# Influential Factors in Students' Aspirations to STEM Careers

Barbara Villatoro
Derek Eckman
Fern Van Vliet
Frankie Sullivan
Lorna Fullmer

December 11, 2019

Applied Regression Analysis
STP 530: 71790

# 1. Introduction

A Science, Technology, Engineering, and Mathematics (STEM) focused work force is necessary for the United States to continue leading the global economy. Obama's administration deemed STEM education a priority and stated that it must be accessible for all students to ensure the Nation's future. (Handelsman & Smith, 2016). Consistent with this position, the U.S. Department of Education increased funding for STEM education programs that increased accessibility to STEM subjects and increase student choice of STEM degrees.

Extensive research has identified gender, race, and socioeconomic status (SES) as key factors in student achievement in STEM subjects and their choice to pursue STEM careers. Farrell & McHugh (2017) investigated STEM-bias among male and female university students. Students were divided into groups and identified as Female-STEM, Female-non-STEM, Male-STEM, or Male-Arts based on their current studies in the university. These students were asked to categorize target words i.e. science, mathematics, physics, chemistry, computing, engineering, arts, English, drama, French, History and music, when answering the question "Men or Women are more suited to". All groups in the study exhibited some level of Male-STEM and Female-Arts bias in response to the questions.

A study by Niu (2017) investigated family SES and choice of STEM majors in college. A focus of the study was to identify influencing factors of choosing a STEM major that are beyond the student's control. One finding of the study was that lower SES is associated with students' choice of STEM majors due to lack of access to information regarding the nature of STEM majors and STEM occupations.

Student self-efficacy in STEM subjects influences students' views of STEM education. Dubriwny, Pritchett, Hardesty, & Hellman (2016) investigated the impact of a STEM program on middle-aged school students and their perceived self-efficacy and attitudes towards technology and engineering. Students were given pre- and post-surveys after participating in a program at a fabrication laboratory building projects using basic cutting, milling and electronic tools. The found a positive association between middle school students' increased self-efficacy and positive attitudes towards engineering and technology.

In light of previous research findings, we investigated through a quantitative analysis the variables we thought may influence student choices in STEM careers. Our data was from the High School Longitudinal Study of 2009 (HSLS:09). A research question posed as part of HSLS summarizes the focus of our analysis: "What factors influence students' decisions about taking STEM courses and following through with STEM college majors?"

# 2. Data Source

The National Center for Education Statistics (NCES) began gathering data about ninth grade students from across the United States as part of the High School Longitudinal Study of 2009 (HSLS:09). The publicly released data from the HSLS includes follow-up data on the same group of students collected in 2016. Since our research focused on what factors in 2009 could be used to predict a student's belief that they will have a STEM career at age 30, we only used data from the base year of data collection, the 2009-2010 school year. This data was collected in the fall of 2009 (Ingels et al., 2011, p. 6).

Data was collected through the administration of surveys to students, students' parents, principals, and students' math and science teachers. Students were also given a mathematics exam to determine their current mathematical skills relative to other ninth graders in the United States. The variables that we chose to consider as predictor variables were found on the student questionnaire, parent questionnaire, and the mathematics exam. The response variable, the student's envisioned career at age 30, was found in the student questionnaire. Student surveys were only offered online while surveys for the other groups could be completed online or over the phone. All but two of the schools allowed students to complete the questionnaire either on a school computer or a project provided computer during school hours. Students at the two remaining schools completed their surveys online in their own time out of school.

The researchers had a two stage process for achieving a random sample. In the first stage, a school was randomly chosen to participate in the study. To ensure that the schools selected would be representative of the United States in general, the schools were stratified by three variables: school type, region of the United States, and locale. "School type" was used to categorize schools as public, private-Catholic, and private schools which were not Catholic. Schools were also split into four regions of the United States: Northeast, Midwest, South, and West. The third variable, "locale", placed schools into categories of city, suburban, town, and rural (Ingels et al., 2011, p.38).

The "target population" for schools was "regular public schools, including public charter schools, and private schools in the 50 United States and the District of Columbia providing instruction to students in both the 9th and 11th grades" (Ingels et al., 2011, p. 35). Study-ineligible schools were any schools meeting any of these criteria:

- Bureau of Indian Affairs (BIA) schools;
- Special education for students with disabilities;
- Career technical education (CTE) schools that do not enroll students directly;
- Department of Defense (DoD) schools that do not enroll students directly;
- Schools without both a 9th and 11th grade;
- Schools not in operation during the fall of 2009;
- Juvenile correction/detention facilities;
- Other schools that address disciplinary issues but do not enroll students directly;
- Ungraded schools (i.e., no metric to define students as being in the ninth grade);
- Schools that only offer testing services for home-schooled students; and
- Schools that do not require students to attend daily classes at their facility.

(Ingels et al., 2011, p. 35).

Once an eligible school was chosen, a random sample of students within the school was selected for data collection. If a school was deemed eligible for the study, then all ninth-grade students at the school were eligible to be part of the sample for the study. Students who could not complete the assessment or questionnaire due to disabilities or language barriers were kept in the sample to collect data from surveys not completed by the student, but these students were classified as "questionnaire-incapable".

One potential concern about the reliability of our data that needs to be further investigated in order to make a claim about STEM careers in the US population as a whole is the amount of missing and incomplete data. In our initial analysis of the predictor variables, we noticed that within the subset of students with complete data, there are fewer students in the lowest income brackets and fewer students with English as a second language than we expected. More analysis would be necessary to determine the extent of bias caused by the missing and incomplete data. Alternatively, as we will suggest later, bias could be controlled by analyzing within sub-populations of the data. For example, we could make a model for the career choices of only students who speak English as a first language, whose family is not in poverty, and live in an urban area. It may be easier to form a model with this specific sub-population rather than trying to create a model that describes every student at every school.

# 3. Data Description

*Handling Missing Data*

Reserve codes of (-9), (-8), and (-7), respectively, apply to scenarios when a participant did not respond to a particular questionnaire item that would have been applicable, data was missing for an entire questionnaire from a potential respondent in the sample, and an item was skipped because it was not applicable. We re-coded data from the sample with any of these codes with any of these codes as NA.

*Psychological Scales*

Three of our predictor variables--namely student math identity, student math utility, and student sense of school belongingness--were variables HSLS designed "to be analyzed as psychological scales". The items informing them, which we specify in their individual descriptions, involve rating a level of agreement on a four-point Likert scale (1=strongly agree, 4=strongly disagree). HSLS analysts then converted the resulting scale values to be on continuous scales with mean 0 and standard deviation 1.

*Descriptions of Variables*

In this section we detail the variables we used in our model, with each variable listed in the format: Name of Variable ("HSLS naming convention"), variable type, followed by a qualitative description of the data collected for measurement. We list the dependent variable, "Student-Occupation at Age 30", first. All other variables in the list are predictors.

- Student-Occupation at Age 30 ("X1STU30OCC_STEM1"), categorical

Our original intent for the "student occupation at age 30" variable was to determine whether a student anticipated working in a field related to STEM--traditionally standing for science, technology, engineering, and mathematics--by age 30. Student participants wrote down the occupation they "thought they would have when they were age 30". HSLS analysts then coded the occupations students wrote using version 13 of the occupational information network (O*NET) taxonomy, and categorized the coded occupations into the categories given above. From these HSLS categories, our research group combined the four STEM sub-domains identified by HSLS into a single category to create a binary variable with the categories "STEM" and "non-STEM". We inferred that HSLS analysts used the category "uncodable" when they could not determine whether or not an occupation was "STEM". We inferred this meaning ourselves since we found no

explicit description of "uncodable" occupations in the HSLS documentation. Using our inferred meaning for "uncodable", and bearing in mind that a fairly small portion of the sample fell into the "uncodable" category, we counted the "uncodable" data as missing data.

After creating our binary dependent variable with the categories "STEM" and "non-STEM", we reflected on what we were really interested in predicting: whether students planned to have a career in mathematics, physical sciences, engineering, or information technology. Consequently, we further examined the "STEM" categories of our binary variable and realized that we were really only interested in occupations falling under the HSLS category (1) Life and Physical Science, Engineering, Mathematics, and Information Technology Occupations, as opposed to the umbrella category "STEM", which contained HSLS categories (1), (2), (5), and (6). As a result, we created a new dependent variable. We used "ME" to denote occupations under the original HSLS category (1), and "non-ME" to denote all other occupations, that is, occupations in HSLS categories (0), (2), (5), and (6). "ME" and "non-ME" were the categories for our new variable.

> Original HSLS categories: (-9) Missing; (-8) Non-response; (0) Non-STEM; (1) Life and Physical Science, Engineering, Mathematics, and Information Technology Occupations; (2) Social Science; (5) Two sub-domains (originally architecture and health); (6) Unspecified sub-domain; (9) Uncodable

**Frequency Distribution of Students' Planned Career by Age 30**

| HSLS Category (1) | non-Stem: HSLS Category (0) |
|---|---|
| 1726 | 13969 |

From the frequency distribution it is evident that the data are unevenly represented; there are far more students who anticipate going into careers that are non-STEM (HSLS category (0)) than students who plan to pursue a career in life/physical sciences, engineering, math, or information technology (HSLS category (1)).

- Parent-Education ("X1PAREDU"), categorical

The parent(s) of the student participants indicated the highest level of education they received. HSLS analysts then assigned a category according to the maximum education level between the parents. When necessary, HSLS analysts used imputed data to infer the highest level of parent education. For the public data, HSLS merged what were originally two separate categories for the master's degree and educational specialist diploma.

> Original HSLS categories: (-9) missing; (-8) non-response; (1) less than high school; (2) high school or G.E.D.; (3) associate degree; (4) bachelor's degree; (5) master's degree or educational specialist diploma; (7) doctoral degree or equivalent

**Frequency Distribution of Parent Education**

| Less Than High School | High School | Associate's | Bachelor's | Master's / Educational Specialist Diploma | PhD |
|---|---|---|---|---|---|
| 1010 | 5909 | 2549 | 4102 | 2116 | 1096 |

The frequency distribution indicates that the most common type of highest level for parent education is a high school education or equivalent. However, quite a few parents in the sample had education beyond high school; a combination of the categories "AA-associate degree", "BA-bachelor's", "MA-master's or educational specialist diploma", and "PhD" reveals more parents in the sample with an education beyond high school (9863 households) than parents with a high school education or below (6919 households). According to the 2010 US Census, only 33.4% of Americans have a Bachelor's degree or higher. The percentage of households which have at least Bachelor's degrees is much higher than expected. If the level of parent education is used as a predictor, analysis should be done to examine the reason for this discrepancy or to control for the potential bias by analysing results within specific parental education levels.

- Family Income ("X1FAMINCOME"), categorical

The parent(s) of student participants estimated, to the best of their ability, the income bracket corresponding to the total income for their household. This gives rise to concern about reliability of values since the data is self reported. HSLS analysts used imputed data to categorize family income when necessary.

> Original HSLS categories: (-9) missing; (-8) non-response; (1) less than $15,000 annually; (2) - (12) Income between $15,000 and $235,000 annually, with each category representing a $20,000 bracket (ie. category (2) denotes an annual income that falls between $15,000 and $35,000); (13) more than $235,000 annually

**Frequency Distribution of Family Income by Bracket**

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1570 | 3043 | 2762 | 2514 | 1855 | 1484 | 964 | 741 | 367 | 237 | 316 | 116 | 792 |

The publically available data gives the income bracket (not income amount) for each student. To minimize error, we assigned to each category the median value of its corresponding income bracket. Since the income values for category (13) had no upper bound, an appropriate value could not be determined, so we decided to treat the 792 responses in category (13) as unknown and coded them as NA. Analysis should be done on the impact of excluding these 792 participants if income is used as a predictor variable. Family income was not a significant predictor for the logistic regression model or the decision trees in our analysis.

An additional concern about bias can be observed from the data in the sample. According to the U.S. census bureau, the median U.S. household income in 2009 was 50,303. The median household income for our sample, however, is in bracket (4): between 55,000 to 75,000. Analysis should be done on the reason for the discrepancy and its significance.

- Student-Gender ("X1SEX"), categorical

HSLS analysts identified student gender by asking student participants to specify their gender, asking the students' parents to specify their child's gender, and/or through the sampling roster from the students' schools. Whenever any source was inconsistent with the other two, HSLS analysts coded a student's gender by manually reviewing the student's first name.

Original HSLS categories: (-9) Missing; (1) male; (2) female

**Frequency Distribution of Student Gender**

| Male | Female |
|---|---|
| 11973 | 11524 |

The frequency distribution reveals that our sample had a fairly even representation of each gender. The ratio of female to male students is approximately 0.96.

- Student-First Language ("X1DUALLANG"), categorical

Students indicated the first language they learned "to speak when you were a child" by selecting one of the following: English, Spanish, Another language, English and Spanish equally, and English and Another language equally. The publically released data categorized the responses into the three categories below.

Original HSLS categories: (-9) Missing, (-8) Non-response, (1) English only, (2) non-English only, (3) English and non-English equally

**Frequency Distribution of Student First Language**

| English Only | non-English Only | English and non-English Equally |
|---|---|---|
| 17863 | 2201 | 1355 |

The frequency distribution reveals significantly more students who spoke only English as a first language than students who spoke only a language other than English and students who spoke both. The proportion of students who spoke English only as a first language is higher than expected. Analysis should be done to confirm, and if this high proportion differs significantly from the national average, bias should be controlled for by analysing within specific categories.

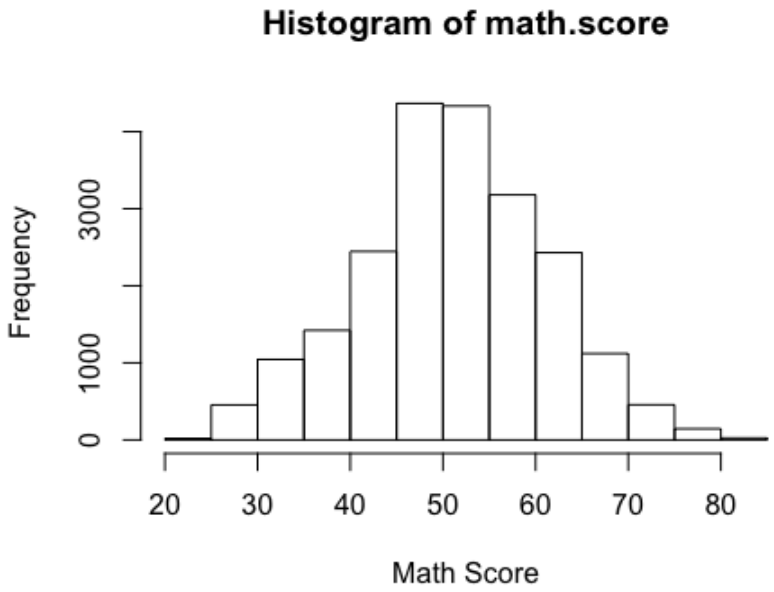
- Student-Standardized Math Theta Score ("X1TXMTSCOR"), numeric/continuous

The standardized math score is a percentile score of a student's performance relative to other students participating in HSLS. The score was transformed (scaled and shifted numerically) from a raw theta score on a test HSLS researchers developed with the intent of assessing students' algebraic reasoning. The data in the transformed score we report has a mean of 50 and a standard deviation of 10. The test consisted of a routing stage with 15 items, and a second stage for assessment, with items whose difficulty depended on a student's performance in the first stage.

Normal Range: 30.4, 70.6

Total Min: 24.1

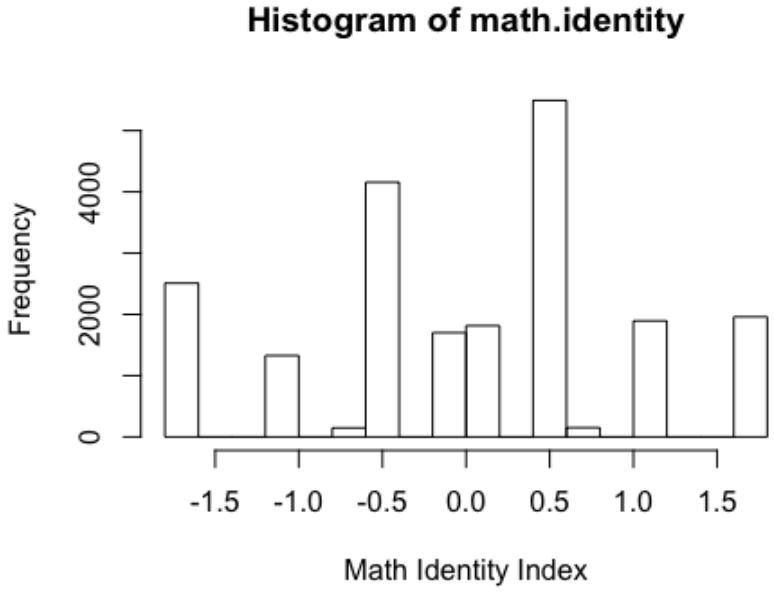
Total Max: 82.2




The percentile math scores of students in our sample are close to being a normally distributed.

- Student-Math Identity ("X1MTHID"), numeric/continuous

Students rated their agreement with the statements "you see yourself as a math person" and "others see you as a math person".




Normal Range (95 percentile interval): -1.73, 1.76

Note that in this case, the bounds of the normal range are equivalent to the min and max values across the entire dataset.

The most common scores are slightly above and slightly below neutral, there are quite a few moderate (approx. 1 std dev above/below neutral) and significant (approx. 1.5 std dev above/below neutral) scores as well.

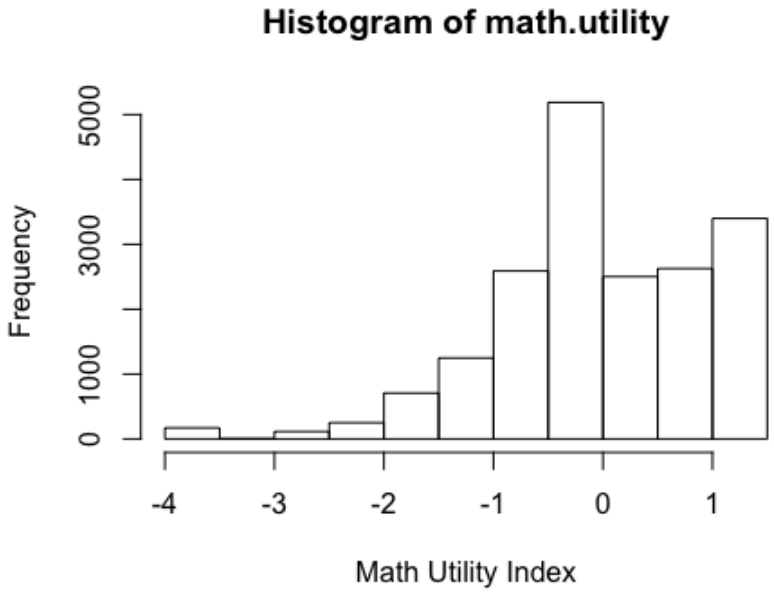
- Student-Math Utility ("X1MTHUTI"), numeric/continuous

Students rated their agreement with the statements "[your math class] is useful for everyday life", "[your math class] will be useful for college", and "[your math class] will be useful for a future career".

Normal Range (95 percentile interval): -2.30, 1.31

Min: -3.51

Max: 1.31




The data for math utility appears to be skewed left according to the histogram. Low scores for math utility are far more extreme (up to four std dev below neutral) than the higher scores (up to 2 std dev above neutral).

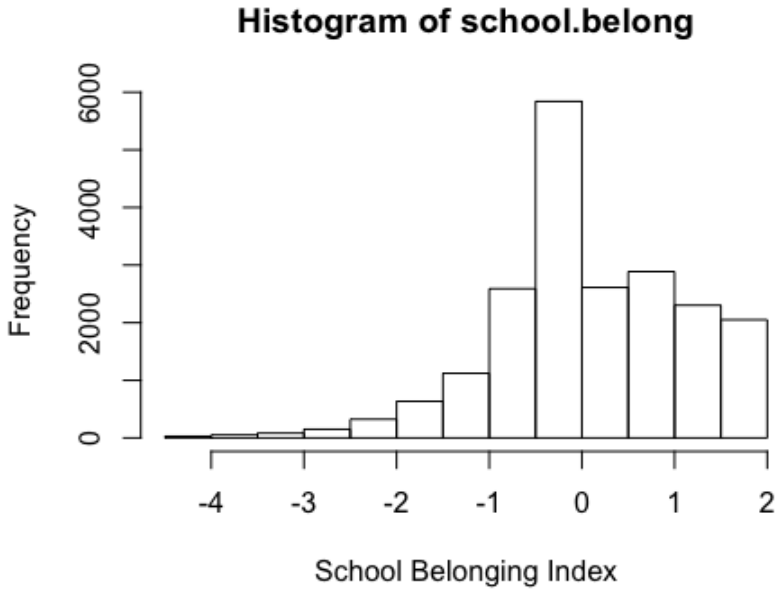
- Student-Sense of School Belonging ("X1SCHOOLBEL"), numeric/continuous

Students rated their agreement with the statements "you feel safe at this school", "you feel proud being part of this school", "there are always teachers or other adults in your school that you can talk to if you have a problem", "school is often a waste of time", and "getting good grades in school is important to you".

Normal Range (95% percentile interval): -2.12, 1.59

Min: -4.35

Max: 1.59




The distribution of data for sense of school belongingness is skewed left.

# 4. Modeling (Methods and Results)

The first step in analyzing the data was recoding the categorical data so that R would recognize it as factors. The data from the source was numerical coding. Additionally the codes -9 (question response missing), -8 (the entire form the question was on was not submitted), and -7 (the questions was not applicable) were recoded as NA since we could not determine what the subject's responses were.

The data for income was partitioned into income brackets so we assigned the middle value for the bracket the data value was in, with the exception of the last bracket, which was assigned NA since the bracket did not include an upper limit so a midpoint could not be determined. If we had used income as an indicator variable we would have limited our study by removing more than 700 students from the analysis since their data set would be incomplete. However, income did not seem to be significant when we included it as a predictor in our model, so we decided to remove it as an indicator variable.

We additionally removed the categorical variable which identified each student's family as above or below poverty level as an indicator variable. We made this decision because the percentage of students' families in the sample whose family would be classified as below the poverty level was a small percentage (<16% of responses that were not missing, 11% of all subjects) with almost 29% of all responses missing. From this we suspect that a significant portion of students whose families would have been scored as below poverty level may not have submitted responses, which we thought might disproportionally effect the regression analysis.

Additionally, in analyzing the way the response variable (the career that the student thought they would have when they are 30) we made a decision to adjust our research question from predicting if the student's choice was coded as STEM to if the student's choice was coded as ME

(Math, Engineering, life, or physical sciences). Students whose responses were Non-STEM or ME were recoded as NA. Part of the reason we made this choice was because careers such as nursing were being coded as a STEM major, but we were interested in looking at students who chose careers in math and fields closely related to math.

We decided to see if a general linear model would be a good model to predict if the career that a student predicted for themselves by age 30 would be scored by the data collectors as Math, engineering, life, or social sciences. We chose to first consider this model because of the potential interpretability and because the coefficients of the model could tell us information about how the indicator variable influences the odds that the student's choice is ME.

We then split the data into 80% training data and 20% testing data. We used the 80% training data to perform all of the analysis. However, most models when tested with testing data predicted no students would select an ME career. We address this issue and potential solutions later in this section.

Prior to building the model we looked at the pairs graphs and the VIF's to verify that there was not an issue with multicollinearity. See Figure 1 and Table 1. Both the pair-wise graphs and the VIF's indicated that there was not an issue with multicollinearity.

Figure 1: Pairwise Graphs

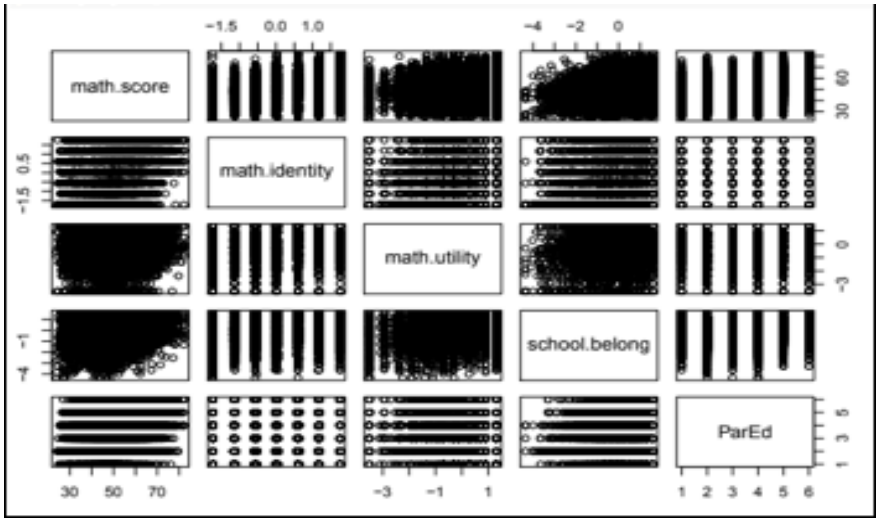



Table 1: VIF Values

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| math.identity | 1.348304 | 1 | 1.161165 |
| math.utility | 1.188418 | 1 | 1.090146 |
| school.belong | 1.149002 | 1 | 1.071915 |
| Gender | 1.007894 | 1 | 1.003939 |
| ParEd | 1.266943 | 5 | 1.023943 |
| Language | 1.077238 | 2 | 1.018774 |
| math.score | 1.407091 | 1 | 1.186209 |

We used a forward selection method, using significance level to build the model. Since the model with the largest absolute value of the test statistic would have the smallest significance level

it is equivalent to compare the test statistics. The p-values were very small so we chose to compare the test statistics.

While looking at the model options involving a single predictor Gender was the most significant predictor. The next variable that was suggested was school belonging, followed by math score, math utility, and finally math identity. When we ran the model with all 5 indicator variables we noticed that school belonging was not statistically significant. We compared the models with and without school belonging as a predictor and the model with school belonging as a predictor had a McFadden's Pseudo R^2 value that was only 0.001 larger than the McFadden's Pseudo R^2 value of the reduced model that did not include school belonging. Since the Pseudo R^2 values for both models were very similar, we decided to remove school belonging as an indicator variable in order to simplify our model.

We also considered adding interaction terms. We ran a significance test to determine if the interaction model was a better fit (alternate hypothesis) or if the models provided equally good fit (null hypothesis). See Table 2. The p-value was 0.003 (indicating the model with the interaction term is a better fit, but when we compared the McFadden Pseduo R^2's there was very little difference. The McFadden Pseudo R^2 without the interaction term was 0.0498. The McFadden Pseudo R^2 for the interaction model was 0.0513 (an increase of only 0.0014). We decided that the small increase in the Pseudo R^2 was not worth the loss of interpretability.

Table 2: Significance Test for Interaction Terms

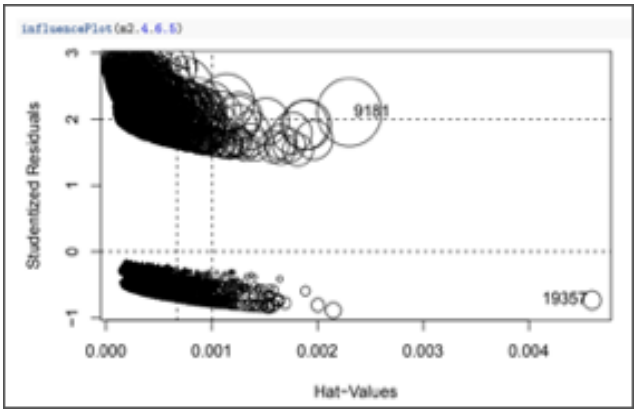
```
Analysis of Deviance Table
Model 1: STEM_numeric ~ Gender + math.score + math.identity + math.utility
Model 2: STEM_numeric ~ math.identity + Gender + math.utility + math.score + math.identity:Gender

     Resid.Df   Resid.Dev    Df    Deviance
1    14902      7955.2
2    14901      7938.8       1     16.445

pchisq(6130.4-6121.6, df=1, lower.tail=F)

[1] 0.003012305
```

We decided that the best general linear model we could get was $\ln(\pi/(1-\pi))$= -3.6106832 - 0.805647(Female)+0.027575(Math score)+0.126673(Math identity)+0.289827(Math utility).

While all the indicator variables are statistically significant, the general linear model is not in general a good predictor of the response variable (as noticed with a very small McFadden Pseudo R^2 of less than 5%). To improve the prediction power additional predictor variables should be considered in addition to the ones we considered in building our model. Additionally, we may have had difficulty because we were trying to build the model on a population which was too diverse. Focusing on specific subclasses (for example urban students whose families are above the poverty level) might allow us to make better predictions of the odds of the student's choice being ME within the specific subclass.

We checked our final model to make sure there weren't any data values that were overly influential. To do this we looked at the influence plot. See Figure 2. From the influence plot there were a couple of data values that indicated a potential issue with being overly influential.

Figure 2: Influential Plot




To investigate further we looked at the DFBetas and the DFFits to see if there were any points that were too influential on the intercepts or on the model itself. We considered a point to be overly influential if the DFBeta score was above 1.22 or the DFFit score was above 0.7 (for our sample size). We then calculated the maximum of the DFBeta's (0.15) and the maximum of the DFFits (0.02). Since neither maximal value met the threshold for significant influence, we concluded that there were not any points which were overly influential.

Since the indicator variables were statistically significant we looked at what we could learn about the effect that the indicators have on the odds the student would anticipate an ME career.

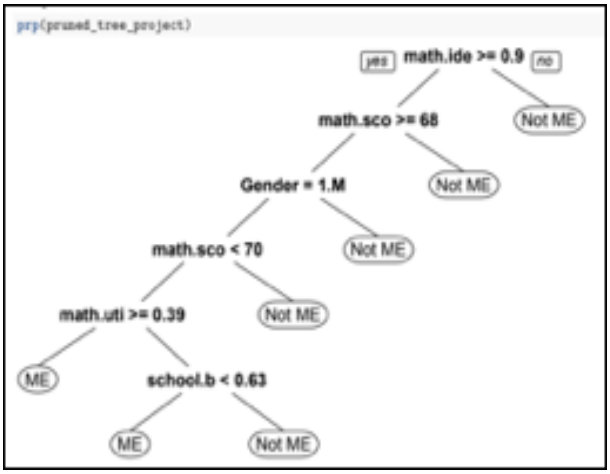
Table 3: Model

```
Call:
glm(formula = STEM_numeric ~ Gender + math.score + math.identity + math.utility,
family = binomial, data = project_train_logistic)

Deviance Residuals:
Min         1Q          Median       3Q         Max
-0.8840     -0.4590     -0.3625      -0.2825    2.8784

Coefficients:
                  Estimate     Std. Error    z value     Pr(>|z|)
(Intercept)       -3.606832    0.184535      -19.546     < 2e-16 ***
Gender2.FEMALE    -0.05647     0.064390      -12.512     < 2e-16 ***
math.score        0.027575     0.003382      8.154       3.53e-16 ***
math.identity     0.126673     0.035749      3.543       0.000395 ***
math.utility      0.289827     0.035134      8.249       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our model suggests that, while holding all other variables constant, the odds of a female student anticipating themselves in an ME career at 30 is decreased by (1-e^(-0.805647))*100% = 55%. See Table 3. Additionally, while holding all other variables constant, the odds that a student with a math identity anticipates an ME career is increased by e^(0.289827)*100% = 34% by for every additional increase by 1 standard deviation in their belief about the utility of the math class they are currently enrolled in.

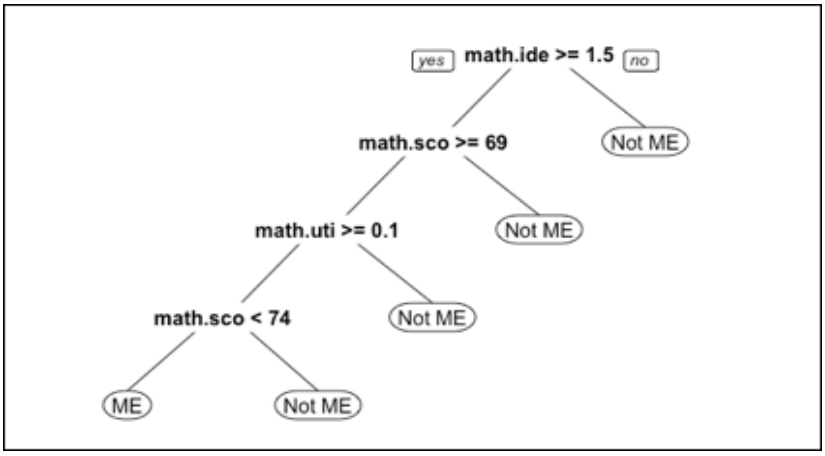
Since the McFadden Pseudo R^2 was very low, we decided to investigate other prediction models. We considered a decision tree as a model (again for the interpretability afforded by a tree).

Figure 3: Decision Tree




Notice that the tree indicates that math identity, math score, gender, and math utility are all significant predictors (consistent with the general linear model (logistic)). However,, we noticed that all female students were routed to not-ME. We decided to use a tree for females. to understand more about the female students. See Figure 4.

Figure 4: Decision Tree for Females



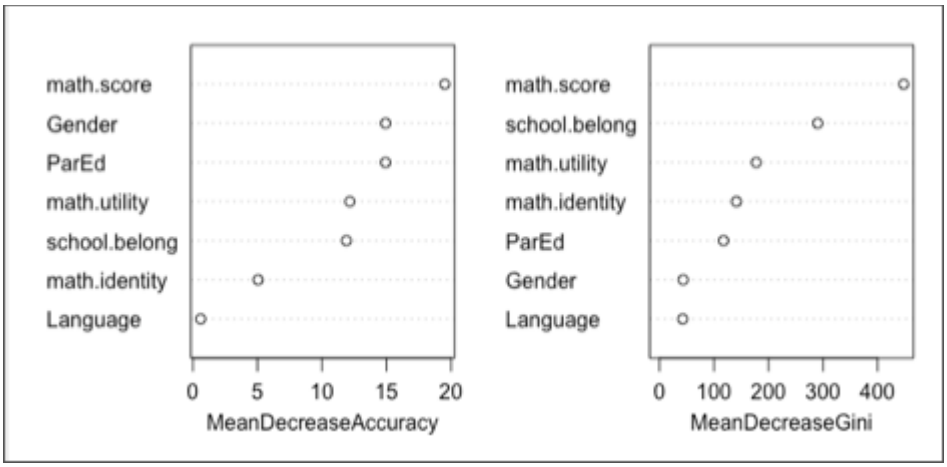

In the tree for the female students, we see one possible reason for why no female students were being routed to ME in the decision tree. Considering only the female students, to be routed to ME the student would need to have a math identity that is 1.5 standard deviations above the mean, and a math score that is almost 2 standard deviations above the mean.

Similar to the logistic regression model, the trees did not predict any students from the testing sample would select an ME career. In future research, we would continue to suggest narrowing the population of students being considered to some sub-class of the total population. Adding additional indicator variables could also be helpful in future research.

We also considered a forest to look at which variables would come up as significant while we control for over fitting. The forest indicates that the student's math score, gender, the parents' education level, math utility and math identity are important predictors. See Figure 5. This is not

surprising as we included all the original predictors because previous research indicated that they are important predictors of students pursuing STEM careers (so should also be important predictors for ME careers).

Figure 5: Forest Analysis




# 5. Discussion

Our group's findings, limitations, and future direction for this project will be the focus of the discussion section. First, we will discuss the significant predictors that persisted throughout the logistic regression analysis, the tree analysis, and the random forest analysis. Second, we will acknowledge the limitations of our findings, which include (1) potential issues that we found or conjecture to be within the data sample, (2) the level of practical significance of our findings, and (3) the difficulty in analyzing only students who were interested in math or engineering careers. We will conclude by discussing future research steps that can be taken by either our group or another researcher.

Our group found four main predictors that persisted throughout the logistic regression analysis, the individual tree analysis and the random forest analysis: (1) gender, (2) math score, (3) math utility, and (4) math identity. We anticipated the significance of several of these variables due to the literature that we reviewed before completing the project. Previous researchers have indicated that female students encounter more difficulties than their male peers in STEM courses (cite), the affective characteristics of mathematics utility (cite) and math identity (cite) play a considerable role in students' attitudes towards mathematics and their personal belief that they can learn mathematics and successfully apply mathematics in context.

Gender appeared to be one of the most significant indicators in our analysis. Depending on our model, either gender or math score was always the first predictor. When conducting our random forest analysis, we also determined that the inclusion of math score and gender would decrease the mean accuracy by the largest amount. However, nearly every decision tree that we created routed all female students to a non-ME career aspiration. Furthermore, the decision tree that we created with just females places an extremely stringent set of qualifications for a female to aspire to an ME career. Combined with our group's finding in the logistic model that holding all other variables

constant, we can conclude being female instead of male decreased the odds that a student might foresee themselves in an ME career.

However, our tree and random forest diagrams typically predicted that most, if not all, students would aspire to a career that did not fall within mathematics, engineering, life or physical sciences (regardless of gender). As we noted previously, our group's realization that our logistic regression model predicted that absolutely no students (male or female) would aspire to an ME career helped us determine that we should include more predictor variables in future models. We also decided that limiting our models to specific subclasses of the original data sample might improve the practical significance of our models.

In the following paragraphs we describe specific limitations in the data and our findings, which include potential issues and bias in the HSLS:09 sample, implications of the original coding scheme for the career variable on the efficacy of our models, and our group's determination that although our models carried high levels of statistical significance, we could do a great deal to improve the practical significance of our findings.

The High School Longitudinal Study of 2009 collected preliminary survey data from 21,444 students. The data collection team also instructed these students' parents, school counselors, and administrators to complete a corresponding survey. However, when our research group analyzed the original data, we determined that only 13,705 students had complete data for each of the variables that we considered in our analysis. Some of the students survey items blank that corresponded to a particular variable and were coded -9 for that variable. Other students (or their parent, counselor, or administrator) failed to complete their survey, resulting in a code of -8 for that student.

We also noted that the proportion of students whose parents claimed high socioeconomic status was much higher than anticipated and that the proportion of students who claimed that English was not their first language was lower than expected (provide number?). These findings lead our group to conjecture that some classes of students were unintentionally but disproportionately excluded from the data set. For example, a student whose parent does not read or write English well might not have completed the parent questionnaire, resulting in that student being excluded from our analysis. Additionally, a counselor at a school with a high percentage of students in poverty could have had an enormous case load and not completed their version of the survey, thus excluding a large proportion of students whose parents likely have low educational backgrounds and socioeconomic status from our survey. Thus, we claim that there is not enough information in the data collection methodology to determine that the final raw sample of complete responses is unbiased in terms of factors such as socioeconomic status and parent education.

If the data were biased in this way it could account for the under-importance of certain variables in our models that we have observed through our teaching experience to be important to student career aspirations, including parental education and sense of school belonging. Our analysis indicated that including both parent education and sense of school belonging in our model would result in large decreases in the mean accuracy of our random forest. Furthermore, school belonging was second only to math score on our mean decrease of the Gini index chart. However, neither of these indicators appeared important when we were using forward selection to create our logistic regression model. Our group was surprised by the disclusion of predictors deemed important by the education field and felt that such a disclusion could only be possible if there were inherent issues with the data sample. Thus, we feel that the snapshot of data that we analyzed could have bias that we did not account for.

Our group's original research question could be paraphrased as "What factors influence a student's perception that they will pursue a STEM career?" However, the researchers who collected the data had a much broader definition of what occupations constituted a "STEM career" than we did. The group that collected the original data determined that "STEM" included careers in social sciences (e.g. psychologists, social workers, education professionals), architecture, and health occupations (nurses, doctors, lab technicians, pharmacists, chiropractors, etc). We wanted to focus our research on math and engineering occupations. The only variable that appeared to match our focus was a career variable that included life and physical sciences, mathematics, engineering, and information technology occupations. As indicated earlier, our research group named this career variable ME. Only 1726 students in the original sample chose an occupation that we considered to be an ME occupation. Thus, only 8.1% of the students with complete responses to their survey chose an ME career. Our team did not conduct analysis to determine what proportion of the 8.1% of students who chose ME were retained after we excluded students who had incomplete responses on the other variables. However, since the available data grouped math and engineering careers with physical sciences, life sciences, and information technology occupations we were not able to investigate students' choice of a mathematics or engineering career separate from the other careers in the category.

Our research group created several logistic regression models to try to predict what factors would result in a student aspiring to a career in ME. Our final model showed that each of the predictor variables involved (gender, math score, math identity, math utility) were highly significant. However, none of our models had a McFadden's $R^2$ value greater than 0.08, which correlates with roughly a 0.1 value for the traditional $R^2$. Thus, none of our models were practically significant. Even our full model with all of the potential predictors did not produce a value for McFadden's $R^2$ that indicated practical significance. The inclusion of interaction terms did not significantly increase the value of McFadden's $R^2$ either. Our group is in the process of planning how to improve the practical significance of our models and increase the reliability of the data we select to analyze.

We outline our group's future direction in the concluding paragraphs. If our group were to continue with this research beyond this project, we would attempt to (1) focus on predicting the odds that certain subpopulations within the original data set will aspire to an ME career (such as students who speak English as a first language and whose families are not in poverty), and (2) examine the correlations between the 2009 initial data and follow-up data.

Our group made several conjectures when discussing potential bias in the original sample. We conjectured that students with low socioeconomic backgrounds and English language difficulties were, unintentionally, disproportionately excluded from the study. We would like to further investigate these conjectures to determine if there is bias within the raw data. If we found bias, we could refine our research question in such a way that we can only apply our findings to the appropriate groups of people.

We would also examine the source of the numerous -9 and -8 codes within the data and determine which survey items were unanswered and which individuals did not complete surveys. If a student was labeled -8 on certain compound variables because one of the sub-variables came from the parent questionnaire, we could explore using different variables that do not take into account any responses from the parent questionnaire.

We would also seek access to the raw student answers that were used to categorize students into their predicted STEM careers. Our group believes that the predicted STEM careers survey

question was open-ended and that the original data collection team sorted the students' open-ended responses into categories. Thus, if our research team could get access to the raw data we would be able to self-determine which students actually wanted to pursue careers in mathematics or engineering. Our team would thus be able to conduct a more rigorous and applicable analysis on students who aspire to the careers we wanted to address in our research question: math and engineering.

In conjunction with our data selection, our research group would also determine a larger number of predictors to add to our model. Social sciences research does not typically produce strong results when only a few predictor variables are considered, which we found when trying to increase the value for McFadden's $R^2$ in our models. Thus, we would increase our predictors and explore both additive models and models that include interaction terms in an effort to produce a model that was both statistically and practically significant.

Finally, we would examine the follow up data to the original study. Follow up data was collected in 2012 (when most of the students in the initial study graduated high school) and in 2016 (when many of the students in the initial study finished undergraduate school). A cursory glance at the change in predicted career responses between 2009 and 2016 reveals that only 8.1% of the students who predicted an ME career for themselves in 2009 still predicted an ME career for themselves in 2016. On the other hand, approximately 4% of students who did not predict themselves in any "STEM" career (as defined by the original data collection team) in the 2009 study predicted themselves in an ME career in 2016. This transition from non-STEM to ME accounted for 53% of the total students in 2016 who said they envisioned themselves in an ME career. This unexpected finding implies that there is a lot more to the story of what factors encourage a student to persist in their desire to pursue an ME career than merely a desire to be a mathematician, engineer, physicist, biologist, chemist, IT professional, or other STEM professional in high school.

The students in the original study sample will turn 30 years old in either 2024 or 2025. Our group is interested to see what percentage of the original students will actually be in an ME career at this stage of their lives and what those students' career aspirations were in 2009. What the original data team finds will likely be unexpected and thought provoking.

# References


Dubriwny, N., Pritchett, N., Hardesty, M., & Hellman, C. (2016). Impact of Fab Lab Tulsa on Student Self-efficacy Toward STEM Education. *Journal of STEM Education : Innovations and Research, 17*(2), 21-25.

Farrell, L., & Mchugh, L. (2017). Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). *Journal of Contextual Behavioral Science, 6*(1), 80-90.

Handelsman, Jo & Smith, Megan (2016). STEM for All [Blog post]. Retrieved from https://obamawhitehouse.archives.gov/blog/2016/02/11/stem-all

High School Longitudinal Study of 2009 (HSLS:09). (n.d.) Retrieved from https://nces.ed.gov/surveys/hsls09/

Ingels, S.J., Pratt, D.J., Herget, D.R., Burns, L.J., Dever, J.A., Ottem, R., Rogers, J.E., Jin, Y., and Leinwand, S. (2011). High School Longitudinal Study of 2009 (HSLS:09). Base-Year Data File Documentation (NCES 2011-328). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved November 26, 2019 from http://nces.ed.gov/pubsearch.

Kutner, M., Nachtsheim, C., Neter, J., & Li,W. (2005). *Applied linear regression models*. New York, NY: McGraw-Hill/Irwin.

National Center for O*NET Development. O*NET 13.0 Database — occupations updated by job incumbents/occupational experts - Occupational Listings. *O*NET Resource Center*. Retrieved December 11, 2019, from https://www.onetcenter.org/listings/13.0/updated.html

Niu, L. (2017). Family Socioeconomic Status and Choice of STEM Major in College: An Analysis of a National Sample. *College Student Journal, 51*(2), 298-312.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. *High School Longitudinal Study of 2009 (HSLS:09).* [Codebook_191111180427.txt]. Retrieved from https://nces.ed.gov/OnlineCodebook/Session/Codebook/c1a62c9b-ae60-4556-90e0-5dea27a1beeb

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. *High School Longitudinal Study of 2009 (HSLS:09).* [https://nces.ed.gov/surveys/hsls09/pdf/2011328_1.pdf]. Retrieved from https://nces.ed.gov/surveys/hsls09/usermanuals.asp

United States Census Bureau. (2009). *Median Household Income by State.* [h08.xslx]. Retrieved from https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

Child Trends Databank. (2015). *Parental education*. Retrieved From: https://www.childtrends.org/?indicators=parental-education