Graduation Rates, a Measure of Student Achievement

David White

Senior Thesis

Abstract

From the 20th into the 21st century there has been extensive research into student achievement and the factors that bring forth such an accomplishment. Much the literature revolves around student test scores as a measure of student achievement. This paper uses graduation rates as a form a student achievement. Using a spline regression analysis, I analysis the graduation rates from 6135 school districts across the United States against demographic and financial factors. Average spending per student, average instructional expenditures per student, average pupil support service, and average instructional salary per student were all correlated with a positive relationship with graduation rates. Average spending on math, science, and teacher quality was correlated with a negative relationship with graduation rates.

Introduction

Why do we go to school? I sure hope it's to learn. We go to school to learn more so that we can get better and get smarter and then be able to do more things. We want to achieve. We live in a world where achievement, growth, moving forward, are all valued. And to be fair, they should be. Moving forward, getting better, doing well, achieving, these are all beneficial things to society. Intelligent

people have the breakthroughs that push society forward, and many of them went to regular public high schools; and many of them graduated.

Graduating high school is extremely important now, and is heavily pushed upon students upon entering middle school; not even high school. Because of this, I think it's important to understand what variables or factors or interactions could be helping or hindering students graduate. Because of the limits to my data, I could only explore financial revenues and expenditures, with the ability to control for and interact demographic variables. My data argues a negative relationship between graduation rates and average spending on math, science and teacher development programs, per student, and a positive relationship between graduation rates and average spending for support services, instruction, higher salaries, and total expenditures, per student.

Theoretical Background

There are many ways to classify student achievement. Academic achievement like test scores is one way to measure achievement (Clark & Jha, 2013; Savasci & Tomul, 2013). Enrollment rates into a school is another way, especially in places where the emphasis is not on doing well in school, but first to get people to school (Washington, 2017). This paper rejects these previous models of student achievement because for test scores, just because a student does not know how to take a test, or is a bad test taker, does not mean that they are not smart, or are not achieving; and for enrollment rates, while there still is an emphasis on being in

school, it is no longer as much about enrolling in school as it is staying in school. This is the reason I chose to look at graduation rates, because, based off of my personal observations, for many public and charter schools there is an emphasis on graduating high school and going to college.

This budding, possibly national, emphasis on graduation is why I chose to set graduation rates as my indicator for student achievement. Graduation rates do not have a barrier to entry, unlike testing where one has to know how to take the test before taking it. Graduation rates step back from an emphasis on knowledge acquirement as achievement, and takes a bigger, more portfolio, stance on achievement.

The theoretical background of my paper is grounded in the idea that the more resources available to a student, the better off they will be. Within the educational world linking resources to student achievement, the factory metaphor is utilized (Greenwald, 1996). In this world, scholars view schools as a means for production of achievement and therefore use the term, *education production function* to describe the relationship between school inputs and student outcomes.

A practical way to measure student resources is by looking at the in school resources available to the students. The first way is to look at the financial reporting of a school. Resources cost money, and the school has to pay for them. A school district receives funding from the federal, state, and local level, and then divides it up amongst the schools. There are also some federal and state fundings that go directly to schools and bypass the school district level. When all the dust has settled

and the money has changed hands, there still is the resounding task of educating the population the money was intended for. If schools aren't spending enough money on their students and providing them with the adequate resources they need to achieve, students will underperform and fall at risk to not graduate.

The teacher's ability, background education, and subsequent motivation, can all play a crucial factor in weather or not a teacher is engaging and drawing their students in. Teachers who are underpaid, under appreciated, overwhelmed because of large classrooms, and/or are inexperienced are not going to be able to perform well and then their students are going to suffer. It is for these reasons that teachers play a crucial role in the development and engagement of students to keep them in school and not dropping out.

Outside factors also offer and take away resources for students. Neighborhood factors like crime can create incentives to drop out of school, taking away resources from children (Christeson, et. al, 2008). Home factors like the socio economic status of a student's parents can also effect what resources are available at home (Lareau, 2003). The skill sets a parent builds with their children provide resources later on to guide them through life. Parents who have to work multiple jobs just to put food on the table for their children do not have time to read with their children, or have the money to enroll or take their children to extra curricular activities. Whereas parents who work jobs that afford them working forty hour or less work weeks can afford to spend more time with their children and build up different and more diverse skill sets.

Literature Review

Over the past century there have been many studies and examinations into what helps/hinders a student's achievement. Student achievement is a very controversial topic, with many scholars on all sides of the arguments. The two touted up until the early 1990s were the Coleman report and Hanushek's studies on school expenditures and academic performance. The Coleman report, published in 1966, surveyed over 600,000 students and found that on the whole academic achievement was less related to the quality of the school, and more related to the social composition of the school (the backgrounds of the other students in the schools and what they and their peers plan on doing in their lives) and family backgrounds (Coleman, 1996). The Coleman report was widely interpreted as concluding that "schools do not matter." (encyclopedia.com)

Hanushek synthesized and analyzed numerous studies on academic achievement the effects of schools. In one 147 study analysis, he found inconsistent effects of class size, teacher education, and teacher experience on achievement, coming to a shaky conclusion (he admits it) that student achievement and school expenditures are not related (Hanushek 1986). In another analysis of close to 400 studies found that there was not a clear relationship between student achievement and school resources. Within the 400 studies there were statistically significant coefficients for opposite markers of the same variable, which raised questions about

the consistency of the significance. Hanushek concluded that he believed that just throwing money at the problem would change anything (Hanushek, 1997).

According to a meta analysis of 60 education production function studies, there is a positive and very uniform relationship between student resources and school achievement. The indicators of student resources for this study were per pupil expenditure, teacher ability, teacher education, teacher experience, teacher salary, teacher pupil ration, and school size. While the study found a positive effect between student resources and school achievement, it could not make budget allocation suggestions (Greenwald, 1996).

A study from Utah found that students with any form of disability were at a greatest risk to graduate within the four year time frame. This study not only found that as a group students with disabilities had poorer outcomes then their non disabled peers, but these outcomes were varied by disability. For example, students with emotions disturbance, multiple disabilities, intellectual disability, traumatic brain injury, or autism were at the greatest risk of failing to graduate; students with autism, multiple disabilities, or an intellectual disability had lower dropout rates than those of general education and those of the students with disabilities as a group; and in contrast students with hearing impairment/deafness reported roughly on part four year graduation rates with the general education. This study highlights the heterogeneity of disability in education, and gives a sense to the specialization of resources needed to tackle special education (Barrat, et al, 2015).

The geographical size of a school district can also have an effect on graduation rates. In one of the only studies I came across that used graduation rates as their measure student achievement, they found that the geographical size of a school has an effect on graduation rates. The analysis found that decreasing the average size of a school district by two hundred square miles was correlated with an increase of graduation rates by 1.64 percentage points. This study delves into the world of bureaucracy, how a bureaucratic reach can only go so far, and the resources needed to achieve bureaucracy and student achievement (Greene & Winters, 2006).

In a research review of class size and environment and student achievement, the authors find a positive relationship between smaller class sizes and student achievement. They found that an educational program that includes students from ages kindergarten to third grade will produce more benefits than a program that reaches only on or two grades, that minority and low income students show fantastic gains when placed in smalls classes during kindergarten through third grade, that the experience and preparation of teachers is a critical factor in the success or failure of class size reduction programs, and that supports like professional development for teachers and a rigorous curriculum enhance the effect of reduced class size on academic achievement. This study yet again helps to highlight the different types of resources available in a school to a student, and to a teacher as well (Center for Public Education, 2017).

Students who drop out of high school will never graduate, unless of course they pass the GED. Answering the questions for why students drop out of high school has been tackled by many scholars (see Doll et all, 2013; Baldwin B et all, 1992, Cairns et all, 1989, Chapman et all, 2011) and there is even a website for the National Dropout Prevention Center/Network (http://dropoutprevention.org). A study in Georgia linked attendance to academic achievement, and truancy to attitudes about school. The study found that just missing five days of school begins to impact student achievement and starts to shape the student's perception of the school. They found that in early high school, attendance was a better prediction of student's likelihood to drop out than were the student's test scores (GaDOE, 2011).

It seems clear that while the meta analysis conducted in the 1980s and 1990s indicated no clear relationship between school resources and student achievement, the data and reports from the 21st century offer pushback to that idea.

Data

I chose to look on the school district level because there was ample demographic and financial statistics and it was relatively easy to obtain the graduation rates (as compared to the county level[1]). The district level is also smaller than the county level allowing for a more accurate representation of the data.

The largest obstacle to my thesis was the data collection. I obtained data from two datasets from the Common Core Data (CCD) database: finance data and

demographic data; and child poverty estimates from the Census Bureau. For the graduation rates, I obtained individual district level data sets from each state via their state websites. For Idaho, I obtained the data from a communications specialist from the Idaho Department of education through an email exchange.

From the CCD I was able to obtain financial and demographic data on the school level. In total there were 14808 observations for the finance data and 18767 observations for the demographic data. The discrepancies in the data observations were due to the demographic data set including specialized districts such as charter, magnet and private. The finance data also included these, however since I was only looking at public schools, and I was certain that both of these datasets contained public school data (as specified by the websites where the data was obtained) I decided that the discrepancies would not affect my final data set. I merged these two data sets and was left with 13078 observations.

From the Census Bureau Small Area and Income Poverty Estimates (SAIPE) I obtained data on estimates of the number of children living in poverty per district. There were 13489 observations in this data set and I matched all of them to the 13078 from the combined demographic and financial data.

The are 4 states that I chose to exclude from my analysis, Alabama, Illinois, Missouri, and Vermont. Alabama and Illinois were excluded because they provide school level graduation rate, howeve, with no accessible way to match these schools to their respective school districts. Missouri was excluded because while the Missouri ed database provided graduation rates by district, each district's graduation

rate was on a separate excel sheet within an excel file. Trying to create a workable graduation rate list from this file would have taken too much time. Finally, Vermont was excluded because the graduation rates were only available on a "click through" database that would have taken too much time to obtain.

Much of my graduation rate data was easy to obtain and was formatted on the district level. I then matched this data to the combined finance and demographic data set from the CCD. Sometimes there would be a district level code attached to my graduation rate data set that would match up to the finance/demographic dataset. When this occurred, I used a statistical program to match up the values by the individual district code. However, when the graduation rate data set did not come with a district level code, then I had to manually go through the two data sets and match up the names of the districts. I could not use a statistical program because the variable for district names under the two data sets had different abbreviates for words like "school" "community" "cooperative" and "district" and therefore would not match up.

There were nine states where I was given the school level graduation cohorts and enrollments: Alaska, Delaware, Maryland, Minnesota, New Hampshire, New York, Ohio, Oregon, and South Dakota. I added up the cohorts and enrollments for each district and calculated the graduation rates for each district. It was only for Oregon and South Dakota did I realize that I could use a statistical program to calculate the total grad cohorts and enrollments for each district; this significantly shortened the data collecting process.

When I finally tallied the number of graduation rates I had collected, factoring out the ones that did not report, I had 9310. When combined with the final financial and demographic data set, we were left with 9310 observations. Unfortunately this does not mean that all my regressions had 9310 observations. For many of my variables there were 3175 missing values, making my total number of viable, workable, regress-able observations 6135. This indicates that my data set is working with 6135 observations.

Methodology

I chose to use a spline regression analysis for my data using the statistical software analysis program Stata. After looking at scatter plots where graduations rates was the dependent variable and some financial variable was the independent variable, I realized that running an interval regression analysis was going to be the best way to fit a model to my data.

I started by generating a two-way scatter plot where graduation rates were the dependent variable and some financial variable was the independent variable. I then overlaid an "mspline" estimation on top of my scatter plot[2]. From this I was given apexes of small cubic relationships within my data. I estimated those apexes and used those x values to break up my independent variable into dummy variables which each corresponded to an interval of the apexes[3]. To do this I used the mkspline command in Stata which generated my dummy variables[4]. I did this for over 75 variables that I thought would garner a statistically significant relationship.

Through trial and error I developed a multiple regression that I thought factored in everything I was trying to account for when comparing graduation rates to a financial independent variable[5]. I first wanted to understand the district location and how that affected the regression model. I used six variables which I manipulated using the demographic variables I was given and the poverty estimates from SAIPE I created three factor variables , one for poverty, one for the local location of the school district, and one for the region of the school district

I chose to include poverty[6] as a factor variable in the regression because I wanted to see if different levels of poverty had different level of graduation rates. I use the estimated number of children in poverty from SAIPE as my poverty indicator because that was the closest indicator I had. There is no median household income data for the school district level (there is on the county level).

I chose to include school district's local location because growing up in NYC I could see how crowdedness and congestion could have an impact on student achievement. The local codes[7] tell us what kind of location the school district is located in either a city, a suburb, a town, or a rural location. The qualifications of a city, suburb, town, or rural location, were based on population, and so understanding how a high school's relative location to other people and other social forces can help us understand a student's desire to graduate or not.

I chose to include a region code to further specify my data. I wanted to know if different regions had higher or lower graduation rates, just so that I could then report that. I don't think they will necessarily have policy implications, but I do

think know what regions have on average higher graduation rates because of some sort of finance is interesting.

I also created one large interaction which 4 variables all surrounding race. I created categorial variables denoting the percentages of race in a given school district[9]. For this I did not want to observe what races had higher or lower graduation rates. Instead I set the white, black, hispanic, and asian % of the graduation class to interact with each other. This let me see certain demographics in a school district and the correlational effect they may have on graduation rates.

After running numerous regressions (I'm thinking upwards of four hundred at least[10]) I came to five independent variables that when regressed garnered higher r-squared and multiple statistically significant intervals. There were many variables that had one or two statistically significant intervals within the spline regression, however almost all the time these individual intervals they did not represent enough observations to be relevant.

Results

Here I have formatted the results of twelve multiple regressions in table format. While my regressions churned out many combinations of correlations and interactions, I have only included those that were statistically significant to 5%. Each regression is separated by the bolded name of the primary independent variable for the regression. Each primary independent variable has been divided by

the total number of students. For example, Total Federal Revenue represents the total federal revenue per student for a school district. After every regression I offer a brief analysis of the individual multiple regression, explaining the first variable from each category mentioned below. Before moving to the regressions, I suggest consulting the chart on the next page with explains the variables names.

Variable Name	Definition
Any monetary value	This is the primary independent variable. The coefficient here represents the correlated change of our dependent variable for a change in the independent variable. The monetary value represents the interval of the independent variable for which the coefficient applies.
"Suburb Large, Large "	This was part of a factor variable indicating the general location of the school district. Because this was a factor variable, the base level is City, Large. The coefficient represents the change in the dependent variable relative to the change between Suburb Large, Large and City, Large

Northeast, Midwest, South, West	This was part of a factor variable indicating the regional location of the school districts. Because this was a factor variable, the base level is Northeast. The coefficient represents the change in the dependent variable relative to the change between Northeast and either South, West, or Midwest.
WH… HI… BL… AS…	This was part of an interaction between white students, black students, hispanic students, and asian students. Each variable represents a combination of student populations. The coefficients represent the correlated change of the dependent variable due to the effect of that interaction.

Variables	Coefficient	P-Value	R-Squared
FEDERAL REVENUE - THRU STATE - MATH, SCIENCE, AND TEACHER QUALITY			0.1516
$0	1.33445	0.000	
$0-$145	-1.367544	0.000	
$145-$165	-1.618337	0.000	
$165-175	-0.7083156	0.205	
$175-$200	-2.109903	0.000	
$200-$220	-0.2697564	0.603	

$220-$250	-1.625998	0.000	
$250-$280	-1.025637	0.039	
$280-$315	-2.181214	0.000	
$315-$370	-0.3178863	0.205	
$370-$550	-2.489537	0.000	
Large Suburb	2.901966	0.014	
Midwest	3.111777	0.000	
South	3.022	0.000	
West	-2.562334	0.000	
WH (25%-50%) HI (50%-75%) BL (0%-25%) AS (0%-25%)	-6.076227	0.021	
WH (50%-75%) HI (0%-25%) BL (0%-25%) AS (25%-50%)	-14.355	0.000	
The r-squared value was .1566 indicating that my regression accounted for 15.66% of the variation in my data. For districts that spent between $0-$145, an increase of $1 was correlated with a decrease of 1.618 in graduation rates. The difference between a large suburb and a large city is correlated with an increase of 2.902. The difference between Midwest and Northeast is correlated at an increase in gradation rates by 3.112. The effect of having a district with 25%-50% white students, 50%-75% hispanic students, 75%-100% black students, and 0-25% asian students is correlated with a decrease in graduation rates by 6.076.			

Variables	Coefficient	P-Value	R-Squared
TOTAL CURRENT EXPENDITURES FOR ELEMENTARY/SECONDARY EDUCATION			0.1451
$0-18000	0.0006736	0.008	
$18000-$20500	0.0013543	0.006	
$20500-$23500	-0.000741	0.253	
$23500-$28000	0.002423	0.000	
$28000-$31500	-0.0016058	0.194	

$31500-$33750	0.0069417	0.008	
$33750-37500	-0.0016456	0.380	
Large Suburb	3.367789	0.004	
Midwest	1.589301	0.001	
West	-4.675228	0.000	
WH (25%-50%) HI (50%-75%) BL (0%-25%) AS (0%-25%)	-6.61719	0.021	
WH (50%-75%) HI (0%-25%) BL (0%-25%) AS (25%-50%)	-15.19973	0.000	
The r-squared value was .1451 indicating that my regression accounted for 14.51% of the variation in my data. For districts that spent between $0-$18000, an increase of $1 was correlated with an increase of .00067 in graduation rates. The difference between a large suburb and a large city is correlated with an increase of 3.368. The difference between Midwest and Northeast is correlated at an increase in gradation rates by 1.589. The effect of having a district with 25%-50% white students, 50%-75% hispanic students, 75%-100% black students, and 0-25% asian students is correlated with a decrease in graduation rates by 6.617.			

Variables	Coefficient	P-Value	R-Squared
CURRENT EXPENDITURES - INSTRUCTION			0.1467
$0-$2200	0.0020209	0.059	
$2200-$3800	0.0005106	0.498	
$3800-$5000	0.0037901	0.000	
$5000-$6500	0.0002452	0.592	
$6500-$8250	0.0006232	0.189	
$8250-$10500	0.0008401	0.094	

$10500-$12000	0.0022956	0.006	
$12000-$14500	-0.0006918	0.287	
$14500-$20500	0.0029557	0.000	
Large Suburb	3.43295	0.005	
Midwest	1.359271	0.011	
South	-4.821919	0.000	
WH (25%-50%) HI (50%-75%) BL (0%-25%) AS (0%-25%)	-5.952922	0.024	
WH (50%-75%) HI (0%-25%) BL (0%-25%) AS (25%-50%)	-14.74687	0.000	
The r-squared value was .1467 indicating that my regression accounted for 1467% of the variation in my data. For districts that spent between $0-$2200, an increase of $1 was correlated with an increase of .00202 in graduation rates. The difference between a large suburb and a large city is correlated with an increase of 3.433. The difference between Midwest and Northeast is correlated at an increase in gradation rates by 1.359. The effect of having a district with 25%-50% white students, 50%-75% hispanic students, 75%-100% black students, and0-25% asian, students is correlated with a decrease in graduation rates by 5.953			

Variables	Coefficient	P-Value	R-Squared
CURRENT EXPENDITURES - SUPPORT SERVICES - PUPILS			0.1432
$0-$200	0.0262502	0.000	
$200-$600	0.0116298	0.018	
$600-$750	0.0132078	0.045	
$750-$950	0.0186264	0.009	
$950-1075	0.014372	0.222	
$1075-$1200	-0.0186512	0.200	
$1200-$1400	0.0465399	0.000	

$1400-$1525	-0.0180584	0.356	
$1525-1650	0.0607122	0.014	
$1650-$1750	-0.017618	0.625	
$1750-$1900	0.0203964	0.448	
$1900-$2050	-0.0012808	0.956	
$2050-$2600	0.0187846	0.044	
Large Suburb	2.197295	0.002	
Midwest	1.493025	0.000	
South	1.467576	0.000	
West	-3.792904	0.000	
WH (25%-50%) HI (50%-75%) BL (0%-25%) AS (0%-25%)	-5.757322	0.021	
WH (50%-75%) HI (0%-25%) BL (0%-25%) AS (25%-50%)	-13.86454	0.000	
The r-squared value was .1432 indicating that my regression accounted for 14.32% of the variation in my data. For districts that spent between $0-$200, an increase of $1 was correlated with a decrease of .026 in graduation rates. The difference between a large suburb and a large city is correlated with an increase of 2.197. The difference between Midwest and Northeast is correlated at an increase in gradation rates by 1.493. The effect of having a district with 25%-50% white students, 50%-75% hispanic students, 0%-25% black students, and asian 0-25% students is correlated with a decrease in graduation rates by 5.757.			

Variables	Coefficient	P-Value	R-Squared
SALARIES - INSTRUCTION			0.1413
$0-$1100	0.0057303	0.023	
$1100-$2800	0.0053695	0.000	
$2800-$5000	0.0061974	0.000	
$5000-$6100	0.0027614	0.036	

$6100-$7625	0.0057993	0.000	
$7625-$8750	0.0020786	0.227	
$8750-$12500	0.006527	0.000	
$12500-$14500	0.006935	0.006	
Large Suburb	3.777322	0.002	
Midwest	2.894673	0.000	
South	2.353312	0.000	
West	-3.064324	0.000	
WH (50%-75%) HI (0%-25%) BL (0%-25%) AS (25%-50%)	-12.85814	0.000	
The r-squared value was . indicating that my regression accounted for 15.66% of the variation in my data. For districts that spent between $0-$200, an increase of $1 was correlated with a decrease of .0057 in graduation rates. The difference between a large suburb and a large city is correlated with an increase of 3.777. The difference between Midwest and Northeast is correlated at an increase in gradation rates by 2.895. The effect of having a district with 50%-75% white students, 0%-25% hispanic students, 0%-25% black students, and 25%-50% asian students is correlated with a decrease in graduation rates by 12.858.			

Analysis

The first thing to note are the r-squared values for my regressions. They are between .14 and .16 indicating that my regressions can explain somewhere between 14% and 16% of the variation of my data, depending on the regression. This is on the lower end of what I was expecting/hoping t find; I would have preferred to see r-squared values between .20 and .40. Therefore any conclusion that are drawn from the data must be understood as not representing the US. school system. They

represent a percentage of my data, which as I explained before, is not a full US school district data set.

Factor and Interaction Variables

I included three factor variables and four interaction variables in my regressions. The factor variables where child poverty (CPOV), a local code (ULOCAL), and a region code (REGION). Notice there is no variable for child poverty in any of the data output tables above. This was because for all my m regressions, the CPOV had no statical significance. I chose to keep it in my regressions however because accounting for child poverty increase my r-squared value, when compared to not including it.

The other two factor variables, ULOCAL and REGION, both had statistically significant relationships in all my regressions. One factor of ULOCAL had a significant relationship in all my regressions, and this was between Large Suburbs vs Large Cities. Because the coefficients were positive as well, this indicates a positive correlational effect on graduation rates for school districts located in a Large Suburb compared to school districts located in a Large City. All factors of REGION were significant in five out of my six regressions, and for the regression of Current Expenditures-Instruction, only the factor of West vs the Northeast was not statistically significant. All the coefficients were positive for the differences between the Midwest and the Northeast and the South and the Northeast, indicating a positive correlational effect on graduation rates for school districts

located in the Midwest and the South vs the Northeast. All the coefficients were negative for the difference between the West and the Northeast indicating a negative correlational effect on graduation rates for schools living in the West compared to schools living in the Northeast.

The differences observed within the factor variables point towards geographic location as an indicator for higher or lower graduation rates. One explanation for the regional differences could be that the Midwest and the South have lower levels and the West have higher levels of children in poverty compared to the Northeast. However, my data says the complete opposite[11]. The West has the lead number of estimated children in Poverty, and the Midwest and the South both have more estimated children in poverty than the Northeast. The same is true for the Local differences; the Large Suburb indicator had the largest estimated number of children in poverty. Therefore estimated number of children in poverty does not have an effect on these indicators (as also seen in the fact that non of the poverty factor variable indicators were not statistically significant).

Probably the best answer is the natural variation of my data. The total student population for the regions was, from highest to lowest, South, Midwest, Northeast, and West. It could make sense that if Midwest and South have higher populations, the they could also have more districts with higher graduation rates. The same could apply for the local difference of Large Suburb to Large City; large Suburban districts have the highest student population in my data.

Interaction Variables

The interaction that I included in my regressions deal with the different % of racial populations of the graduation class. The interaction is based on comparing five levels of each race variable against each other. The five levels where (0%, 0-25%, 25-50%, 50-75%, 75-100%. There were two intervals of the interaction that were statistically significant. One, occurring in all my regressions, shows a negative correlation effect on graduation rates when a school district has 50-75% white students, 0-25% hispanic students, 0-25% black students, and 25-50% asian students. The other one, occurring in all but one of the regressions, shows a negative correlational effect on graduation rates when a schools district has 25-50% white students, 50-75% hispanic students, 0-25% black students, and 0-25% asian students.

When we look at these populations of races and look at the estimated children of poverty, we can see patterns emerging that can explain the interaction. If we take a look at poverty counts by race[12] we can find the counts of estimated children in poverty of the first and second intervals of the interaction. We can see that in the first one interval (50-75% white students, 0-25% hispanic students, 0-25% black students, and 25-50% asian students), the counts for hispanic, black and asian are very high and only medium for white. When we look at the second interval (25-50% white students, 50-75% hispanic students, 0-25% black students, and 0-25% asian students) the counts for hispanic and white have gone down drastically, black has stayed the same, and asian has increased only slightly. I believe this shift of

poverty estimates is why we can see a larger negative coefficient for all of the first intervals of the interaction compared to the second interval.

Dependent variables

I started out my project with over seventy five variables, however by the end I chose five to keep in my final regressions. These variables represent what forces I found to have a significant effect and I higher r-squared.

The first dependent variable is Federal Revenue-Thru State-Math, Science, and Teacher Quality. This represents how much average funding per student a school district receives for math and science programs, and for the training of effective teachers. The breakdown of this variable shows all negative coefficients for every interval. Only three out of the ten intervals are not statistically significant, but still are have negative effects. This indicates that spending more per student on math, science, and teacher development programs is correlated with a decrease in graduation rates.

This correlation is one that is difficult to explain. One possible explanation is the these schools, while spending more, have inefficient math, science, and teacher development programs. However, without being on the ground, inside the schools, we cannot say that for certain. This variable also represents the money the districts were allocated, not how much they spent. If they are not actually spending more on math, science, and teacher development, but still receiving more funding, then that could explain the negative relationship. However, without being in the

schools in person, it is impossible to determine the effectiveness of spending on math, science, and teacher quality.

The second dependent variable is Total Current Expenditures for Elementary/Secondary Education. This represents the total average amount of money a school district spent in 2014. The breakdown for this variable shows both positive and negative coefficients. The negative coefficients are not statistically significant, and they are much smaller than the statistically significant positive coefficients right next to them. Because the negative coefficients are representative of far fewer observations than the positive coefficients, and because their coefficients are much smaller than those of the statistically significant positive coefficients, this indicates an overall positive effect of spending more on average per student and graduation rates.

The third dependent variable is Current Expenditures-Instruction. This represents the average amount of money a school district spent on instruction per student. The breakdown for this variable shows all but one positive coefficient, with three of the positive coefficients being statically significant. This indicates an overall correlational positive effect of spending more on instruction per student and graduation rates.

The fourth dependent variable is Current Expenditures-Support Services-Pupil. This represents the average amount a school district spent on the support services for students, per student. The breakdown of the variable shows the first four coefficients as statically positive effects. We can see that these coefficients

belong to intervals of the spline regression that basically fall under a vast majority of the data[13]. Because the rest of the intervals cover so the little of the data, this indicates an overall positive correlated effect on graduation rates. This means that an increase in average spending on support services for students can be correlated with an increase in gradation rates.

The fifth dependent variable is Current Expenditures-Salaries-Instruction. This represents the average amount a school spends on salaries for instruction, per student. The breakdown of the variable shows all positive coefficients, and all but two statistically significant, indicating an overall positive effect. This means that an increase in spending on teachers' salaries per student is correlated with a positive increase in graduation rates.

The easiest way to explain the second through fifth dependent is to think about resources. In my experience as an educator, having more resources makes teaching easier. A study which synthesized many other studies on student achievement found that a broad range of resources were positively related to student outcomes and that moderate increases in spending may be associated with significant increases in achievement (Greenwald, 1996). For the schools in my data, their increase in spending correlated with increases in graduation rates. It makes sense that spending more on instruction and student support would correlate with graduation rates because increases in these areas usually leads to increases in student resources. An increase in teacher salaries would could also make teachers happier

to work, leading to more engagement with their students, and eventually leading to higher achievement and higher graduation rates.

Conclusion

I will admit, my data does not show much. My r-squared values were much lower than what I had expected them to be. This means that any statistically significant effect that I found does not account for a whole lot of the variation in my data. I was able to get my r-squared above 10%, so my data can explain at least 10% of the variation of data. I think for a national sample that is not terrible. I was able to show that there was a negative correlational relationship between spending money on math, science, and teacher development programs and graduation rates; and a positive correlational relationships between average spending per student and graduation rates, average instructional expenditures per student and graduation rates, average pupil support services expenditures per student and graduations rates, and average expenditures on salaries per student and graduation rates. It feels a little weird that only five out of the over seventy five independent variables that I had were as statistically significant as my final five.

My knowledge of data analysis is very limited and therefore I did not know a lot about complex statical analysis. Further research would require more complex statistical analysis to tease out more of the nuanced effects of my data (if there were any). Take for example a study conducted by Spyros Konstantopoulos in which he examined the school effects on student achievement. He realized that a two level

hierarchical linear model would be better suited than a multiple regression to study the clustering nature of schools. This is something that I would have never thought about, but possibly had I been able to learn more about statistical analysis. Konstantopoulos found that a substantial proportion of the variation in student achievement was within the schools themselves, not in-between schools (Konstantopoulos, 2006).

Works Cited

1. Baldwin, Beatrice, and Moffit. "The High School Dropout: Antecedents and Alternatives." *Journal Of School Leadership* 3rd ser. 2 (1992): 355-62. *ERIC*. Web. 28 Apr. 2017. <https://eric.ed.gov/?id=EJ447133>.

2. Barrat, Vanessa X., Beth Ann Berliner, and Adam Voit. "School Mobility, Dropout, and Graduation Rates across Student Disability Categories in Utah." *Proquest*. Proquest, 2014. Web. 28 Apr. 2017. <http://search.proquest.com/eric/docview/1651828973/5E92CAA29DAB4E84PQ/173?accountid=11321>.

3. Chapman, Chris, Jennifer Laird, and Nicole Ifill. *Trends in High School Dropout and Completion Rates in the United States: 1972–2009*. Rep. NCES, Oct. 2011. Web. 28 Apr. 2017. <https://nces.ed.gov/pubs2012/2012006.pdf>.

4. Christen, Bill, Brian Lee, and Stephanie Schaefer. School or the Streets Crime and America's Dropout Crisis. Rep. Washington DC: JL Blackmon II, 2008. Print.

5. Cobb-Clark, Deborah A., and Nikhil Jha. *Educational Achievement and the Allocation of School Resources*. *Conference IZA*. Melbourne Institute of Applied Economic and Social Research, 31 July 2013. Web. 28 Apr. 2017. <http://conference.iza.org/conference_files/ESSLE2013/cobb-clark_d77.pdf>.

6. Coleman, James S. *Equality of Educational Opportunity*. Rep. Washington DC: US Government Printing Office, 1966. Print.

7. Doll, Jonathan Jacob, Zohreh Eslami, and Lynne Walters. "Understanding Why Students Drop Out of High School, According to Their Own Reports." *SAGE Open* 3.4 (2013): 215824401350383. Web. 28 Apr. 2017. <http://journals.sagepub.com/doi/pdf/10.1177/2158244013503834>.

8. Greene, Jay P., and Marcus A. Winters. "The Effect of Residential School Choice on Public High School Graduation Rates." *Peabody Journal of Education* 81.1 (2006): 203-16. *Proquest*. Web. 28 Apr. 2017. <http://search.proquest.com/eric/docview/62100313/5E92CAA29DAB4E84PQ/81?accountid=11321>.

9. Greenwald, Rob, Larry V. Hedges, and Richard D. Laine. "The Effect of School Resources on Student Achievement." *Review of Educational Research* 66.3 (1996): 361-93. *JSTOR*. Web. 28 Apr. 2017. <https://www.jstor.org/stable/pdf/1170528.pdf>.

10. Hanushek, Eric A. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19.2 (1997): 141-64. *Stanford.edu*. Stanford. Web. 28 Apr. 2017. <http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%201997%20EduEvaPolAna%2019(2).pdf>.

11. Hanushek, Eric A. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24.3 (1986): 1141-177. *JSTOR*. Web. 28 Apr. 2017. <http://www.jstor.org/stable/10.2307/2725865?ref=search-gateway:548d8c0df1de0adc80e051fc09fe2aee>.

12. Konstantopoulos, Spyros. "Trends of School Effects on Student Achievement: Evidence from NLS:72, HSB:82, and NELS:92." *Teachers College Record* 108.12 (2006): 2550-581. *Proqueset*. Web. 28 Apr. 2017.

13. Savasci, Havva Sebile, and Ekber Tomul. "The Relationship between Educational Resources of School and Academic Achievement." *International Education Studies* 6.4 (2013): n. pag. Web. 28 Apr. 2017. <http://files.eric.ed.gov/fulltext/EJ1067588.pdf>.

14. US Fed News Service. US State News. *GEORGIA DEPARTMENT OF EDUCATION RESEARCH SHOWS STUDENT ATTENDANCE SIGNIFICANTLY IMPACTS STUDENT ACHIEVEMENT*. *Proquest*. Proquest, 08 Sept. 2011. Web. 28 Apr. 2017. <http://search.proquest.com/pqrl/docview/887905003/1047BDD59164416EPQ/3?accountid=11321>.

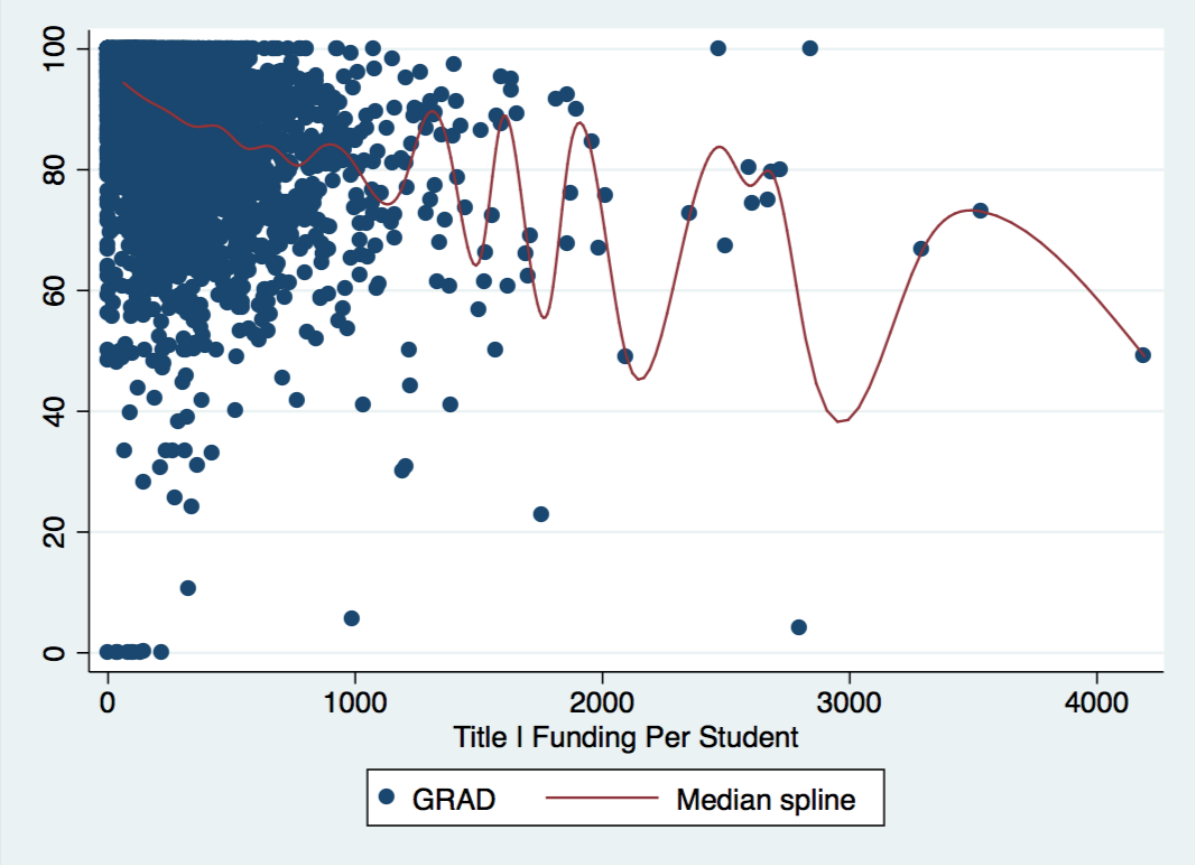
Appendixes

1. I first chose to look graduation rates on the county levels. After reading the book *City and County Extra* by Rowman and Littlefield, I thought that the county level

would generate the most data. There was ample resources for county level data from the census and from other resources across the internet. However what I had not accounted for was that a lot of the reporting of graduation rates was done on the school district level, and not on the county level. This required me to tally up all the graduates of school districts of one county, and divide them by the total enrollment for the county. Then I would have to match that state's data set to my master data set. This was before I knew the magical powers of Stata and the merge function, so I was doing this all by hand. It was a horrible, time consuming process. When winter break come around I was nowhere near done with my data collecting. I knew something had to change, and so that's when I changed my data level. Over winter break I started collecting graduation rates on the school district level. It was way easier and a lot less time consuming. After meeting with my advisor in the first weeks of February, I learned about the Stata "merge" function and it changed my life. There were still some states where the function would not work, but it shaved off tens of hours of data entry.

2. Stata entry: *twoway (scatter GRAD TitleI) (mspline GRAD TitleI*
   Here we have a scatter plot of graduation rates as a product of Title I Funding Per Student. Overplayed is the median spline estimation line. By telling Stata *(mspline GRAD TitleI)* I have asked Stata to split up my scatter plot into sections,

find the median of those sections, and then connect those points using a cubic



estimation model.

3. For some reason Stata does NOT have a function where you can just like on a point on the scatter plot and Stata will tell you the x and y values. This meant that I had to estimate those apexes in the graph from appendix 2. Thankfully Stata has an option to add vertical reference lines at specified x values. I used guess and check to eventually estimate the apexes.

4. The process for creating an mkspline dummy variable is quite simple. Take Title Funding Per Student. The name of the variable is Title. You create a Title0

variable by telling Stata: *gen Title0=Title*. Then you type the mkspline command:

mkspline Title1 # Title2 # Title3 # Title4 # Title5 # Title6 = Title0

The # represents whatever numeric value the interval falls on. From this, when running the regression, just add Title0-Title6 (or whatever number of intervals you choose) to the regression and you're good to go. The coefficient represents the correlational change of y with a change of 1 of x for that interval.

5. Figuring out which categorical variables to include in my data set through me a curveball. I started using a variable named "TYPE" thinking that it meant school district location type. I was wrong. I did a bunch of regressions using TYPE instead of ULOCAL for my district location type variable. I also tried factoring a categorical variable denoting levels of revenue per student from property taxes.. that did not work either. I eventually settled on ULOCAL, CPOV, and REGION.

6. My poverty categorical variable (CPOV) is derived from estimates of number of children in poverty for a given district. I categorized CPOV as:

Label	0	1	2	3	4	5	6	7
Value	0	0-50	50-100	100-200	200-300	300-500	500-1000	1000+

7. My local codes variable (ULOCAL) is the location type for the kind of school district. There are twelve location types: City (large, midsize, small), Suburb

(large, midsize, small), Town (fringe, distant, remote), and Rural (fringe, distant, remote).

1. City, Large: Territory inside an urbanized area and inside a principal city with population of 250,000 or more.
2. City, Midsize: Territory inside an urbanized area and inside a principal city with population less than 250,000 and greater than or equal to 100,000.
3. City, Small: Territory inside an urbanized area and inside a principal city with population less than 100,000.
4. Suburb, Large: Territory outside a principal city and inside an urbanized area with population of 250,000 or more.
5. Suburb, Midsize: Territory outside a principal city and inside an urbanized area with population less than 250,000 and greater than or equal to 100,000.
6. Suburb, Small: Territory outside a principal city and inside an urbanized area with population less than 100,000.
7. Town, Fringe: Territory inside an urban cluster that is less than or equal to 10 miles from an urbanized area.
8. Town, Distant: Territory inside an urban cluster that is more than 10 miles and less than or equal to 35 miles from an urbanized area.
9. Town, Remote: Territory inside an urban cluster that is more than 35 miles of an urbanized area.
10. Rural, Fringe: Census-defined rural territory that is less than or equal to 5 miles from an urbanized area, as well as rural territory that is less than or equal to 2.5 miles from an urban cluster.
11. Rural, Distant: Census-defined rural territory that is more than 5 miles but less than or equal to 25 miles from an urbanized area, as well as rural territory that is more than 2.5 miles but less than or equal to 10 miles from an urban cluster.
12. Rural, Remote: Census-defined rural territory that is more than 25 miles from an urbanized area and is also more than 10 miles from an urban cluster.

8. My region code is tabulated by:

Label	1	2	3	4
Value	Northeast	Midwest	South	West

State	Connecticut Maine Massachusetts New Hampshire Rhode Island Vermont Delaware New Jersey New York Pennsylvania	Illinois Indiana Michigan Ohio Wisconsin Iowa Kansas Minnesota Missouri Nebraska North Dakota South Dakota	Florida Georgia Maryland North Carolina Virginia District of Columbia West Virginia Alabama Kentucky Mississippi Tennessee Arkansas Lousiana Oklahoma Texas	Arizona Colorado Idaho Montana Nevada New Mexico Utah Wyoming Alaska California Hawaii Oregon Washington

9.

10. I started by taking the variable for white, black, hispanic, or asian count in twelfth grade and dividing it by the total count of twelve grade. I then created a categorical variable off the percentages with these values:

Label	0	1	2	3	4
Value	0	0-25%	25-50%	50-75%	75-100%

11.

12. So this is where a large part of my analysis time went… into trying to figure out how to analyze my data. I was new to the game, so I first tried a regular linear regression, and I added a bunch of categorical or interaction variables. I wasn't getting high enough r-squared values, and I thought that my data looked logarithmic, so I started playing around with natural logarithmic regression. The idea behind this was that there was some way to transform my variable by, taking the natural log of it, to make the regression smoother. The statistical analysis software can tell you the right way to transform the variable, however it's really

hard to do that in Stata, so I tried learning SPSS, XLSTAT, and Wolfram Mathematic to try and solve the issue. Eventually I resorted to trying and finding the right combination of stretching and shrinking the graph of y=ln(x) by guessing and checking. Needless to say that did not work either. Finally I just took simply the natural log of the variable and regressed that. I was getting higher r-squared values and statistically significant coefficients, however I didn't really understand what I was doing. Finally I remembered my thesis advisor telling me about spline regression analysis. I looked into it, thought it might work, and well the rest is this thesis.

13. This is the tabulation of CPOV by REGION. We would think that higher coefficients for 2 and 3 (Midwest and South) would indicate lower levels of poverty, but this tabulation says otherwise. There are higher numbers of school districts in levels 2 and 3.

	REGION				
CPOV	1	2	3	4	Total
0	2	0	0	0	2
1	78	500	276	304	1,158
2	234	496	345	178	1,253
3	475	567	512	189	1,743
4	385	331	308	110	1,134
5	355	363	412	143	1,273
6	245	292	548	189	1,274
7	177	225	723	348	1,473
Total	1,951	2,774	3,124	1,461	9,310

CPOV	White					
	0	1	2	3	4	Total
0	**0**	**0**	**0**	**1**	**1**	**2**
1	**6**	**66**	**55**	**125**	**545**	**797**
2	**16**	**65**	**84**	**163**	**553**	**881**
3	**29**	**95**	**112**	**193**	**772**	**1,201**
4	**14**	**62**	**69**	**135**	**492**	**772**
5	**11**	**55**	**71**	**133**	**585**	**855**
6	**17**	**86**	**74**	**148**	**542**	**867**
7	**22**	**75**	**113**	**191**	**603**	**1,004**
Total	**115**	**504**	**578**	**1,089**	**4,093**	**6,379**

14. Tabulations of CPOV by White; Hispanic; Black; Asian

CPOV	Hispanic					
	0	1	2	3	4	Total
0	**1**	**1**	**0**	**0**	**0**	**2**
1	**199**	**507**	**44**	**25**	**22**	**797**
2	**227**	**534**	**69**	**31**	**20**	**881**
3	**285**	**750**	**102**	**41**	**23**	**1,201**
4	**168**	**506**	**53**	**29**	**16**	**772**
5	**215**	**551**	**55**	**20**	**14**	**855**
6	**212**	**521**	**74**	**29**	**31**	**867**
7	**272**	**584**	**89**	**37**	**22**	**1,004**
Total	**1,579**	**3,954**	**486**	**212**	**148**	**6,379**

	Black					
CPOV	0	1	2	3	4	Total
0	0	2	0	0	0	2
1	321	406	37	19	14	797
2	340	453	48	17	23	881
3	467	603	64	35	32	1,201
4	288	398	47	17	22	772
5	312	460	41	19	23	855
6	316	462	42	25	22	867
7	408	473	61	28	34	1,004

	Asian					
CPOV	0	1	2	3	4	Total
0	1	1	0	0	0	2
1	397	396	4	0	0	797
2	462	403	14	2	0	881
3	604	590	6	1	0	1,201
4	392	371	8	1	0	772
5	434	412	7	1	1	855
6	451	406	9	1	0	867
7	538	458	6	2	0	1,004
Total	3,279	3,037	54	8	1	6,379

15. We can see how many observations are represented by the first four intervals by creating a copy of our starting variable, Current Expenditures- Support Services-Pupils, per student. Lets call this VAR1 Then we know that the fourth interval goes up to 950, so we set VAR1 = 1 when VAR1<=950. There were 5552 changes that were made, indicating that the first four intervals cover just over 90% of the observations.