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 the inclusion of more potential confounding variables. Then, to investigate differences between the two groups of predicts, we will compare the coefficients of significant predictors.

Because of the number of variables used in the models, there are many observations that have missing data for at least one of the variables. Therefore, we use the `mice` package to impute the data using the predictive mean-matching method.

Due to the small sample sizes within the levels of combinations of the categorical variables and to better take into account their ordinal nature, we transform them into numeric variables. The variables of how far the parents pushed the students to go in school and the parents' highest level of education are transformed into numeric variables that represent the number of additional years of education past 9th grade for each level indicated. We acknowledge the limitations of this, as there are assumptions made that may not be accurate for all observations. For example, for levels of education that denote only attending, and not completing, high school or college, we arbitrarily set a number of years of education that would fall under the category, as the exact number of years is unknown. Additionally, there is a set of categorical variables that denote how much learning is hindered by a particular environmental factor. We change these into numeric variables as well, where each variable ranges from integers 0 to 4 based on how much the factor affects student learning. We also ackowledge the limitations of this, since this assumes a linear change from level to level. However, we believe there may be some validity to the constant effect changes between levels and believe it to be more important to have sufficient sample sizes.

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:

$$
\begin{aligned}
\text{test composite}_i =& \beta_0 + \beta_1 1(\text{school type=private}) + \beta_2 1(\text{school type=public}) + \beta_3 1(\text{race=API}) \\
&+ \beta_4 1(\text{race=Black}) + \beta_5 1(\text{race=Hispanic}) + \beta_6 1(\text{race=White}) \\
&+ \beta_7 1(\text{family composition=two parents}) + \beta_8 \text{SES} + \beta_9(\text{parents' \# years ed.}) \\
&+ \beta_{10}(\text{\# years ed. parents desire for student}) + \beta_{11} 1(\text{has computer at home}) \\
&+ \beta_{12} 1(\text{has internet at home}) + \beta_{13}(\text{teachers' tech access score}) \\
&+ \beta_{14} 1(\text{library has computer lab}) + \beta_{15}(\% \text{ sophomores in college prep program}) \\
&+ \beta_{16}(\text{lowest teacher salary (thousands)}) + \beta_{17}(\text{learning hindrance by poor buildings}) \\
&+ \beta_{18}(\text{learning hindrance by poor heating/air/light}) + \beta_{19}(\text{learning hindrance by poor science labs}) \\
&+ \beta_{20}(\text{learning hindrance by poor fine arts facilities}) + \beta_{21}(\text{learning hindrance by lack of space}) \\
&+ \beta_{22}(\text{learning hindrance by poor library}) + \beta_{23}(\text{learning hindrance by lack of text/supplies}) \\
&+ \beta_{24}(\text{learning hindrance by too few computers}) + \beta_{25}(\text{learning hindrance by lack of tech equipment}) \\
&+ \beta_{26} 1(\text{free lunch percentage} > 50\%) \\
&+ \beta_{27} 1(\text{free lunch percentage} > 50\%)(\text{learning hindrance by lack of space}) + \epsilon_i, \text{ where } \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

OLS models have model assumptions of independence, normality, homoscedasticity (constant variance), and linearity between predictors and the response. Since each of the observations in our data comes from a single student, the independence assumption is satisfied. 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:

$$
\begin{aligned}
&logit(P(\text{SES quantile} \leq i)) = \beta_0 + \beta_1 1(\text{school type=private}) + \beta_2 1(\text{school type=public}) + \beta_3 1(\text{race=API}) \\
&\quad + \beta_4 1(\text{race=Black}) + \beta_5 1(\text{race=Hispanic}) + \beta_6 1(\text{race=White}) \\
&\quad + \beta_7 1(\text{family composition=two parents}) + \beta_8 \text{SES} + \beta_9(\text{parents' \# years ed.}) \\
&\quad + \beta_{10}(\text{\# years ed. parents desire for student}) + \beta_{11} 1(\text{has computer at home}) \\
&\quad + \beta_{12} 1(\text{has internet at home}) + \beta_{13}(\text{teachers' tech access score}) \\
&\quad + \beta_{14} 1(\text{library has computer lab}) + \beta_{15}(\text{\% sophomores in college prep program}) \\
&\quad + \beta_{16}(\text{lowest teacher salary (thousands)}) + \beta_{17}(\text{learning hindrance by poor buildings}) \\
&\quad + \beta_{18}(\text{learning hindrance by poor heating/air/light}) + \beta_{19}(\text{learning hindrance by poor science labs}) \\
&\quad + \beta_{20}(\text{learning hindrance by poor fine arts facilities}) + \beta_{21}(\text{learning hindrance by lack of space}) \\
&\quad + \beta_{22}(\text{learning hindrance by poor library}) + \beta_{23}(\text{learning hindrance by lack of text/supplies}) \\
&\quad + \beta_{24}(\text{learning hindrance by too few computers}) + \beta_{25}(\text{learning hindrance by lack of tech equipment}) \\
&\quad + \beta_{26} 1(\text{free lunch percentage} > 50\%) + \beta_{27} 1(\text{free lunch percentage} > 50\%)(\text{learning hindrance by lack of space}) \\
&\quad \forall i \in 1, 2, 3, 4
\end{aligned}
$$

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:

$$
\begin{aligned}
&logit(P(\text{Dropped Out HS}_i = Yes)) = \beta_0 + \beta_1 1(\text{school type=private}) + \beta_2 1(\text{school type=public}) + \beta_3 1(\text{race=API}) \\
&\quad + \beta_4 1(\text{race=Black}) + \beta_5 1(\text{race=Hispanic}) + \beta_6 1(\text{race=White}) \\
&\quad + \beta_7 1(\text{family composition=two parents}) + \beta_8 \text{SES} + \beta_9(\text{parents' \# years ed.}) \\
&\quad + \beta_{10}(\text{\# years ed. parents desire for student}) + \beta_{11} 1(\text{has computer at home}) \\
&\quad + \beta_{12} 1(\text{has internet at home}) + \beta_{13}(\text{teachers' tech access score}) \\
&\quad + \beta_{14} 1(\text{library has computer lab}) + \beta_{15}(\text{\% sophomores in college prep program}) \\
&\quad + \beta_{16}(\text{lowest teacher salary (thousands)}) + \beta_{17}(\text{learning hindrance by poor buildings}) \\
&\quad + \beta_{18}(\text{learning hindrance by poor heating/air/light}) + \beta_{19}(\text{learning hindrance by poor science labs}) \\
&\quad + \beta_{20}(\text{learning hindrance by poor fine arts facilities}) + \beta_{21}(\text{learning hindrance by lack of space}) \\
&\quad + \beta_{22}(\text{learning hindrance by poor library}) + \beta_{23}(\text{learning hindrance by lack of text/supplies}) \\
&\quad + \beta_{24}(\text{learning hindrance by too few computers}) + \beta_{25} 1(\text{learning hindrance by lack of tech equipment}) \\
&\quad + \beta_{26} 1(\text{free lunch percentage} > 50\%) + \beta_{27} 1(\text{free lunch percentage} > 50\%)(\text{learning hindrance by lack of space})
\end{aligned}
$$

$$
\begin{aligned}
&logit(P(\text{Received Bachelor's}_i = Yes)) = \beta_0 + \beta_1 1(\text{school type=private}) + \beta_2 1(\text{school type=public}) + \beta_3 1(\text{race=API}) \\
&\quad + \beta_4 1(\text{race=Black}) + \beta_5 1(\text{race=Hispanic}) + \beta_6 1(\text{race=White}) \\
&\quad + \beta_7 1(\text{family composition=two parents}) + \beta_8 \text{SES} + \beta_9(\text{parents' \# years ed.}) \\
&\quad + \beta_{10}(\text{\# years ed. parents desire for student}) + \beta_{11} 1(\text{has computer at home}) \\
&\quad + \beta_{12} 1(\text{has internet at home}) + \beta_{13}(\text{teachers' tech access score}) \\
&\quad + \beta_{14} 1(\text{library has computer lab}) + \beta_{15}(\text{\% sophomores in college prep program}) \\
&\quad + \beta_{16}(\text{lowest teacher salary (thousands)}) + \beta_{17}(\text{learning hindrance by poor buildings}) \\
&\quad + \beta_{18}(\text{learning hindrance by poor heating/air/light}) + \beta_{19}(\text{learning hindrance by poor science labs}) \\
&\quad + \beta_{20}(\text{learning hindrance by poor fine arts facilities}) + \beta_{21}(\text{learning hindrance by lack of space}) \\
&\quad + \beta_{22}(\text{learning hindrance by poor library}) + \beta_{23}(\text{learning hindrance by lack of text/supplies}) \\
&\quad + \beta_{24}(\text{learning hindrance by too few computers}) + \beta_{25} 1(\text{learning hindrance by lack of tech equipment}) \\
&\quad + \beta_{26} 1(\text{free lunch percentage} > 50\%) + \beta_{27} 1(\text{free lunch percentage} > 50\%)(\text{learning hindrance by lack of space})
\end{aligned}
$$

The assumptions for logistic regression are independence, no multicollinearity, and a large sample size. All observations are independent, as they are sampled from individual students. Multicollinearity is tested for using VIF (see Appendix). Finally, 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, this assumption is not fully met due to imbalanced data. However, since we are only performing inference, we will not be analyzing the less frequent combinations of variables.