_i + \beta_{13}(\text{\% sophomores in college prep program})_i \\
&\quad +\beta_{14}(\text{lowest teacher salary (thousands)})_i + \beta_{15} 1(\text{percentage students with free lunch} > 50\%) \\
&\quad +\beta_{16} 1(\text{learning hindered by lack of space})_i + \beta_{17} 1(\text{learning hindered by poor building conditions})_i \\
&\quad +\beta_{18}(\text{learning hindered by poor heating/air/light})_i + \beta_{19}(\text{learning hindered by lack of text/supplies})_i \\
&\quad +\beta_{20}(\text{learning hindered by poor facilities})_i + \beta_{21}(\text{learning hindered by poor technology})_i \\
&\quad +\beta_{22} 1(\text{percentage students with free lunch} > 50\%)_i 1(\text{learning hindered by lack of space})_i) \\
&\quad \forall x \in 1, 2, 3, 4
\end{aligned}
$$

Ordinal regression has four assumptions: response variable is ordinal, explanatory variables are continuous or categorical, no multicollinearity, and that odds are proportional. The response variable is ordinal because of the nature of quantiles. All explanatory variables used are continuous or categorical. The variables that have an ordinal nature are accounted for by making them numerical or treating them as categorical variables.

Multicollinearity is tested for using VIF, and a full likelihood ratio test is used to test for proportionality (see Appendix).

Education Attainment

Our third measure of success is the level of education the student attained. Since there are 6 levels that we coded from the data and our variables include lots of categorical variables, there is a high possibility of a small sample size for many of the groupings of levels between the predictors and outcome. Therefore, we analyze models for outcomes of whether a student dropped out of high school and whether a student graduated a 4-year college using logistic regression models, as defined below:

$$
\begin{aligned}
logit(P(\text{Dropped Out HS}_i = Yes)) &= \beta_0 + \beta_1 1(\text{race=API})_i + \beta_2 1(\text{race=Black})_i + \beta_3 1(\text{race=Hispanic})_i + \beta_4 1(\text{race=White})_i \\
&+\beta_5 1(\text{family composition=two parents})_i + \beta_6 \text{SES}_i + \beta_7(\text{parents' \# years ed.})_i \\
&+\beta_8(\text{\# years ed. parents desire for student})_i + \beta_9 1(\text{has computer/internet at home})_i \\
&+\beta_{10} 1(\text{school type=public})_1 + \beta_{11}(\text{\# years ed. math teacher desires for student})_i \\
&+\beta_{12}(\text{\# years ed. English teacher desires for student})_i + \beta_{13}(\text{\% sophomores in college prep program})_i \\
&+\beta_{14}(\text{lowest teacher salary (thousands)})_i + \beta_{15} 1(\text{percentage students with free lunch} > 50\%) \\
&+\beta_{16} 1(\text{learning hindered by lack of space})_i + \beta_{17} 1(\text{learning hindered by poor building conditions})_i \\
&+\beta_{18}(\text{learning hindered by poor heating/air/light})_i + \beta_{19}(\text{learning hindered by lack of text/supplies})_i \\
&+\beta_{20}(\text{learning hindered by poor facilities})_i + \beta_{21}(\text{learning hindered by poor technology})_i \\
&+\beta_{22} 1(\text{percentage students with free lunch} > 50\%)_i 1(\text{learning hindered by lack of space})_i
\end{aligned}
$$

$$
\begin{aligned}
logit(P(\text{Received Bachelor's}_i = Yes)) &= \beta_0 + \beta_1 1(\text{race=API})_i + \beta_2 1(\text{race=Black})_i + \beta_3 1(\text{race=Hispanic})_i + \beta_4 1(\text{race=White})_i \\
&+\beta_5 1(\text{family composition=two parents})_i + \beta_6 \text{SES}_i + \beta_7(\text{parents' \# years ed.})_i \\
&+\beta_8(\text{\# years ed. parents desire for student})_i + \beta_9 1(\text{has computer/internet at home})_i \\
&+\beta_{10} 1(\text{school type=public})_1 + \beta_{11}(\text{\# years ed. math teacher desires for student})_i \\
&+\beta_{12}(\text{\# years ed. English teacher desires for student})_i + \beta_{13}(\text{\% sophomores in college prep program})_i \\
&+\beta_{14}(\text{lowest teacher salary (thousands)})_i + \beta_{15} 1(\text{percentage students with free lunch} > 50\%) \\
&+\beta_{16} 1(\text{learning hindered by lack of space})_i + \beta_{17} 1(\text{learning hindered by poor building conditions})_i \\
&+\beta_{18}(\text{learning hindered by poor heating/air/light})_i + \beta_{19}(\text{learning hindered by lack of text/supplies})_i \\
&+\beta_{20}(\text{learning hindered by poor facilities})_i + \beta_{21}(\text{learning hindered by poor technology})_i \\
&+\beta_{22} 1(\text{percentage students with free lunch} > 50\%)_i 1(\text{learning hindered by lack of space})_i
\end{aligned}
$$

The assumptions for logistic regression are a binary nature of the dependent variable, independence, and no linearity between independent variables and the log odds. The independence assumption may be violated, as mentioned in assumptions for the OLS model for standardized test score. There is no linearity between the independent variables and the log odds, as there are over 16,000 observations. In addition, our model output shows that there is no linearity, as there are no extremely inflated coefficients.

As for other conditions to note, multicollinearity is tested for using VIF (see Appendix). Additionally, as mentioned earlier, since there are many categorical variables used in our model, there are many combinations of levels. Although many categorical variable levels were combined to reduce the total number of samples needed, there are still a few levels that have only a couple of samples. Therefore, there is still imbalanced data. However, since we are only performing inference, we will not be analyzing the less frequent combinations of variables.