Investigating Factors that Influence High School Math Scores
Adrianna Lucero   Takehiro Matsuzawa
Stat 139

Abstract
Math scores in the U.S. both lag behind international competitors and exhibit achievement gaps based on race and income, making it imperative that researchers discover the roots of the problem and the most effective methods to fix it. This study uses multiple linear regression to examine some of the variables that might be associated with associated with math achievement. It uses a large national data set, the High School Longitudinal Study of 2009, to explore which of 20 factors are most highly associated with students' test scores. Race and income remain significant predictors of test scores even when holding other variables constant. Additionally, individual-level factors such as interest in math, math self-efficacy, how much time one spends on math homework, and whether one takes calculus are associated with higher math scores. Also, though the years that a teacher spends teaching math is associated with higher test scores, earning a higher degree is not. Finally, the percent of students at a school who go on to higher education is associated with higher test scores, indicating there may be school-wide effects on students. We suggest that future research should delve further into these relationships and test them using experimental methods.

I. Introduction
Individuals concerned with both equity and international competitiveness have reason to be concerned about math scores in the US. Compared to other developed countries, the US ranks 35th out of 64 countries in their Program for International Student Assessment (PISA) math scores (Pew Research Center 2015). Domestically, gaps in math achievement on the National Assessment of Education Progress (NAEP) between high-income and low-income students, racial groups, and slightly between boys and girls (US D.O.E. 2012), persist.

Researchers are not certain about the precise causes for these problems. For instance, different scholars provide different hypotheses for why the U.S. lags behind other countries, with some attributing the problem to school practices, and others believing the biggest contributing factor is high poverty rate (Fensterwald 2013). Similarly, the gaps between students of different racial or class backgrounds might be due to the quality of schools in low-income, racially-segregated neighborhoods, or to the involvement of parents, or to the effects of poverty upon students.

In this study, we examine what factors are correlated with high math scores in order to hypothesize about what the causes of some of these problems might be. We examine:

1) What are the effects of individual-related factors, such as interest in math?
2) What are the effects of family-related factors, such as parents' level of education?

3) What are the effects of teacher-related factors, such as the years of math taught?
4) What are the effects of school quality-related factors, such as the number of students the school sends to higher education?

To examine these questions, we use a unique data set - the High School Longitudinal Study (HSLS) of 2009 - that follows a randomly selected sample of nearly 20,000 9th graders from the 2009-2010 school year in 900 public and private high schools. Two waves of data collection, in 9th and 11th grade, have been completed so far. The study is particularly designed to track the evolution of students' interest and outcomes in science, technology, engineering, and mathematics (STEM). Using this large, multi-year data set to examine the variables associated with high individual math scores might help future researchers pinpoint how exactly policymakers should intervene to improve academic equity and excellence in the United States.

II. Methods

To examine our hypothesis, we use multiple linear regression to fit a model relating 20 different individual, family, teacher, and school-related factors to students' math scores. Within schools, students' scores are not independent and therefore likely to be highly correlated. Due to the limitations of our data set - information about the exact school and state of each student were removed from publically available data - we were not able to use hierarchical models to examine how effects might be clustered. Instead, we used multiple linear regression to create a model.

The outcome variable we examine is the math score from students' junior year of high school, calculated out of 118 possible points. This test measures multiple domains of algebraic reasoning, such as linear equations, systems of equations, sequences, and so on (HSLS 2009 Data File Documentation).

	Range	Mean	Standard Deviation
mathScore	25.9-115.1	64.4	19.00

To select predictor variables, we drew from school-provided data on students and families, and student, teacher, counselor, and administrator questionnaires. We each selected variables we thought would be likely to influence students' math score, and then discussed our selection and finalized a list of 21 variables. We summarize them roughly by category below, noting that some variables are difficult to place into one single category. For instance, taking AP Calculus may be a result of student interest, parent pressure, as well as whether the school offers the class.

Individual	Sex, Race, Students' interest in their math class, Students' feelings of self-efficacy in math, How many friends get good grades, Hours spent on math homework, Whether student took AP calculus
Family	English spoken at home, Parents' highest level of education, Family

	income, Number of siblings, Whether parents divorced, How often student and parents discuss college applications
Teacher	Years taught high school math, Highest degree earned
School	Percent of students who entered higher education, Caseload of each school counselor, Whether school offers scholarships, Average schoolwide math ACT score, Average schoolwide math SAT score

We cleaned up the data, creating binary and factor variables where appropriate. A summary of the variables is provided in Appendix A . We examined histograms and qqplots of continuous variables, and corrected right-skewed variables famIncome (family income) and mathYears (years of math taught by teacher) by using a log(X) transformation. We corrected left-skewed variable studentsHE (percent of graduates attending higher education) with a X^2 transformation. Graphs are included in Appendix B.

After fitting a first model including all variables, we found that the model had only 1158 degrees of freedom - very few for a data set including 23,000 observations. We determined this was because the lm() command does not use observations with missing data, and certain variables had very high rates of missing data, particularly variables from the parent and counselor questionnaires. Thus, we decided to remove all variables with a nonresponse rate over 30%, both to improve the degrees of freedom of our model and because this indicates there might be nonrandom factors influencing nonresponse to that particular question (Appendix C). We removed number of siblings, divorce, times discussed college with parents, avg. ACT Math, and avg. SAT Math. This change ensured that our initial model had 9708 degrees of freedom.

The full model:
fullmodel <- lm(score ~ race + sex+homeLang+logFamIncome+mathInt +
        mathEff+fGrade+hoursHW+calculus+degreeTE+logmathYears+scholarship
      +sqrtstudentsHE+caseLoad, data=newdata)

We conducted a step backward procedure to add main effects to our model and remove nonessential variables, and then utilized a step forward variable selection procedure to incorporate interaction effects. We compared these two models using the Extra-Sum of Squares test and concluded that the second model had the best explanatory power. Then, we checked the assumptions of linear regression by checking whether the mean of the residuals was 0, whether variance of the residuals was constant, and whether the residuals were normally distributed using a plot of the residuals vs. fitted values and a qqplot. These assumptions were satisfied (see Appendix E), although observations are unlikely to be independent.

III. Results

All results are reported while holding the other variables in the model constant. The full results are reported in Appendix D. As this is an observational study, all relationships between predictor variables and the outcome are associations, and cannot be interpreted as causal effects.

1. The Persistent Effects of Race. Even considering the many other variables in our model, the racial identity of students was still a significant predictor of their test score outcomes. Asian students scored 5.3 more points than White students on the test (p-value 0.004), Black students scored 7.1 points lower (p-value .000), and Hispanic students scored scored -4.7 points lower than White students (p-value .001). All the other categories were not associated with significant effects on students' scores.

In general, this indicates that there still remains something unique about the experiences of different racial groups in America that shapes students test score outcomes. The association between race and test scores ranges from a quarter of a standard deviation for Hispanic students to .38 of a standard deviation for Black students. Our analysis does not provide any further evidence to indicate why.

2. Individual-related Factors. There are many characteristics of individuals that are significantly associated with test scores. Interest in math is positively associated with test scores (p-value 0.018), as is taking AP Calculus BC (p-value 0.000) - a 17 point higher score. It makes sense that students who take higher levels of math and are interested in math would have better scores. Additionally, students who reported higher feelings of self-efficacy in math had higher scores (p-value 0.018). There was a positive interaction between interest and math and math self-efficacy. However, the relationship between having higher self-efficacy and math scores was smaller in magnitude when considering females instead of males (interaction term p-value (0.000), although the main effect for females was not significant. This is difficult to interpret; though it is understandable that students who feel more confident about their math abilities do better on math tests, it is unclear why confidence would have a smaller association with test scores for women than for men.

The more time an individual spends working on their math homework per week, the higher their score is predicted by this model to be; for instance, individuals who spent less than half an hour on their homework scored an average of nearly 5 more points on their exam compared to students who spent no time on their homework. From there, each unit more of time has an increasing effect on scores, with the maximum association of 11.4 points at 7 to 9 hours of math homework per week. Finally, though having less than half of one's friends have high grades is positively associated with 3.2 higher points on the math test relative to individuals with few or no friends with high grades, there seems to be diminishing effects of friends on grades from there.

3. Family-related Factors. Income is the primary characteristic of families that was significant in the final model. The log of family income was positively associated with an increase in the

median math score of students (p-value 0.000). Each one unit increase in the X scale is associated with an increase of $15,000-$20,00 of income given the way the variable was coded, making it difficult to interpret an exact association between income and math scores. Parents' level of education was not selected for the final model. Whether English is or is not the first language spoken at home is also not significant, although the interaction term between a students' first language not being English and the student being Native American is positive and significant (.011), indicating that Native American students being exposed to their native language at home is associated with higher test scores among these students.

4. Teacher-related Factors. The years a math teacher has spent teaching math is a statistically significant predictor of the math score of a student (p-value 0.000). However, the math score changes by only 1.882 points. The highest degree earned by the teacher was not included in the final model. Thus, it seems that experience is associated with students' math scores, though additional higher education degrees are not.

5. School-related Factors. The percentage of students who entered the higher education is statistically is positively correlated with the math score of a student (p-value 1.83e-06). Thus, this indicates that there are may be school related predictors of student achievement, not merely individual or family predictors. Average caseload for school's counselors, which we thought would measure the amount of personal attention and advice students received, was not included in the final model. Whether the school offered scholarships to students also did not have a significant relationship with students' math score.

IV. Limitations

This study examines correlations between observational data, and thus cannot make causal claims between the relationship between variables. There are likely many confounding factors and variables we did not consider that are related to students' math scores. Instead, what we hope is that examination of this data set will allow future researchers to generate hypotheses that can then be tested through experimental research. For instance, as self-efficacy is associated with higher test scores, researchers might test experimentally whether different interventions to boost students' self-confidence can improve their math performance. Or, as hours spent on homework are associated with higher scores, researchers might design an experiment where students are incentivized for how long they spend on their homework.

One other concern with this study was dependence between observations. Students' scores in each school are likely highly correlated. Since we do not have cluster data about which school students belong to, as that information is not released to public, we could not use hierarchical models for this study. Instead, we used a multiple linear regression model, which assumes independence between observations. We anticipate that the researchers involved in the HSLS project will consider the importance of independence when analyzing the complete data set.

Moreover, this data set has many missing observations, as demonstrated in Appendix C. Thus, when fitting a multiple linear model, the model lost a lot of degrees of freedom. Had we tried to fit all the predictors at first, R deleted 22219 of 23,415 observations due to missingness. As a result, we needed to delete some variables which had many missing observations, but which may have had important explanatory power in our model if they could have been incorporated.

Another weakness might result from the outcome variable, a math score. This math score comes from the cohort of students' junior year of high school, calculated out of 118 possible points, but it is hard to predict students' math abilities only with one variable. Also, this test is taken during junior year, so it does not show the achievement or improvement of the student.

V. Conclusions

This data analysis indicates that race and family income remain important factors in predicting students' math score, even when controlling for other individual or school-related factors. Moreover, the years of teaching experience of math teachers and the percent of students from the school who attend college are significant predictors to predict math scores. Students are not able to change these variables. Thus, test scores are not solely dependent on students' abilities. However, there are also many individual variables that are associated with test scores, such as confidence in math and who spends more hours on homework, which are hopefully under students' control.

Futures studies might utilize experimental methods to examine the causal relationships between the many influential variables we uncovered and students' test scores. Researchers will need to look deeper into exactly what factors that are associated with race and income lead to achievement gaps. They also might examine variables we did not have access to, such as the racial and income makeup and geological location of the school. There remain enduring puzzles over how to improve math scores in this country, which future research will hopefully uncover.

Works Cited

Desiver, Drew. "U.S. students improving – slowly – in math and science, but still lagging
     internationally". *Pew Research Center,* February 2, 2015. Web. Accessed May 5, 2015.

Fensterwald, John. "U.S. scores stagnant, other nations pass us by in latest international test".
EdSource, December 3, 2013. Web. Accessed May 5, 2015.

National Center for Education Statistics. *The Nation's Report Card: Trends in Academic
     Progress 2012.* Institute of Education Sciences, U.S. Department of Education,
     Washington, D.C., 2013. Web. Accessed May 5, 2015.

National Center for Education Statistics. "High School Longitudinal Study of 2009
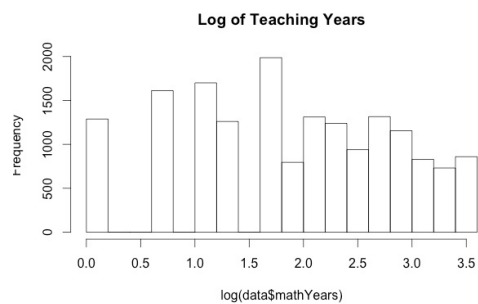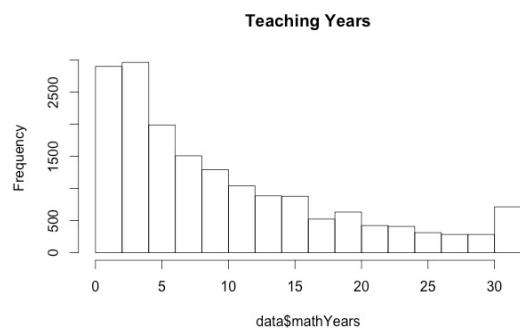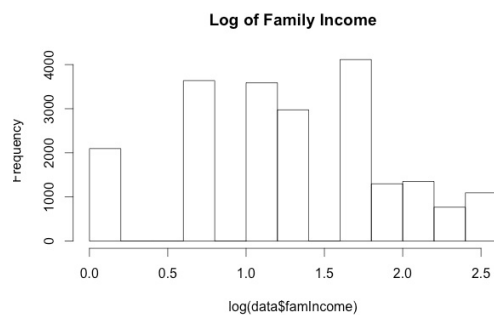(HSLS:2009/12) Student File". U.S. Department of Education, December 2009
     http://nces.ed.gov/ Accessed May 5, 2015.

Appendix A: Summary of Variables

Sex	Individual	sex	"Male", "Female"
Race	Individual	race	"White", "Native", "Asian", "Black", "Hispanic", "Multiracial", "Pacific"
Student's interest in math course	Individual	mathInt	Continuous composite variable of related survey questions
Students' math self-efficacy	Individual	mathEff	Continuous composite variable of related survey questions
How many friends get good grades	Individual	friendGrade	"None", "Less than half", "About half", "More than half", "All of them"
Hours spent on math homework	Individual	hoursHW	"No time", "Less than 1/2 hour", "1/2 to 1 hour", "1 to 2 hours", "2 to 3 hours", "4 to 6 hours", "7 to 9 hours", "More than 9 hours"
Taking calculus	Individual	calculus	"No", "Yes"
English spoken at home	Family	homeLang	"English", "NotEnglish"
Parent's highest level of education	Family	parEdu	"HighSchool", "LessHS", "Vocat", "AA", "BA", "MA", "DR"
Family Income	Family	famIncome	Used as continuous, though each unit means
Number of siblings	Family	sibling	Continuous **[removed]**
Teenager's parents / guardians divorced	Family	divorce	"Yes", "No" **[removed]**
How often discussed applying to college with parents	Family	discussCollege	"Never", "Once or Twice","Three/four", "More" **[removed]**

Math teacher's highest degree earned	Teacher	degreeTE	"Bachelor", "Master","Specialist", "PhD,MD"
Years taught math in HS	Teacher	yearsTE	Continuous
Percent of students who entered higher education	School	studentsHE	Continuous
School supports high achievers w / scholarships	School	scholarship	"No", "Yes"
Average SAT Math score	School	avgSATMath	Continuous **[removed]**
Average ACT Math score	School	avgACTMath	Continuous **[removed]**

Appendix B: Plots of Transformed Variables

Percent of students who entered higher education

Square of Percent of students who entered higher education

Family Imcome

Log of Family Income

Teaching Years

Log of Teaching Years
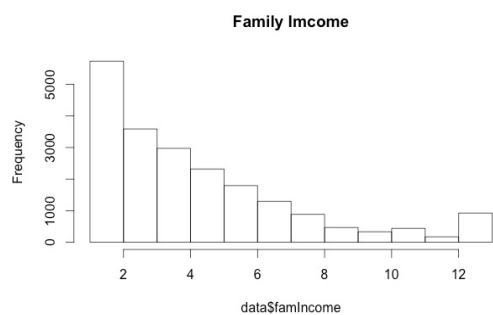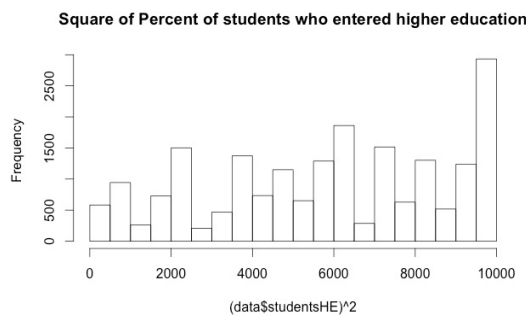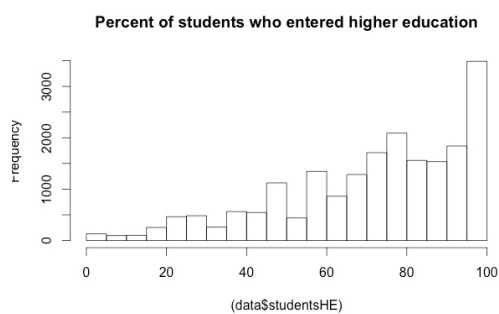
Appendix C: Percent of Observations Missing

race	sex	homeLang	parEdu	logfamIncome
0	0	0.00047	0.106598	0.106598
enroll	mathInt	mathEff	fGrade	hoursHW
0.120564	0.280504	0.155627	0.155456	0.226564
calculus	**sibling**	**divorce**	discuss	degreeTE
0.255605	**0.65458**	**0.655648**	0.656118	0.271108
logmathYears	scholarship	sqrtstudentsHE	studentsHE	caseLoad
0.273116	0.182148	0.138544	0.138544	0.182404
avgACTMath	**avgSATMath**	score		
0.648003	**0.580354**	0.120478		

Appendix D: Final Model Regression Output

```
Call:
lm(formula = score ~ race + sex + homeLang + logFamIncome + mathInt +
    mathEff + fGrade + hoursHW + calculus + logmathYears + scholarship +
    sqrtstudentsHE + sex:mathEff + race:homeLang + logFamIncome:sqrtstudentsH
E +
    sex:fGrade + mathInt:scholarship + mathInt:mathEff, data = newdata)

Residuals:
    Min     1Q  Median     3Q     Max
-44.212  -9.313   0.072  10.009  42.571

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.515e+01  1.488e+01   1.018 0.308848
raceNative                      -1.818e+01  1.035e+01  -1.757 0.079220 .
raceAsian                        5.311e+00  1.829e+00   2.904 0.003755 **
raceBlack                       -7.096e+00  1.718e+00  -4.130 3.90e-05 ***
raceHispanic                    -4.657e+00  1.420e+00  -3.280 0.001068 **
raceMultiracial                  2.984e-01  1.480e+00   0.202 0.840234
racePacific                      8.843e+00  1.039e+01   0.851 0.394997
sexFemale                        4.880e+00  2.080e+01   0.235 0.814569
homeLangNotEnglish              -2.567e+00  7.314e+00  -0.351 0.725654
logFamIncome                     9.292e+00  1.879e+00   4.945 8.76e-07 ***
mathInt                          1.517e+00  6.402e-01   2.370 0.017955 *
mathEff                          4.390e+00  7.472e-01   5.875 5.52e-09 ***
fGradeLess than half             3.236e+01  1.509e+01   2.145 0.032165 *
fGradeAbout half                 2.748e+01  1.487e+01   1.848 0.064836 .
fGradeMore than half             2.781e+01  1.486e+01   1.871 0.061605 .
fGradeAll of them                2.514e+01  1.493e+01   1.683 0.092570 .
hoursHWLess than 1/2 hour        4.962e+00  2.131e+00   2.329 0.020038 *
hoursHW1/2 to 1 hour             4.357e+00  2.051e+00   2.125 0.033835 *
hoursHW1 to 2 hours              7.380e+00  2.137e+00   3.454 0.000573 ***
hoursHW2 to 3 hours              1.025e+01  2.207e+00   4.646 3.78e-06 ***
hoursHW4 to 6 hours              8.710e+00  2.331e+00   3.736 0.000196 ***
hoursHW7 to 9 hours              1.141e+01  2.979e+00   3.831 0.000135 ***
hoursHWMore than 9 hours         1.104e+01  4.239e+00   2.605 0.009310 **
calculusYes                      1.724e+01  1.958e+00   8.807  < 2e-16 ***
logmathYears                     1.882e+00  4.844e-01   3.886 0.000108 ***
scholarshipYes                  -1.387e+00  8.761e-01  -1.583 0.113683
sqrtstudentsHE                   1.879e-03  3.917e-04   4.796 1.83e-06 ***
sexFemale:mathEff               -3.542e+00  8.886e-01  -3.985 7.16e-05 ***
raceNative:homeLangNotEnglish    3.909e+01  1.527e+01   2.560 0.010595 *
raceAsian:homeLangNotEnglish     1.070e+01  7.868e+00   1.360 0.174198
raceBlack:homeLangNotEnglish    -1.774e+01  1.268e+01  -1.399 0.162101
raceHispanic:homeLangNotEnglish  3.651e+00  7.725e+00   0.473 0.636630
raceMultiracial:homeLangNotEnglish -8.685e+00 1.265e+01 -0.686 0.492589
racePacific:homeLangNotEnglish          NA         NA      NA       NA
logFamIncome:sqrtstudentsHE     -5.566e-04  2.440e-04  -2.281 0.022721 *
sexFemale:fGradeLess than half  -1.651e+00  2.120e+01  -0.779 0.436319
sexFemale:fGradeAbout half      -9.360e+00  2.088e+01  -0.448 0.653985
sexFemale:fGradeMore than half  -6.726e+00  2.084e+01  -0.323 0.746999
sexFemale:fGradeAll of them     -3.478e+00  2.091e+01  -0.166 0.867943
mathInt:scholarshipYes           1.462e+00  8.607e-01   1.699 0.089615 .
mathInt:mathEff                  6.542e-01  3.874e-01   1.689 0.091513 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.43 on 1156 degrees of freedom
Multiple R-squared:  0.4176,   Adjusted R-squared:  0.398
F-statistic: 21.26 on 39 and 1156 DF,  p-value: < 2.2e-16
```

Appendix E: Model Checking Plots

Residuals vs Fitted
~ race + sex + homeLang + logFamIncome + mathInt + ma

Normal Q-Q
~ race + sex + homeLang + logFamIncome + mathInt + ma