Ursinus College
Digital Commons @ Ursinus College

Business and Economics Honors Papers | Student Research

4-23-2015

Characteristics of STEM Success: A Survival Analysis Model of Factors Influencing Time to Graduation Among Undergraduate STEM Majors

Riley K. Acton
Ursinus College, riacton1@ursinus.edu

Adviser: Jennifer VanGilder
Second Adviser: Nicholas Scoville

Follow this and additional works at: http://digitalcommons.ursinus.edu/bus_econ_hon

Recommended Citation
Acton, Riley K., "Characteristics of STEM Success: A Survival Analysis Model of Factors Influencing Time to Graduation Among Undergraduate STEM Majors" (2015). *Business and Economics Honors Papers.* Paper 1.

This Paper is brought to you for free and open access by the Student Research at Digital Commons @ Ursinus College. It has been accepted for inclusion in Business and Economics Honors Papers by an authorized administrator of Digital Commons @ Ursinus College. For more information, please contact aprock@ursinus.edu.

Characteristics of STEM Success:
A Survival Analysis Model of Factors Influencing Time to Graduation among Undergraduate STEM Majors

Riley Acton

April 22, 2015

Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Mathematics and Business & Economics

Abstract

Producing more graduates in Science, Technology, Engineering, and Mathematics (STEM), as well as ensuring students complete college in a timely manner are both areas of national public policy interest. In order to improve these two outcomes, it is imperative to understand what factors lead undergraduate students to persist in, and ultimately graduate with STEM degrees. This paper uses data from the Beginning Postsecondary Students Longitudinal Study, provided by The National Center of Education Statistics, to model the time to baccalaureate degree among STEM majors using a Cox proportional hazard model.

I. Introduction

In late 2012, U.S. President Barack Obama announced a Cross-Agency Priority (CAP) goal of increasing the number of students that receive undergraduate degrees in science, technology, engineering and mathematics (STEM) disciplines by more than 30% over the next decade. If accomplished, this will send one million new STEM graduates into the labor force by 2020. The Obama administration recognizes that "science and innovation are key components of a strong American economy" and that "increasing opportunities for young Americans to gain STEM skills can both create jobs and enhance our national competitiveness."[1] The question, then, lies in how our nation can accomplish this goal.

The demand for college-educated workers in STEM fields has recently been very strong. Between 2000 and 2010, the number of STEM jobs in the U.S. grew at a rate of 7.9%, compared to a 2.6% growth rate in non-STEM occupations.[2] The Joint Economic Committee of the U.S. Congress expects this disparity to continue between the two sectors over the next ten years, projecting 17% growth in STEM occupations and only 14% for those not in STEM.[3]

This evidence of strong demand is further supported by the high wage premiums reported by STEM workers. Americans employed in STEM earn more than their counterparts in non-STEM occupations at every education level and the wage differential has been growing over the past two decades.[4] After controlling for various demographic factors, STEM workers earned 26% more than non-STEM workers in 2010, up from 18% in 1994.

Moreover, unemployment rates in STEM fields have been significantly lower than those in non-STEM fields, even during the most recent recession. The national STEM unemployment rate peaked at just 5.5% in 2009, while the unemployment rate for non-STEM occupations continued to climb to above 10% in 2010.[5] A study by Change the Equation, a non-profit organization dedicated to improving STEM education in the U.S., found that over the past three years, the number of online job postings in STEM occupations outnumbered the number of unemployed STEM workers by a

ratio of 1.9 to 1. That is, the demand for STEM workers was greater than the supply of those willing and able to fill the jobs. In contrast, the number of unemployed workers in non-STEM fields outnumbered the number of non-STEM online job postings by a factor of 3.6, indicating that the demand for workers in these occupations was significantly weaker than the supply.[6] Even more startling, at the height of the 2008 recession, over one third of manufacturers reported a shortage of scientists and engineers in their firm.[7]

The supply of college graduates prepared to work in STEM occupations, however, has not kept up with this increased demand. Over the past several decades, the U.S. has lost its footing as the world's leader in science and technology, particularly when it comes to the achievements of our nation's youth. In 2012, the U.S. ranked 17th in the world for 15-year olds' achievement in science according to the Program for International Student Assessment (PISA), which is administered annually by the Organization for Economic Cooperation and Development (OECD). The ranking for achievement in mathematics was even worse at 25th in the world. Furthermore, the U.S. is ranked 27th in the OECD for the percentage of all bachelor degree recipients that major in STEM, at just 15% of graduates.[8] This phenomenon is predominantly caused by the high attrition rates in STEM fields; it is estimated that less than 40% of college graduates who begin their college careers majoring in STEM will graduate with a STEM major.[9]

Moreover, the make-up of the STEM workforce has not kept pace with recent U.S. demographic shifts. While women have made great strides in the labor market over the past 50 years, now making up 48% of the aggregate workforce, they still only account for 23% of STEM workers. A similar story can be told for African-Americans, who make up only 6% of STEM workers despite having twice that share in the overall labor market. Additionally, Latinos are underrepresented in STEM compared to their proportions in the U.S. population, while Asians are drastically overrepresented.[10] A complete racial breakdown of American workers, both in STEM and in total, is provided below in Table 1.

Table 1: U.S. Labor Force Demographics

	White	African-American	Latino	Asian	Other
Proportion of workers in all occupations by race (%)	65	12	16	5	2
Proportion of workers in STEM occupations by race (%)	71	6	6	16	2

* Reproduced from (Carnevale, Smith, & Melton, 2011)

Given these labor market imperfections and the government's goal, it is vital to understand what factors influence a student's decision to major in, persist in, and ultimately graduate with a STEM degree. In particular, it is important for policy considerations to determine whether certain racial, ethnic, and gender groups are more at risk for dropping out of a STEM degree program. This paper attempts to understand students' progress through STEM programs by isolating the variables that influence a student's graduation in a STEM major.

III. Literature Review

Extensive research on college attrition, time-to-degree among undergraduate students, STEM major selection, and STEM graduation has been published across economic, education, and public policy journals. Some authors have investigated college graduation rates and time-to-degree in a historical context and others have attempted to determine the specific underlying factors that influence student outcomes within STEM fields. No currently published studies, however, have attempted to combine time-to-degree evaluation with research on STEM completion rates as is the goal of this paper.

Prior to evaluating the literature, it is important to understand what is meant by the term STEM, and which disciplines are included in this paper's analysis. Judith Ramaley, a former director of the National Science Foundation (NSF), is credited with first coining the acronym STEM in 2001,

although the concept was referred to by the NSF as SMET as early as 1993.[11] Today, the NSF uses a broad definition of STEM that includes selected social sciences such as psychology and economics. However, other government agencies, scholarship programs, and educational initiatives use a narrower definition that limits the STEM acronym to only the "hard sciences."[12] Further, research on attrition and completion rates in STEM have used differing definitions, making it difficult to reach definitive conclusions from the vast body of literature available. For this paper, we use the National Center for Education Statistics definition of STEM which includes: mathematics, physical sciences, biological/life sciences, engineering, and computer/information sciences. Social sciences and health sciences are not included.

Returning to the literature, we begin by reviewing historical time to graduation rates in the United States. In comparing national longitudinal data from the high school classes of 1972 and 1992, Bound, Lovenheim, and Turner (2010) find that the average time to complete a bachelor's degree increased from 4.69 to 4.97 years over the two decades. Their research suggests that declines in per-student spending and increases in student-faculty ratios at less-selective public institutions, as well as the rise of the non-traditional working student, are reasons for the increase. They find that, on average, students were equally prepared for college-level work in 1972 and 1992 but do not investigate micro-level factors – such as race, ethnicity, socioeconomic status, and high school curriculum – that may influence college completion.

However, there is an extensive body of research on what contributes to students graduating from college. Adelman (1999) reports that a student's "academic resources" – his or her high school curriculum, test scores, and class rank – matter far more than his or her socioeconomic status in the probability of completing a bachelor's degree. He also finds that finishing a math course beyond Algebra II during high school more than doubles the chance that a student who enters college will graduate. Furthermore, his work shows that delaying college entrance negatively impacts the probability of graduation, but attending multiple schools has no impact on the outcome. Ahlburg

and McCall (2002) confirm many of Adelman's findings, noting that higher ability students are more likely to complete college and that delayed entrance leads to a higher probability of dropout. Zhu (2004) has similar findings when looking at undergraduate students at SUNY College at Brockport. Her analysis shows that students with higher high school GPAs and SAT scores, along with female students, are significantly more likely to graduate college in four years. In addition, students who took out loans and those who worked at least 15 hours a week are more likely to graduate "on time" (i.e. in four years).

In order to study graduation and time-to-degree specifically within STEM fields, it is important to understand why students initially choose to study STEM. Crisp, Nora, and Taggart (2009) find demographic differences between students who choose to major in STEM and non-STEM fields, noting that females are less likely to declare a STEM major than males while Hispanic and Asian students are more likely to do so than White students. The authors also indicate that early exposure to, attitude towards, and achievement in math and science are important factors in a student's declaration of a STEM major. Crisp et al. find that a student's score on the math section of the SAT is a significant predictor of majoring in STEM. Wang (2013) confirms the author's findings. He further reports that students who perceive that their high school math and science classes have adequately prepared them for college-level work, as well as those who have a positive attitude toward math, are more likely to major in a STEM field.

Factors influencing STEM attrition and graduation are similar. In a report prepared for the National Center for Education Statistics (NCES), Chen (2013) breaks down a nationally representative sample of college students into three distinct groups: those who declare a STEM major and later switch majors, those who declare a STEM major and later drop out of college, and those who persist towards a STEM degree. He finds that women are more likely to switch majors while men are more likely to drop out of school completely and that Black students are the most likely to leave STEM fields while Asian students are the least likely. Additionally, those who persist

in STEM take more STEM classes in the first year than STEM "leavers" (on average, 18 credits vs. 11 credits) and STEM "persisters" are more likely to have taken calculus or more advanced math in their first year of college. Chen confirms that high school academic preparation, particularly in math, influences a student's probability of completing a STEM degree, reporting that those who have a higher high school GPA and have taken a higher level of math before college are less likely to drop out of a STEM program.

Chen's (2013) research summarizes and builds upon what many other authors have found. Several studies have shown that high school academic preparation is the largest single factor influencing a student's attainment of a STEM degree. Crisp et al. (2009) also write that the likelihood of earning a STEM degree is uniquely associated with a student's gender, ethnicity, SAT math score, high school percentile, first semester college GPA, and enrollment in math and science classes. Rask (2010) bolsters this claim by finding that higher grades in preliminary STEM courses increase persistence in a STEM major among students at a liberal arts college. Whalen and Shelley (2010) have similar findings to Chen, Crisp et. al, and Rask, but add that students who live on-campus are more likely to graduate in STEM and that out-of-state students are less likely than in-state students.

IV. Data

Data for this paper comes from the Department of Education's National Center for Education Statistics. The 2004/09 Beginning Postsecondary Students Longitudinal Study (BPS) follows a cohort of over 18,000 students entering college for the first time during the 2003-2004 school year, and asks several questions related students' pathways into and out of STEM fields. The study draws its initial student sample from the National Postsecondary Student Aid Study (NPSAS), a nationally representative dataset that aims to explain how students of varying backgrounds finance their

educations. In its final form, the BPS dataset provides hundreds of variables on student demographics, academic preparation, and undergraduate experiences over a six year time period, as well as degree completion and time to degree data.

In order to investigate our population of interest – first-time bachelor degree seeking students majoring in a STEM field – it is necessary to understand the construction of the BPS dataset and extract a sample that fits our paper's goal. Students in the BPS:04/09 cohort were initially identified from 1,630 institutions meeting NPSAS criteria, such as location in the U.S. or Puerto Rico and the exclusion of U.S. service academies. A total of 44,670 college students were contacted for the study, with 23,090 agreeing to participate. These students were interviewed during their first year of postsecondary study, and contacted again at three and six year intervals. Transcript data was also collected for these students when available. In total, 18,640 students completed all three interview phases and remained in the study dataset in 2009.

Of these 18,640 students, 4,870 indicated that they were seeking a bachelor's degree rather than pursuing an associate's degree or a non-degree program. We further identify 1,126 **STEM beginners** based on the number of STEM credits a student completed in their first two years of college.[13] Students who completed more than 16 credits in STEM classes are included in our dataset and analyzed in the paper.

Descriptive statistics on the dataset reveal that this group of students has a make-up similar to the U.S. college population at large. Below, Table 1 shows that there are slightly more females than males in the dataset. This is expected given the increase in college attendance by women over the past half century.[14]

Table 1: Frequency by Gender

Gender	Frequency	Percent
Male	486	43.16%
Female	640	56.84%

While the majority of students in the sample identify as Caucasian, there is a fairly even distribution of students identifying as Black, Hispanic, Asian, and other minorities. Table 2 provides the frequencies of these groups.

Table 2: Frequency by Race

Race	Frequency	Percent
White	845	75.04%
Black	84	7.46%
Hispanic	75	6.66%
Asian	60	5.33%
Other	62	5.52%

Academically, the students in our sample are representative of all college students in the U.S. The average SAT score of our dataset is 1091, which is just above the national average of 1000.[15] Most took three or four years of math and science in high school, and enrolled in college immediately following graduation from high school. Further, there is a wide range of institution enrollments included in the sample and students come from households at every income level. Table 3 provides summary statistics on these variables, among others.

Table 3: Summary Statistics

Variable	Mean	SD	Minimum	Maximum
Age of first college enrollment	18.37	0.60	17	22
Years delayed prior to enrollment	0.06	0.38	0	5
Number of institutions attended	1.39	0.68	1	5
SAT composite score	1091.47	187.84	500	1570
SAT math score	547.95	106.43	220	800
SAT verbal score	543.53	101.97	210	800
Years of math in high school	3.75	0.58	0	4
Years of science in high school	3.44	0.73	0	4
Annual household income	76,530.53	58,449.65	0	483,748
First institution enrollment	13,829.18	11,883.57	163	47,952
Year 1 STEM credits	14.13	6.70	0	49
Year 2 STEM credits	14.07	7.83	0	45
Year 3 STEM credits	10.11	9.71	0	50
Year 4 STEM credits	7.85	9.33	0	42

The only area where the STEM Beginners dataset deviates from the U.S. college population at large is in the students' parents' educational attainment levels. A large percentage of the sample has parents who completed a bachelor's degree, and an even larger percentage, a graduate of professional degree. While this is surprising given the relatively low share of the U.S. adult population with such degrees, it may be indicative of those students who choose to pursue STEM degrees and who agree to participate in a higher education research survey. However, as with all dataset peculiarities, this should be kept in mind when interpreting the final results. The breakdown of parental education levels of our dataset is provided below in Table 4.

Table 4: Frequency by Parent's Education

Parent's Education	Frequency	Percent
Do not know	15	1.33%
Did not complete high school	25	2.13%
High school diploma or equivalent	147	13.06%
Some college or vocational training	238	21.14%
Bachelor's degree	334	29.66%
Graduate or professional degree	368	32.68%

V. Methods

In order to effectively analyze this dataset, we implement techniques from the field of survival analysis, a statistical method originally designed for the interpretation of biological and health data. Survival regression models relate the time until an event (e.g. graduation from college) occurs to some set of covariates (e.g. sex, race, parent's education level, SAT score, number of STEM credits earned in the first year of college, etc.). A brief overview of the two key functions used in survival analysis is provided below.

The *survival function* is the probability that an individual "survives" in the study to time x. This may also be interpreted as the probability than an individual experiences the event of interest after time x. The function is typically expressed as:

$$S(x) = P(X > x).$$

For the purposes of this paper, the survival function lacks a strong interpretive meaning. Instead, we focus on the *hazard function*, or *hazard rate*. The hazard function is the chance that an individual of "age" x experiences the event in the next instant of time. The function is defined as:

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X \leq x + \Delta x \mid X \geq x]}{\Delta x}.$$

The hazard function of our dataset can be defined as the chance that an individual who has been enrolled in college for *x* month graduates with a STEM degree in the following month. For example, if *x=45*, then *h(45)* is the probability that the individual of interest will graduate with a STEM degree in the 46^{th} month of college enrollment.

We use this hazard function to derive a regression model and analyze the data. The general multiplicative hazard rate model is expressed as:

$$h(x|\boldsymbol{z}) = h_0(x)c(\boldsymbol{\beta}^t\boldsymbol{z}),$$

where $h(x|\boldsymbol{z})$ is the conditional hazard rate of an individual with covariate vector $\boldsymbol{z}$, $h_0(x)$ is a baseline hazard rate, and $c(\boldsymbol{\beta}^t\boldsymbol{z})$ is a non-negative function of the covariates. In particular, we use the Cox proportional hazard model which specifies $c(\boldsymbol{\beta}^t\boldsymbol{z}) = \exp(\boldsymbol{\beta}^t\boldsymbol{z})$. A key feature of this specification is that, when all the covariates are fixed at time 0, the hazard rates of two individuals with distinct sets of covariates are proportional. Observe that:

$$\frac{h(x|\boldsymbol{z}_1)}{h(x|\boldsymbol{z}_2)} = \frac{h_0(x)c(\boldsymbol{\beta}^t\boldsymbol{z}_1)}{h_0(x)c(\boldsymbol{\beta}^t\boldsymbol{z}_2)} = \frac{c(\boldsymbol{\beta}^t\boldsymbol{z}_1)}{c(\boldsymbol{\beta}^t\boldsymbol{z}_2)}$$

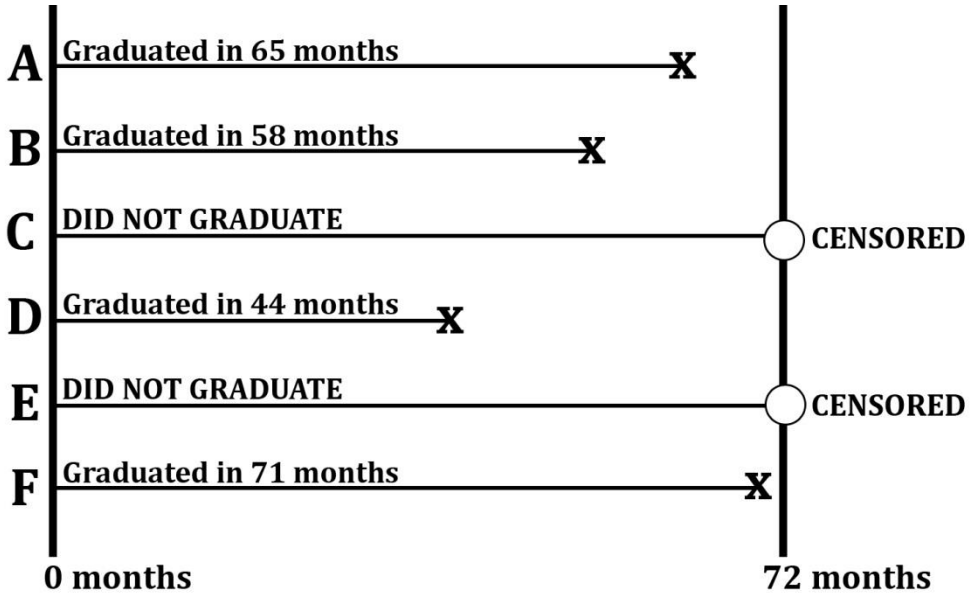
which is constant independent of time.

V.1 Censoring

An important feature of survival analysis is the method's ability to handle observations that have yet to experience the event of interest. In survival analysis, these types of observations that lack a time of entry and/or exit into a given study are said to be

censored. In particular, those that lack a time of exit are called *right censored* observations and those that lack a time of entry are called *left censored* observations. Observations that are both left and right censored are said to be *interval censored*. In our STEM beginner's dataset, only right censoring occurs as all participants enter the study at the same time: the beginning of the 2003-2004 academic year. All participants are followed for the full 6 year duration of the study, so no left or interval censoring occurs. Students who had not graduated in STEM at the end of the study are censored at 72 months. A schematic of which observations are censored is provided below in Figure 1.

Figure 1: Censored Data



In survival data, all observations contain information on (1) the subject's time to event, and (2) whether or not the observation is censored. These two pieces of information are captured in the regression model's dependent variable, and as such, the survival analysis method allows for correct incorporation of both censored and uncensored observations into its parameter estimates. This approach differs significantly from a linear

regression model, where observations that have not experienced the event are not included in the estimation of parameters, or a logistic regression model, where observations must be grouped together and time to event data is lost, and provides a much more accurate and robust interpretation of the data.

This model is implemented using SAS 9.2. The regression results and an interpretation of the coefficients are provided in the following sections.

VI. Results

Seven variables, designed to collectively capture a student's demographic information, academic preparation, and institutional characteristics, are selected from the dataset and included in the proportional hazards regression model. Descriptive information on these variables is provided in Table 5.

Table 5: Model Variable Information

Variable	Type	Description
Gender	Categorical	1 if student is male; 2 if student is female
Asian	Categorical	1 if student is Asain; 0 otherwise
Poverty	Categorical	1 if student's household income is below the 2004 federal poverty line for a family of 4; 0 otherwise
SAT	Continuous	Student's score on the SAT test, or ACT equivalent score
HSCalculus	Categorical	1 if student took calculus in high school; 0 otherwise
InstPrivate	Categorical	1 if the first institution a student attended is private; 0 otherwise
NumInst	Continuous	Number of institutions the student attended through 2009

Additionally, a theoretical framework of the model using these variables is provided below:

$$h(x) = h_0(x) * \exp(\beta_1 \text{Gender} + \beta_2 \text{Asian} + \beta_3 \text{Poverty} + \beta_4 \text{SAT} + \beta_5 \text{HSCalculus} + \beta_6 \text{InstPrivate} + \beta_7 \text{NumInst})$$

Based on the results of previous literature, we hypothesize that $\beta_2, \beta_4, \beta_5, \beta_6 > 0$ and $\beta_1, \beta_3, \beta_7 < 0$. Several authors have shown that (1) female students are less likely to declare and persist in a STEM major than male students, (2) Asian students have a higher propensity to

graduate in one than students of other ethnic backgrounds, and (3) students from households of poverty face a challenging road to graduate from college (Crisp et al., Chen). Thus, we would expect the coefficients on *Gender* and *Poverty* to be negative, and that on *Asian* to be positive. Adelman (1999), Zhu (2004), and Ahlburg and McCall (2002) cite the importance of a student's high school academic background on his or her chances of completing college. They find that a student's SAT score and the level of math completed during high school both have a positive correlation with graduating college, particularly in a STEM field. This implies that the expected coefficients on both *SAT* and *HSCalculus* should both be positive. Further, we expect that the coefficient on *InstPrivate* will be positive given that private institutions are typically able to provide more resources per student, and in turn have higher graduation rates, than public institutions. We also hypothesize that the coefficient on *NumInst* will be negative, as it is expected that the time it takes a student to graduate from college will increase as the number of institutions attended increases.

The results of the model, with significance level included, are outlined below in Table 6.

Table 6: Proportional Hazard Model Results

Variable	Coefficient	Standard Error	Chi-Square	Pr > Chi-Square	Significance
Gender	-0.08167	0.15540	0.2762	0.5992	
Asian	0.55336	0.26599	4.3281	0.0375	**
Poverty	-0.58155	0.30443	3.6491	0.0561	*
SAT	0.00260	0.00049	27.6560	<0.0001	***
HSCalculus	1.01957	0.17645	33.3873	<0.0001	***
InstPrivate	0.29895	0.15547	3.6976	0.0545	*
NumInst	-0.67875	0.15792	18.4741	<0.0001	***

* Significant at 10% level ** Significant at 5% level *** Significant at 1% level

Due to the exponential form of the hazard rate model, the coefficient estimates shown in this table cannot be easily interpreted as given. Instead, we refer to the hazard ratios for each variable. These ratios can be interpreted in two useful ways. First, for categorical variables, they

may be seen as an equivalent to the likelihood ratios that are given in traditional logistic regression models. However, instead of the term "likelihood" we use the term "hazard" to refer to the chance that an individual experiences the event of interest. Second, for both categorical and continuous variables, the hazard ratios may be easily transformed to be seen as a percentage change in the hazard corresponding with a one-point increase in the given variable. We simply subtract 1 from the hazard ratio and multiply by 100 to obtain this useful figure that may be interpreted similarly to coefficients in traditional log-log or log-linear regression models. The hazard ratio for each variable coefficient is given in Table 7 and an interpretation of the results follows.

Table 7: Hazard Ratios

Variable	Hazard Ratio	Percentage
Gender	0.922	-7.8%
Asian	1.739	73.9%
Poverty	0.559	44.1%
SAT	1.003	0.3%
HSCalculus	2.772	177.2%
InstPrivate	1.348	34.8%
NumInst	0.507	-49.3%

A student's demographic background appears to have a substantial bearing on his or her success in STEM. However, we do not interpret the hazard ratio for *Gender* as the regression model revealed that it is not a significant predictor of a student's time to graduation. Asian students have a hazard of graduation in STEM 1.739 times larger than non-Asian students. Equivalently, they can be said to have a 73.9% greater hazard of graduation than their non-Asian counterparts. Reversely, students from a household with income below the federal poverty line have a hazard of graduation 0.559, or 44.1%, lower hazard of graduation as those from a household with income above the federal poverty line.

Unsurprisingly, a student's academic background prior to entering college plays an important role in their collegiate success. For every one-point increase on the SAT examination, a student's hazard of graduating increases by 0.3%. While this may appear like a minimal increase, it

is important to note that the SAT is scored on an 800-point scale in 10-point increments. Thus, a 100-point score increase could increase a student's hazard by over 3%. Additionally, students who completed a calculus course in high school have a hazard of graduating that is an astounding 2.772 times, or 177.2%, higher as likely to graduate in STEM as those who did.

The institutions a student attends throughout his or her college career may also have a bearing on his or her graduation outcome. Students who attended a private institution at the beginning of their college careers have a hazard 1.348 times as large as those who began their studies at public institutions. Moreover, a one-unit increase in the number of institutions attend decreases a student's hazard of graduating by 49.3%.

VII. Conclusion

In many regards, the results of this paper confirm the findings of previous research on STEM education. We find that a student's academic preparation, measured here by his or her SAT score and calculus class completion, is arguably the most important variable in determining whether or not he or she will graduate with a STEM degree. Additionally, demographic and institutional factors also appear to play a role in determining educational outcomes. Students from households below the poverty line are significantly less likely to graduate in STEM while those of Asian heritage are significantly more likely to do so. Further, students at private institutions experience more favorable outcomes, and those who transfer institutions multiple times are significantly less likely to earn a STEM degree.

However, we find that a student's gender plays no significant role in his or her obtainment of a STEM degree. This finding comes in stark contrast to a plethora of studies which show that women are less likely to graduate in STEM, and ultimately, pursue a scientific career. There are two likely reasons for this disparity. First, it is possible that by the time a student reaches college, his or

her propensity to pursue STEM and succeed in the field has already been established. That is, student experiences in primary and secondary school are of more importance than those that occur during the college years. If this is indeed the case, we should focus more resources on developing STEM interest and aptitude in young girls rather than initiating higher education programs aimed at increasing the number of women in STEM. Second, it is important to note the BPS:04/09 dataset provides information on a much more recent cohort of college students than many previous studies on STEM education. Thus, it could be that women no longer face as many challenges in studying STEM fields as they once did. This would be a remarkable conclusion, but certainly needs further research to claim.

VIII. Future Work

Without question, additional research is needed in order to obtain a clear picture of STEM education in the United States and ensure a healthy STEM workforce in the future. First and foremost, it is necessary to determine whether or not the sample of students analyzed in this paper is indeed the population group of interest. Given the wide variety of institutional protocols surrounding major declaration and course enrollment, as well as the delay in response times to major and degree changes, it is extremely difficult to determine if and when a student begins working on a STEM degree. While some students may follow a clear, linear trajectory in a specific area of study, the vast majority are likely to change their majors along the way or take classes outside of their declared area of interest. Determining how to statistically asses these situations, and creating a standard method of identifying STEM majors, are of vital importance in furthering STEM education research.

It would also be beneficial to work with a dataset that includes additional variables on the student's academic department and family characteristics. For example, knowing the gender

composition of the faculty could help analyze the claim that the dearth of women pursuing STEM degrees and careers is due to a lack of female role models in such fields. Additionally, knowing the careers and detailed educational backgrounds of a student's parents may also shed light on to what factors influence an individual to complete a STEM degree. Overall, the more complete and more specific that we can make the model, the more information we will be able to glean about STEM outcomes and the future of our nation's scientific and research workforces.

---

[1] (Feder, 2012)
[2] (Langdon, McKittrick, Beede, Khan, & Doms, 2011)
[3] (U.S. Congress Joint Economic Committee, 2012)
[4] (Langdon, McKittrick, Beede, Khan, & Doms, 2011)
[5] (U.S. Congress Joint Economic Committee, 2012)
[6] (Change the Equation, 2012)
[7] (U.S. Congress Joint Economic Committee, 2012)
[8] Ibid
[9] (Carnevale, Smith, & Melton, 2011)
[10] Ibid
[11] (Donahoe, 2013)
[12] (Gonzalez & Kuenzi, 2012)
[13] Given the significant amount of time between interviews in the BPS study, it is difficult to determine when, and in what subjects, students declared majors. Additionally, a student's reported major does not always correspond with the classes taken. Thus, it is more effective to identify a student's entrance in to a STEM field by the number of STEM classes taken within the first two years of undergraduate study.
[14] (U.S. Department of Education, 2013)
[15] (Average SAT Scores, 2015)

Appendix A: Bibliography

Adelman, C. (1999). *Answers in the Toolbox: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment.* Washington, D.C.: National Institute on Postsecondary Education, Libraries, and Lifelong Learning.

Ahlburg, D. A., & McCall, B. P. (2002, April). Time to Dropout from College: A Hazard Model with Endogenous Waiting.

Average SAT Scores. (2015). Retrieved February 17, 2015, from The College Board: http://professionals.collegeboard.com/testing/sat-reasoning/scores/averages

Carnevale, A. P., Smith, N., & Melton, M. (2011). *STEM: Science Technology, Engineering, Mathematics.* Washington, DC: Georgetown University Center on Education and the Workforce.

Change the Equation. (2012). *STEM: Help Wanted.*

Chen, X. (2013). *STEM Attrition: College Students' Paths Into and Out of STEM Fields (NCES 2014-001).* Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Crisp, G., Nora, A., & Taggart, A. (2009, December). Student Characteristics, Pre-College, College, and Environmental Factors as Predictors of Majoring in and Earning a STEM Degree: An Analysis of Students Attending a Hispanic Serving Institution. *American Educational Research Journal, 46*(4), 924-942.

Donahoe, D. (2013, December). *Today's Engineer*. Retrieved October 30, 2014, from The Definition of STEM?: http://www.todaysengineer.org/2013/Dec/STEM-definition.asp

Feder, M. (2012, December 18). *One Decade, One Million More STEM Graduates*. Retrieved June 30, 2014, from The White House.

Gonzalez, H. B., & Kuenzi, J. J. (2012). *Science, Technology, Engineering, and Mathematics (STEM) Education: A Primer.* Congressional Research Service.

Kokkelenberg, E. C., & Esha, S. (2010, December). Who Succeeds in STEM Studies? An Analysis of Binghamton University Undergraduate Students. *Economics of Education Review, 29*(6), 935-946.

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good Jobs Now and for the Future.* U.S. Department of Commerce, Economics and Statistics Administration.

NLSF Data Overview. (n.d.). Retrieved July 16, 2014, from National Longitudinal Survey of Freshmen.

Rask, K. (2010, December). Attrition in STEM Fields at a Liberal Arts College: The Importance of Grades and Pre-Collegiate Preferences. *Economics of Education Review, 29*(6), 892-900.

Ronco, S. L. (1994, May). Meandering Ways: Studying Student Stopout with Survival Analysis. AIR 1994 Annual Forum Paper.

Snyder, T. D., & Dillow, S. A. (2013). *Digest of Education Statistics 2012.* Washington, D.C.: National Centerfor Education Statistics, Institute of Education Sciences, U.S. Department of Education.

U.S. Congress Joint Economic Committee. (2012). *STEM Education: Preparing for the Jobs of the Future.* U.S. Congress.

U.S. Department of Education, N. C. (2013). *Digest of Education Statistics, 2012 (NCES 2014-015).*

Wang, X. (2013, October). Why Students Choose STEM Majors: Motivation, High School Learning, and Postsecondary Context of Support. *American Educational Research Journal, 50*(5), 1081-1121.

Whalen, D. F., & Mack, C. S. (2010). Academic Success for STEM and Non-STEM Majors. *Journal of STEM Education, 11*(1).

Zhu, L. (2004). *Exploring the Determinants of Time-to-Degree in Public 4-Year Colleges.* Brockport, NY: SUNY College at Brockport.

Appendix B: SAS Code

```
LIBNAME F09 'C:\ECBW\F09';

DATA X1;INFILE 'C:\ECBW\F09\DATA\F9DERIV.DAT' LRECL=1024 PAD;INPUT ID 1-6
  AGE 745-746 DELAYENR 757-758 GENDER 767-768 RACE 769-770
  HOMEDIST 963-968 LOCALRES 975-976 /  ATBAM6Y 29-30
  ATBAEN6Y 31-32 ATTYPE6Y 91-92 ATHTY6Y 99-100 ENINUM6Y 311-312
  ACAD04A 469-470 ACAD04C 471-472 FREQ04A 483-484
  HIGHLVEX 537-538 TESATDER 555-558 TESATMDE 563-566
  TESATVDE 567-570 CRDHS04 571-572 CRDCL04 573-574
  HCGPAREP 583-584 HSTYPE 589-590 HCMATH 591-592
  HCYSMATH 599-600 HCYSSCIE 601-602 MAJ06DEC 647-648
  MAJ09DEC 649-650 MAJ04A 655-656 MAJ06A 657-658
  MAJ09A 661-662 PROUT6 721-722 UGDEG 815-816 UGDEG06 817-818
  UGDEG09 819-820 FALLDGFT 821-822 NPFST04 929-930
   /  CAGI 643-648 PCTALL 711-714 /  SAMESTAT 213-214
  ENRLSIZE 219-223 SELECTV2 224-225 CC2000A 300-301
  PDADED 374-375 PMOMED 376-377 PAREDUC 380-381  / ;

LABEL
  AGE='Age first year enrolled'
  DELAYENR='Delayed enrollment into PSE: Number of years 2003-04'
  GENDER='Gender'
  RACE='Race/ethnicity'
  HOMEDIST='Distance from first institution 2003-04'
  LOCALRES='Housing 2003-04'
  ATBAM6Y='First bachelor^s degree months elapsed through 2009'
  ATBAEN6Y='First bachelor^s degree months enrolled through 2009'
  ATTYPE6Y='Degree types attained through 2009'
  ATHTY6Y='Highest degree attained anywhere through 2009'
  ENINUM6Y='Number of institutions attended through 2009'
  ACAD04A='Academic 2004: graduate student instructors'
  ACAD04C='Academic 2004: large classes'
  FREQ04A='Frequency 2004: Faculty informal meeting'
  HIGHLVEX='Highest degree ever expected 2003-04'
  TESATDER='Admissions test scores (ACT or SAT)'
  TESATMDE='Derived SAT math score'
  TESATVDE='Derived SAT verbal score'
  CRDHS04='Earned any college level credits in high school'
  CRDCL04='Earned credits for courses at a college while in high school'
  HCGPAREP='High school grade point average (GPA)'
  HSTYPE='High school type attended'
  HCMATH='Highest level of high school mathematics'
  HCYSMATH='Years of mathematics in high school'
  HCYSSCIE='Years of science in high school'
  MAJ06DEC='Major declared as of 2006'
  MAJ09DEC='Major declared as of 2009'
  MAJ04A='Major when first enrolled in 2003-04 (comparable to 2006, 2009)'
  MAJ06A='Major when last enrolled 2006'
  MAJ09A='Major when last enrolled 2009'
  PROUT6='Cumulative persistence and attainment anywhere 2008-09'
  UGDEG='Degree program during 2003-04'
  UGDEG06='Degree program when last enrolled 2006'
  UGDEG09='Degree program when last enrolled 2009'
  FALLDGFT='Fall 2003 beginners'
```

```
  NPFST04='First choice was NPSAS school 2004'
  CAGI='Adjusted Gross Income (AGI) 2003-04'
  PCTALL='Income percentile rank for all students 2003-04'
  SAMESTAT='Attend institution in state of legal residence 2003-04'
  ENRLSIZE='Enrollment size 2003-04'
  SELECTV2='First institution selectivity 2003-04'
  CC2000A='Carnegie code (2000) with control 2003-04'
  PDADED='Father^s highest education level 2003-04'
  PMOMED='Mother^s highest education level 2003-04'
  PAREDUC='Parent^s highest level of education';

PROC FORMAT;
 VALUE DELAYENF        0='{zero}'
                       -3='{Skipped}';

 VALUE GENDERF         1='Male'
                       2='Female';

 VALUE RACEF           1='White'
                       2='Black or African American'
                       3='Hispanic or Latino'
                       4='Asian'
                       5='American Indian or Alaska Native'
                       6='Native Hawaiian / other Pacific Islander'
                       7='Other'
                       8='More than one race';

 VALUE HOMEDISF        0='{zero}';

 VALUE LOCALREF        0='Attended more than one institution'
                       1='On campus'
                       2='Off campus'
                       3='Living with parents';

 VALUE ATBAM6YF        0='{zero}';
 VALUE ATBAEN6F        0='{zero}';

 VALUE ATTYPE6F        0='No degree'
                       1='Certificate only'
                       2='Associate^s degree only'
                       3='Certificate and associate^s degree'
                       4='Bachelor^s degree only'
                       5='Certificate and bachelor^s degree'
                       6='Associate^s degree and bachelor^s degree'
                       7='Certificate, associate^s degree and bachelor^s
                        deg';

 VALUE ATHTY6YF        0='No degree'
                       1='Certificate'
                       2='Associate^s degree'
                       3='Bachelor^s degree';

 VALUE ACAD04AF        0='Never'
                       1='Sometimes'
                       2='Often'
                       -3='{Skipped}';
```

```
VALUE ACAD04CF        0='Never'
                      1='Sometimes'
                      2='Often'
                      -3='{Skipped}';

VALUE FREQ04AF        0='Never'
                      1='Sometimes'
                      2='Often'
                      -3='{Skipped}';

VALUE HIGHLVEF        1='No degree or certificate'
                      2='Certificate'
                      3='Associate^s degree'
                      4='Bachelor^s degree'
                      5='Post-BA or post-master certificate'
                      6='Master^s degree'
                      7='Doctoral degree'
                      8='First-professional degree';

VALUE TESATDEF        -3='{Skipped}';
VALUE TESATMDF        -3='{Skipped}';
VALUE TESATVDF        -3='{Skipped}';

VALUE CRDHS04F        0='No'
                      1='Yes'
                      -3='{Skipped}';

VALUE CRDCL04F        0='No'
                      1='Yes'
                      -3='{Skipped}';

VALUE HCGPAREF        1='0.5-0.9 (D- to D)'
                      2='1.0-1.4 (D to C-)'
                      3='1.5-1.9 (C- to C)'
                      4='2.0-2.4 (C to B-)'
                      5='2.5-2.9 (B- to B)'
                      6='3.0-3.4 (B to A-)'
                      7='3.5-4.0 (A- to A)'
                      -3='{Skipped}';

VALUE HSTYPEF         0='No high school diploma or certificate'
                      1='Public'
                      2='Private'
                      3='Attended a foreign high school';

VALUE HCMATHF         0='None of these'
                      1='Algebra 2'
                      2='Trigonometry/Algebra II'
                      3='Pre-calculus'
                      4='Calculus'
                      -3='{Skipped}';

VALUE HCYSMATF        0='Did not take or took half-year'
                      1='One year to one and a half years'
                      2='Two years to two and a half years'
                      3='Three years to three and a half years'
                      4='Four or more years'
```

```
                    -3='{Skipped}';

VALUE HCYSSCIF      0='Did not take or took half-year'
                    1='One year to one and a half years'
                    2='Two years to two and a half years'
                    3='Three years to three and a half years'
                    4='Four or more years'
                    -3='{Skipped}';

VALUE MAJ06DEF      0='Not in a degree program'
                    1='Yes, I have declared a major'
                    2='Yes, I have declared a double major'
                    3='No, I have not declared a major yet'
                    -3='{Skipped}';

VALUE MAJ09DEF      0='Not in a degree program'
                    1='Yes, I have declared a major'
                    2='Yes, I have declared a double major'
                    3='No, I have not declared a major yet'
                    -3='{Skipped}';

VALUE MAJ04AF       0='Undeclared or not in a degree program'
                    1='Agriculture/natural resources/related'
                    2='Architecture and related services'
                    3='Area/ethnic/cultural/gender/grp studies'
                    4='Visual and performing arts'
                    5='Biological and biomedical sciences'
                    6='Business/management/marketing/related'
                    7='Communication/journalism/related tech'
                    8='Computer/information science/support'
                    9='Construction trades'
                    10='Education'
                    11='Engineering'
                    12='English language and literature/letters'
                    13='Family/consumer sciences/human sciences'
                    14='Foreign languages/literature/linguistics'
                    15='Health professions and related programs'
                    16='Legal professions and studies'
                    18='Mathematics and statistics'
                    19='Mechanic/repair technologies/technicians'
                    20='Multi/interdisciplinary studies'
                    21='Parks/recreation/leisure/fitness studies'
                    22='Precision production'
                    23='Personal and culinary services'
                    24='Philosophy/theology/religious studies'
                    25='Physical sciences'
                    26='Psychology'
                    27='Public administration/social service'
                    28='Science technologies/technicians'
                    29='Homeland security/law enforce/protective'
                    30='Social sciences/history'
                    31='Transportation and materials moving'
                    33='Liberal arts/sci/gen studies/humanities'
                    34='Engineering technologies/related fields';

VALUE MAJ06AF       0='Undeclared or not in a degree program'
                    1='Agriculture/natural resources/related'
```

```
                    2='Architecture and related services'
                    3='Area/ethnic/cultural/gender/grp studies'
                    4='Visual and performing arts'
                    5='Biological and biomedical sciences'
                    6='Business/management/marketing/related'
                    7='Communication/journalism/related tech'
                    8='Computer/information science/support'
                    9='Construction trades'
                    10='Education'
                    11='Engineering'
                    12='English language and literature/letters'
                    13='Family/consumer sciences/human sciences'
                    14='Foreign languages/literature/linguistics'
                    15='Health professions and related programs'
                    16='Legal professions and studies'
                    18='Mathematics and statistics'
                    19='Mechanic/repair technologies/technicians'
                    20='Multi/interdisciplinary studies'
                    21='Parks/recreation/leisure/fitness studies'
                    22='Precision production'
                    23='Personal and culinary services'
                    24='Philosophy/theology/religious studies'
                    25='Physical sciences'
                    26='Psychology'
                    27='Public administration/social service'
                    28='Science technologies/technicians'
                    29='Homeland security/law enforce/protective'
                    30='Social sciences/history'
                    31='Transportation and materials moving'
                    32='Other'
                    33='Liberal arts/sci/gen studies/humanities'
                    34='Engineering technologies/related fields'
                    -3='{Skipped}';

VALUE MAJ09AF       0='Undeclared or not in a degree program'
                    1='Agriculture/natural resources/related'
                    2='Architecture and related services'
                    3='Area/ethnic/cultural/gender/grp studies'
                    4='Visual and performing arts'
                    5='Biological and biomedical sciences'
                    6='Business/management/marketing/related'
                    7='Communication/journalism/related tech'
                    8='Computer/information science/support'
                    9='Construction trades'
                    10='Education'
                    11='Engineering'
                    12='English language and literature/letters'
                    13='Family/consumer sciences/human sciences'
                    14='Foreign languages/literature/linguistics'
                    15='Health professions and related programs'
                    16='Legal professions and studies'
                    18='Mathematics and statistics'
                    19='Mechanic/repair technologies/technicians'
                    20='Multi/interdisciplinary studies'
                    21='Parks/recreation/leisure/fitness studies'
                    22='Precision production'
                    23='Personal and culinary services'
```

```
                    24='Philosophy/theology/religious studies'
                    25='Physical sciences'
                    26='Psychology'
                    27='Public administration/social service'
                    28='Science technologies/technicians'
                    29='Homeland security/law enforce/protective'
                    30='Social sciences/history'
                    31='Transportation and materials moving'
                    33='Liberal arts/sci/gen studies/humanities'
                    34='Engineering technologies/related fields'
                    -3='{Skipped}';

VALUE PROUT6F       1='Attained bachelor^s degree'
                    2='Attained associate^s degree'
                    3='Attained certificate'
                    4='No degree, still enrolled'
                    6='No degree, left without return';

VALUE UGDEGF        1='Certificate'
                    2='Associate^s degree'
                    3='Bachelor^s degree'
                    4='Not in a degree program or others';

VALUE UGDEG06F      1='Undergraduate certificate/diploma'
                    2='Associate^s degree'
                    3='Bachelor^s degree (4 year)'
                    4='Bachelor^s degree (5 year)'
                    5='First-professional/graduate'
                    6='Other undergraduate'
                    -3='{Skipped}';

VALUE UGDEG09F      1='Undergraduate certificate/diploma'
                    2='Associate^s degree'
                    3='Bachelor^s degree (4 year)'
                    4='Bachelor^s degree (5 year)'
                    6='Other undergraduate'
                    -3='{Skipped}';

VALUE FALLDGFF      1='Certificate, full time fall'
                    2='Associate^s, full time fall'
                    3='Bachelor^s, full time fall'
                    4='No degree, full time fall'
                    5='Certificate, not full time fall'
                    6='Associate^s, not full time fall'
                    7='Bachelor^s, not full time fall'
                    8='No degree, not full time fall';

VALUE NPFST04F      0='No'
                    1='Yes';

VALUE CAGIF         0='{zero}';

VALUE SAMESTAF      1='Yes'
                    2='No'
                    3='Foreign or international student';

VALUE SELECTVF      0='Not public or private nfp 4-year'
```

```
                    1='Very selective'
                    2='Moderately selective'
                    3='Minimally selective'
                    4='Open admission'
                    -9='{Missing}';

 VALUE CC2000AF     0='Not degree granting'
                    1='Public 2-year associate^s'
                    2='Public nondoctoral'
                    3='Public doctoral'
                    4='Private nfp nondoctoral except lib arts'
                    5='Private nfp doctoral and liberal arts'
                    6='Other public degree granting'
                    7='Other private nfp degree granting'
                    8='Private for-profit degree granting';

 VALUE PDADEDF      0='Do not know father^s education level'
                    1='Did not complete high school'
                    2='High school diploma or equivalent'
                    3='Vocational or technical training'
                    4='Less than two years of college'
                    5='Associate^s degree'
                    6='2 or more years of college but no degree'
                    7='Bachelor^s degree'
                    8='Master^s degree or equivalent'
                    9='First-professional degree'
                    10='Doctoral degree or equivalent';

 VALUE PMOMEDF      0='Do not know mother^s education level'
                    1='Did not complete high school'
                    2='High school diploma or equivalent'
                    3='Vocational or technical training'
                    4='Less than two years of college'
                    5='Associate^s degree'
                    6='2 or more years of college but no degree'
                    7='Bachelor^s degree'
                    8='Master^s degree or equivalent'
                    9='First-professional degree'
                    10='Doctoral degree or equivalent';

 VALUE PAREDUCF     0='Do not know parent^s education level'
                    1='Did not complete high school'
                    2='High school diploma or equivalent'
                    3='Vocational or technical training'
                    4='Less than two years of college'
                    5='Associate^s degree'
                    6='2 or more years of college but no degree'
                    7='Bachelor^s degree'
                    8='Master^s degree or equivalent'
                    9='First-professional degree'
                    10='Doctoral degree or equivalent';

DATA X4;INFILE 'C:\ECBW\F09\DATA\F9SCH.DAT' LRECL=1024 PAD;INPUT ID 1-6
  SCHIPEDS 11-16 SCHNAME $ 17-116;

LABEL
  SCHIPEDS='IPEDS number'
```

```
  SCHNAME='School name';

PROC FORMAT;
 VALUE $SCHNAMEF         '01'='{Alpha string}'
                        -9='{Missing}';

DATA X5;INFILE 'C:\ECBW\F09\DATA\F9CODE.DAT' LRECL=1024 PAD;INPUT ID 1-6
  MCMAJ1 $ 7-86 MCMJ1SPE $ 89-92;

LABEL
  MCMAJ1='Primary major: string'
  MCMJ1SPE='Primary major: specific CIP code';

PROC FORMAT;
 VALUE $MCMAJ1F         '01'='{Alpha String}'
                        -3='{Skipped}'
                        -9='{Missing}';

 VALUE $MCMJ1SPF         '01'='Total number of valid codes'
                        -3='{Skipped}'
                        -9='{Missing}';

LIBNAME PETS 'D:\Documents\School Work\Honors Research\BPS 04-09 Data\BPS
PETS\Data Files\SAS';

DATA X6;
SET PETS.DERIVED;
KEEP ID QE1STSTM QE2NDSTM QE3RDSTM QE4THSTM QE5THSTM QE6THSTM;
RUN;

DATA ALLSTUDENTS;
SET X1;
SET X4;
SET X5;
SET X6;
RUN;

*** Select only full-time bachelor's degree students at 4-year schools ***;
DATA ALLSTUDENTS;
SET ALLSTUDENTS;
IF FALLDGFT ne 3 THEN DELETE;
IF UGDEG ne 3 THEN DELETE;
IF SELECTV2=0 THEN DELETE;
RUN;

*** Rename variables ***;
DATA ALLSTUDENTS;
SET ALLSTUDENTS;
RENAME AGE = Age
         GENDER = Gender
         RACE = Race
         DELAYENR = DelayedEntry
         LOCALRES = Housing2003
         ATBAM6Y = DegreeMonthsElapsed
         ATBAEN6Y = DegreeMonthsEnrolled
         ATTYPE6Y = DegreesAttained2009
         ATHTY6Y = HighestDegree2009
```

```
          ENINUM6Y = NumInstAttended
          ACAD04A = GraduateInstructors
          ACAD04C = LargeClasses
          FREQ04A = FacultyMeet
          HIGHLVEX = HighestDegreeExpected
          TESATDER = SAT
          TESATMDE = SATMath
          TESATVDE = SATVerbal
          CRDHS04 = CollegeCredInHS
          CRDCL04 = CollegeClassInHS
          HCGPAREP = HSGPA
          HSTYPE = HSTYPE
          HCMATH = HighestMathHS
          HCYSMATH = YearsMathHS
          HCYSSCIE = YearsSciHS
          NPFST04 = FirstChoice
          CAGI = HHIncome
          PCTALL = HHIncomePct
          SAMESTAT = InState
          ENRLSIZE = InstEnrollment
          SELECTV2 = InstSelectivity
          CC2000A = InstCarnegie
          PDADED = FatherEd
          PMOMED = MotherEd
          PAREDUC = ParentEd
          SCHIPEDS = InstIPEDS
          SCHNAME = InstName
          MCMJ1SPE = FinalMajorCIP
          QE1STSTM = STEMCreditsY1
          QE2NDSTM = STEMCreditsY2
          QE3RDSTM = STEMCreditsY3
          QE4THSTM = STEMCreditsY4
          QE5THSTM = STEMCreditsY5
          QE6THSTM = STEMCreditsY6;
RUN;

*** Delete observations with missing variables ***;
DATA ALLSTUDENTS;
SET ALLSTUDENTS;
IF SAT=-3 THEN DELETE;
IF SATMath=-3 THEN DELETE;
IF SATVerbal=-3 THEN DELETE;
IF YearsMathHS=-3 THEN DELETE;
IF YearsSciHS=-3 THEN DELETE;
IF STEMCreditsY1<0 THEN DELETE;
IF STEMCreditsY2<0 THEN DELETE;
IF STEMCreditsY3<0 THEN DELETE;
IF STEMCreditsY4<0 THEN DELETE;
IF STEMCreditsY5<0 THEN DELETE;
IF STEMCreditsY6<0 THEN DELETE;
IF DelayedEntry=-3 THEN DELETE;
RUN;

*** Create STEM major variable ***;
DATA ALLSTUDENTS;
SET ALLSTUDENTS;
IF STEMCreditsY1+STEMCreditsY2 >= 16 THEN InitialSTEM=1;
```

```
ELSE InitialSTEM=0;
RUN;

*** Create STEM dataset ***;
DATA STEM;
SET ALLSTUDENTS;
IF InitialSTEM=0 THEN DELETE;
RUN;

*** Delete variables with missing data ***;
DATA STEM;
SET STEM;
IF SAT=-3 THEN DELETE;
IF SATMath=-3 THEN DELETE;
IF SATVerbal=-3 THEN DELETE;
IF YearsMathHS=-3 THEN DELETE;
IF YearsSciHS=-3 THEN DELETE;
RUN;

*** Create graduation variable ***;
DATA STEM;
SET STEM;
IF HighestDegree2009=3 THEN GraduatedALL=1;
ELSE GraduatedALL=0;
RUN;

DATA STEM;
SET STEM;
IF GraduatedALL=1 and (MAJ09A=1 OR MAJ09A=5 or MAJ09A=8 or MAJ09A=11 or
MAJ09A=18 or MAJ09A=25 or MAJ09A=34) THEN GraduatedSTEM=1;
ELSE GraduatedSTEM=0;
RUN;

*** Create survival variable ***;
DATA STEM;
SET STEM;
IF DegreeMonthsElapsed=0 THEN SurvivalMonths=72;
ELSE SurvivalMonths=DegreeMonthsElapsed;
RUN;

*** Create censoring variable ***;
DATA STEM;
SET STEM;
IF GraduatedStem=1 THEN CENSORED=0;
IF GraduatedStem=0 THEN CENSORED=1;
RUN;

*** Collapse race variable ***;
DATA STEM;
SET STEM;
IF RACE=4 THEN ASIAN=1; ELSE ASIAN=0;
RUN;

*** Collapse parent education variable ***;
DATA STEM;
SET STEM;
IF HHIncome<=18850 THEN Poverty=1;
```

```
ELSE Poverty=0;
RUN;

*** Collapse carnegie variable ***;
DATA STEM;
SET STEM;
IF InstCarnegie=4 or InstCarnegie=5 or InstCarnegie=7 THEN InstPrivate=1;
ELSE InstPrivate=0;
RUN;

*** Collapse highest math variable ***;
DATA STEM;
SET STEM;
IF HighestMathHS=4 THEN CalcHS=1;
ELSE CalcHS=0;
RUN;

*** Run proportional hazard model on entire dataset ***;
PROC PHREG DATA=STEM;
MODEL SurvivalMonths*CENSORED(1)= Gender Asian Poverty SAT CalcHS InstPrivate
NumInstAttended;
RUN;

*** Run proportional hazard model on those who didn't transfer ***;
PROC PHREG DATA=STEM;
WHERE NumInstAttended=1;
MODEL SurvivalMonths*CENSORED(1)= Gender Asian Poverty SAT CalcHS
InstPrivate;
RUN;

*** Run proportional hazard model on those who DID transfer ***;
PROC PHREG DATA=STEM;
WHERE NumInstAttended>1;
MODEL SurvivalMonths*CENSORED(1)= Gender Asian Poverty SAT CalcHS InstPrivate
NumInstAttended;
RUN;

*** Test for correlation between SAT and calculus ***;
PROC CORR DATA=STEM;
VAR SAT CalcHS;
RUN;
```