AN INTEGRATED MULTIPLE STATISTICAL TECHNIQUE FOR PREDICTING POST-SECONDARY EDUCATIONAL DEGREE OUTCOMES BASED PRIMARILY ON VARIABLES AVAILABLE IN THE 8TH GRADE

by

Gillian M. Nicholls

B.S. in Industrial Engineering, Lehigh University, 1987

Master of Business Administration, Pennsylvania State University, 1996

M.S. in Industrial Engineering, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Gillian M. Nicholls

It was defended on

July 1, 2008

and approved by

Mary E. Besterfield-Sacre, Associate Professor, Industrial Engineering Department

Henry W. Block, Professor, Statistics Department

Kim L. Needy, Associate Professor, Industrial Engineering Department

Larry J. Shuman, Associate Dean (Faculty), School of Engineering

Dissertation Director: Harvey Wolfe, Professor Emeritus, Industrial Engineering Department

AN INTEGRATED MULTIPLE STATISTICAL TECHNIQUE FOR PREDICTING POST-SECONDARY EDUCATIONAL DEGREE OUTCOMES BASED PRIMARILY ON VARIABLES AVAILABLE IN THE 8TH GRADE

Gillian M. Nicholls, PhD

University of Pittsburgh, 2008

There is a class of complex problems that may be too complicated to solve by any single analytical technique. Such problems involve so many measurements of interconnected factors that analysis with a single dimension technique may improve one aspect of the problem while overall achieving little or no improvement. This research examines the utility of modeling a complex problem with multiple statistical techniques integrated to analyze different types of data. The goal was to determine if this integrated approach was feasible and provided significantly better results than a single statistical technique. An application in engineering education was chosen because of the availability and comprehensiveness of the NELS:88 longitudinal dataset. This dataset provided a huge number of variables and 12,144 records of actual students progressing from 8th grade to their final educational outcomes 12 years later.

The probability of earning a Science, Technology, Engineering, or Mathematics (STEM) degree is modeled using variables available in the 8th grade as well as standardized test scores. The variables include demographic, academic performance, and experiential measures. Extensive manipulation of the NELS:88 dataset was conducted to identify the student outcomes, prepare the covariates for modeling, and determine when the students' final outcome status

occurred. The integrated models combined logistic regression, survival analysis, and Receiver Operating Characteristics (ROC) Curve analysis to predict obtaining a STEM degree vs. other outcomes. The results of the integrated models were compared to actual outcomes and the results of separate logistic regression models. Both sets of models provided good predictive accuracy. The feasibility of integrated models for complex problems was confirmed. The integrated approach provided less variability in incorrect STEM predictions, but the improvement was not statistically significant.

The main contribution of this research is designing the integrated model approach and confirming its feasibility. Additional contributions include designing a process to create large multivariate logistic regression models; developing methods for extensive manipulation of a large dataset to adapt it for new analytical purposes; extending the application of logistic regression, survival analysis, and ROC Curve analysis within educational research; and creating a formal definition for STEM that can be statistically verified.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation was the culmination of a long process of reorienting my focus from that of an industry practitioner to that of an academic seeking to add to the body of industrial engineering knowledge. It would not have been possible without the encouragement, support, and guidance from my advisor Dr. Harvey Wolfe. When my industry career brought me to Pittsburgh, PA, Dr. Wolfe suggested I consider pursuing a doctorate at the University of Pittsburgh at some point. Neither of us anticipated then how soon I would act upon his advice. After consulting a series of professional colleagues who had worked in industry, completed a doctoral program, and started a second career in academia; I decided to take the plunge. Throughout the long process of taking classes, working as a research/teaching assistant, developing the dissertation proposal, conducting the analysis, writing the dissertation, and defending the dissertation; Dr. Wolfe was there to help as an advisor, mentor, and friend. It has been an honor to work with him, and I feel particularly lucky that he delayed his retirement to see me through the process as his final doctoral student.

Returning to school as a full time graduate student was made possible by the financial support I received as a research assistant for Dr. Mary Besterfield-Sacre on a grant from the National Science Foundation (Math and Science Partnership HER-0227016) entitled "System-wide Change for All Learners and Educators" (SCALE). The research I conducted on this project for Dr. Besterfield-Sacre directly led to my interest in testing for statistically significant

differences between STEM and Not-STEM students. Dr. Besterfield-Sacre's participation in my dissertation committee and the guidance she provided were very valuable.

Dr. Larry J. Shuman provided material support and advice in setting the dissertation goals, sharpening its focus, and communicating the findings. His input in shaping one of the dissertation's key precursor journal articles was particularly helpful. Dr. Kim LaScola Needy was another valued member of the dissertation committee. She was also one of the people I consulted about pursuing a doctorate and was among the most persuasive in encouraging me to take this path. The final member of my dissertation committee, Dr. Henry Block, was instrumental in exposing me to Survival Analysis methods and conveying just how powerful this analytical technique could be. He is a gifted teacher and provided valuable insights into adapting survival analysis and categorical data analysis methods for this research.

During several semesters of the doctoral program I received financial support from the Industrial Engineering Department through teaching assistantships. I am very grateful to Dr. Bidanda Bopaya for providing both financial support and the experience of assisting Dr. Bryan Norman, Dr. Mary Besterfield-Sacre, and Dr. Lisa Maillart with various classes. Each course was a learning opportunity, and I appreciate the guidance provided by the professors.

During my time as a graduate student, I had ample opportunity to observe that the Industrial Engineering Department, like every successful organization, relies on the often unrecognized contributions of the administrative staff. Minerva Pilachowski sorted out issues with work assignments, payroll, benefits, and office space with quiet efficiency and good humor. Richard Brown, Lisa Bopp, and Nora Siewiorek helped with the myriad of administrative details that make the difference between having things go right or wrong, and they were invariably kind in dealing with all the graduate students. Chalice Zavada made sure that the department's

funding from research projects such as the one that supported me for three years went smoothly. Jim Segneff and Frank Kremm provided critical support in keeping the computers, printers, audio/visual equipment, data storage servers, licensing of software, and general telecommunications working properly.

Several other people encouraged me to consider returning to school to pursue a doctorate. Dr. Ching-Chung Kuo was my M.B.A. thesis advisor and began urging me in 1995 to continue my graduate studies. He has been a good friend and mentor ever since. Dr. Jessica Matson gave me my first teaching job and has been very supportive of my quest for the doctorate. Keith Robbins and Jerry Vest were supervisors during my time in the railroad industry and have become good friends. Having worked with me during my first graduate degree program, Keith was uniquely positioned to advocate that I embark on the second program. Both gentlemen shared the benefits of their long experience in freight railroading throughout our work together. Without the career development provided by Conrail, CSX Transportation, and Bessemer & Lake Erie, it would have been very hard to return to school full-time.

There were a number of graduate students that I was privileged to meet through this process. Their friendship and support was one of the nicest fringe benefits of completing the degree. They include Renee Walwender Clark, Erin Gross Claypool, Bradley Golish, Guiping Hu, Rob Koppenhaver, Jennifer Kreke Sindelar, Chen Li, Gorkem Saka, Rebeca Sandino, Natalie Scala, Mark Sindelar, Susan Olsen Sullivan, Natasa Vidic, Lizhi Wang, and Tuba Pinar Yildirim.

Last but not least, I must acknowledge the great debt that I owe to my family. My parents, Drs. Joan and Richard Nicholls, gave me the unparalleled gifts of a wonderful family and unflagging support throughout the different phases of my life. One of their best gifts was my

brother Rick Nicholls. Rick is a treasured friend as well as my favorite brother (that he is my only sibling does not diminish that sentiment). I wish my maternal grandmother had lived long enough to see me complete this degree since Gram was one of the most influential people in my life. Going forward, I hope to live up to the best career advice my father ever shared with me: "Always leave them laughing, Gillian."

1.0 ANALYSIS OF COMPLEX PROBLEMS

1.1 INTRODUCTION

There is a class of complex problems that may be too complicated to be solved by any single analytical technique. Such problems involve so many measurements of interconnected factors that single dimension techniques are very limited in their ability to conduct significant analysis. Sometimes sophisticated modeling can provide an excellent or very good solution to these types of problems. However, frequently utilization of a single analytical technique may lead to an improvement in one aspect of the problem, but it risks achieving no improvement over the remaining portions of the system or conceivably worsening the overall status. Such complex problems involve assorted data collected at many points in the system and at various times. In order to analyze this class of problems and obtain significant improvements, a methodology that employs multiple modeling techniques and an extensive dataset might be valuable.

Two questions are raised by this scenario. Is it feasible to develop a methodology integrating multiple analytical techniques and test it against a standard single analytical technique in solving a complex problem application? If the methodology is feasible, would it provide a solution that was significantly better than that of the standard single analytical technique?

A “complex” problem may be defined as a problem that is sufficiently large and intricate that creating a smaller version with simplifying assumptions for analysis is of limited usefulness. Analysis of such a simplified sub-problem does not allow for the interactions with other factors that affect the problem and risks producing results that lead to a significantly sub-optimal solution. Complexity in this sense does not refer to the information theory definition of complexity which is the amount of time required to solve a problem of a given size with an algorithm.

Characteristics of Complex Problems:

- The process contains numerous interconnected factors such that changing one aspect has ramifications for other factors
- The process occurs over an extended time period that exceeds a threshold number of time units
- The process has numerous inputs and outputs such that creating a simplified sub-problem that realistically mirrors the original problem is difficult
- Measuring the state of the process requires extensive data collection at many different points throughout the process functionally and temporally

Attributes of a Dataset for a Complex Problem:

- May be high dimensional in terms of numbers of variables collected
- Data is often measured at different times
- Dataset may include repeated measures
- May contain many categorical variables as well as continuous variables

- May have multiple sources of data such that variables are obtained from different parts of the system
- May consist of an extensive number of records

In some instances a methodology is sufficiently comprehensive to provide an excellent solution if the modeling is done carefully. While this is difficult using very large datasets it is possible. In other instances additional techniques need to be applied concurrently or sequentially to enhance the solution. In either case, the models need to be applied in an integrated way.

The concept behind the integrated methodology is similar to approaches developed to solve other problems. For example, integer optimization is one of the most challenging classes of problems in operations research. Optimizing a solution in which numerous variables are constrained to be integers is far more difficult than doing so for a problem in which the variables have a continuum of values. Algorithms have been developed to solve these problems by decomposing them into simpler problems and applying a series of techniques in combination. Benders Decomposition[1] works by simplifying the problem into the original objective function with a subset of the original constraints. Once the simplified problem is solved additional constraints are added to successively shape the problem until an acceptable answer to the original problem is found. This combines delayed constraint generation with the cutting plane method to produce better answers than if a single technique had been applied. In the methodology of this dissertation, different statistical techniques were linked sequentially and in parallel such that each was matched to appropriate data from portions of the problem and the output of one technique became an input to the other.

The methodology in this dissertation integrated statistical techniques to analyze a complicated problem involving different types of data measures. The application chosen was the process of students completing their education during a 12 year period from the 8th grade in order to predict which students would get a four year college degree in a quantitative subject. Variables that measured demographic, academic, attitudinal, and experiential factors for the students were analyzed using multiple regression methods[2] including nonlinear regression[3] to determine which variables were significant predictors of educational outcomes. For example, data collected during the 8th grade were examined with logistic regression[4] to predict which students earned a bachelors degree in a subject such as engineering vs. another outcome within the 12 year period. Data that reflected the educational outcome status of the students at different points over the period were examined using survival analysis methods[5] to estimate the probability of a student remaining on course to achieve a degree in a quantitative subject beyond a given point in time. An integrated model was developed that linked different statistical techniques to analyze multiple factors from 8th grade, standardized test scores collected by 12th grade, and the time at which the students' educational outcomes occurred in order to predict the students' final outcomes. Models were also fitted with the 8th grade variables and without the standardized test scores. The predicted outcomes were compared to actual outcomes to test the research hypothesis that an integrated multiple technique methodology would provide a better solution than a standard single technique methodology.

1.2 THE NATURE OF A COMPLEX PROBLEM: AN EXAMPLE

An example of the class of complex system problems for which an integrated model is applicable is the network of a large transportation provider. For example, a Class One[6] freight railroad in the United States typically maintains a network of thousands of miles of rail tracks, hundreds of railroad yards, thousands of employees, and an enormous fleet of railcars, locomotives, and maintenance equipment. Every day freight trains originate at one point on the railroad, operate across the network, and terminate at another point. The termination of the train occurs with the final delivery of the cars, the further sorting of the cars to other destinations, or the interchange of cars to a different railroad. Cars circulate through the system with empty cars delivered to shippers for loading and loaded cars transported to receivers. Locomotives are assigned to trains as required by the tonnage to be hauled, the elevation traveled, and the later demands for locomotives at the destination. Train crewmembers are assigned to specific trains based on work requirements, federal hours of service regulations, agreements with labor unions, and the need to balance the supply of crews at origin/destination locations. The movement of trains through the network is controlled by train dispatchers that decide the sequence of trains traveling through the territory they are responsible for based upon the train schedule, priority of traffic, the performance of individual trains, and responses to unexpected events (derailments, accidents, mechanical failures, etc.).

An attempt to improve one aspect of the transportation network by applying a single analytical technique risks ignoring the interconnectedness of the system and degrading its overall performance for the sake of a small improvement in one narrow area. For instance, consider a single scheduled train operating from point A five days per week and traveling to point B over the course of 24 hours. If the objective were to analyze the on time record of this scheduled train

and improve it by reducing late arrivals then a variety of single techniques are available. The train's performance is measured against its schedule at various geographic points in its progress to forecast its total "lateness" as a function of its accumulated late/early arrival times at a series of intermediate points. Factors that are involved in the train's performance include the number of cars picked up/dropped off at each intermediate point, the mix of locomotives assigned, and competition from other trains for transit across the trackage.

Any attempt to optimize the performance of this single scheduled train risks potentially adverse effects on other trains or support functions within the network. For example, assume the analysis reveals that lateness typically occurs near the middle of the train's journey where it encounters congestion among other trains seeking to use the same trackage. The train under study can be optimized by increasing its priority to the train dispatcher controlling that territory. As a consequence, other trains are delayed. The net effect of delaying other trains while reducing lateness for the train of interest may be higher expenses. If the most significant factor causing lateness is the typical assignment of lower powered locomotives to the train this can be ameliorated by assigning additional or higher powered locomotives to the train. Since locomotives are a scarce resource, the assignment of more resources to the train in question reduces locomotives available for assignment to other time-sensitive trains. This is an obvious risk of making improvement decisions on the basis of a single analytical technique applied to a complex problem.

The use of single dimensional analytical techniques to solve problems in complex networks is more successful if a specific sub-problem can be identified that presents limited interactions with other aspects of the network. If the problem is too complex to isolate it from the other portions of the network, single techniques are limited in their usefulness.

Developing a methodology to integrate multiple statistical techniques in modeling complex problems required working with a practical application to create a model and validate the methodology. A model that was successful in analyzing one application could lead to a general model applicable to other complex problems. The application chosen for development and testing of the model required extensive background knowledge to understand the problem's aspects and select appropriate statistical techniques to be integrated. The more complex the problem area, the greater the background knowledge required to work with it.

As discussed, transportation networks can be very extensive involving thousands of personnel, expensive capital equipment, and a myriad of interconnected support functions. Other complex problems exist on a smaller scale sufficient to develop the integrated methodology and demonstrate its viability. An aspect that was weighed in selecting the application was the accessibility of data. The voluminous and varied data required to examine a complex problem was difficult to obtain in the transportation network instance. Commercial freight transportation networks operate in a highly competitive environment where such data is a sensitive, proprietary resource to be shielded from competitors, customers, and potential hostile actors such as terrorists. Gaining access to such data and permission to conduct publishable academic research was not feasible.

1.3 EDUCATION AS AN APPLICATION AREA

The education of students in the U.S. is another example of the type of complex system problem for which the integrated methodology was suitable. The education system is a network of teachers, students, administrators, and support staff working to instill a basic level of competency

in subjects including reading, English, history, mathematics, science, and physical education. Additional training in the arts, sciences, and vocational/technology topics are provided for students to partake as their interests direct. Teachers present information in a given subject at each grade level that builds upon knowledge students have acquired in earlier grades. Teachers also work to correct deficiencies in students' learning and assist those struggling with concepts. Administrators work with students, teachers, staff, school boards, parents, and local communities to manage educational resources. Students progress through grades crossing between elementary, middle, and high schools with learning at later points affected by what was learned earlier. Many students progress further to additional vocational training or institutions of higher education.

Students' educational progress is assessed regularly through assignments, tests, and course grades. Students' attitudes and behaviors are periodically measured with survey instruments. Variables that measure school resources such as the numbers of teachers by subject, budget dollars, etc. are recorded at regular intervals. Local, state, and federal education entities measure attributes of schools and their overall educational performance. Because of the significant amount of government funded research being done on education, varied and voluminous datasets were readily available to a researcher as opposed to the proprietary data for the freight transportation network instance. Aspects of the education system have complexity similar to the freight railroad example given previously, but with the greater data availability this area was more conducive for development of the proposed methodology.

A study of the complex process involved in U.S. students' acquisition of college degrees in Science, Technology, Engineering, and Math (STEM) was the education application chosen for the methodology development. Currently, insufficient numbers of American students are

achieving college degrees in STEM topics[7]. The number of U.S. citizens and permanent residents who earned bachelors, masters, and doctoral degrees in Engineering increased very little from 1996 through 2005 despite greater increases in undergraduate enrollment levels from 1992 through 2004[8] on. Degrees in mathematics followed a similar trend. Graduate degrees in other areas of Science increased slightly or declined through 2001 despite increasing numbers of full-time graduate students[9]. The inputs to the STEM degree acquisition process are the quantity and quality of students, educators, and institutions. The production processes are the mechanisms by which American students are recruited, educated, and developed into STEM degree-holders. The students are a prime area upon which to focus initial analysis since they represent the "raw material" to be transformed by the education system into STEM degree graduates. As with all raw materials to a system, there may be issues with supply and input quality affecting the system's output. For college students, this would translate into the number of American students entering STEM degree programs, their educational performance capabilities, and their motivation to persevere until achieving a bachelor's degree. Examining this issue with a single statistical technique may ignore the complex nature of the problem.

Performance capability can influence interest in pursuing a course of study since people are naturally inclined towards subjects within which they feel confident of success. Astin[10] found that students with a higher probability of persisting in engineering majors or switching to engineering from another major had high self-confidence in their mathematical abilities. Astin's results indicated that students were more apt to switch majors out of disappointment or frustration if their college grades were poor. Performance capability develops over time as students acquire greater knowledge and experience. Thus an attempt to increase the number of STEM degree-holders must explore significant factors affecting students' learning in core STEM

areas over time, their interest in pursuing a STEM degree, and their persistence in achieving the degree. These factors may encompass the students' academic experience throughout their pre-college education, demographic information, and education institutions' characteristics.

Berryman[11] states that fewer Americans achieving degrees in STEM will mean reduced competitiveness of America in the fields that rely heavily on skills acquired through the study of STEM topics. These fields have traditionally offered high-paying salaries and prestige. Having fewer Americans competitive in these fields limits their ability to compete for these higher wage/prestige jobs. Data to examine this application is available from the U.S. Department of Education's National Center for Education Statistics (NCES). NCES has conducted a series of extensive longitudinal studies to collect data about selected students, their families, their schools, and their teachers in each decade since the 1970's. Among them is the National Education Longitudinal Study of 1988 (NELS:88)[12] conducted between 1988 and 2000. The dataset from this study is impressively comprehensive and captures many aspects of the complex education problem including demographic characteristics, school characteristics, coursework taken, and cognitive test results.

Overall, education is a valid example of a complex problem involving many different interconnected factors in a process that occurs over time. Educational attainment by the general public and the numbers of STEM degree-holders are of great importance to society. Examining the process of STEM degree acquisition was a useful test of the integrated model approach.

1.4 THE RESEARCH PROTOCOL

This dissertation examines the utility of modeling a complex problem with a set of statistical techniques integrated to analyze different types of data measured from the problem. The goal was to determine if this integrated approach was feasible and provided significantly better results than employing a single statistical technique. An application in engineering education was chosen because of the availability and comprehensive nature of the NELS:88 dataset. The NELS:88 dataset provided a huge number of potential predictor variables and 12,144 records of actual students progressing from 8th grade through a 12 year period showing their actual educational outcomes.

In order to address the inconsistency in defining STEM by prior research, the potential educational outcomes were classified as earning a STEM bachelor degree; earning a bachelor degree in a major that involved quantitative coursework similar to STEM (STEM-Related); earning a bachelor degree in another four year program; earning an associate or certificate degree; or earning no college degree. The no college degree category was further subdivided into students that dropped out of high school, students that dropped out of college, students that were still in college at the study's conclusion, and students for whom completing high school was the highest educational attainment. Various combinations of the categories were created so that multiple pairs of outcomes could be modeled. A rigorous approach was developed to construct multiple samples of fit and test data so that different versions of the same model could be created and compared to one another in how well the fitted models performed when applied to the test data samples.

Logistic regression was chosen as the statistical technique used to create the initial set of predictive models and to serve as the standard by which the integrated modeling approach was

judged. A set of potential covariates was selected after examining the variables available from the students' 8th grade and standardized test scores by 12th grade. The variables selected as potential covariates measured factors that prior educational research found significant in analyzing STEM students. Extensive manipulation of the NELS:88 dataset was conducted to identify the student outcomes, prepare the set of potential covariates for modeling, and determine the time at which students experienced their final educational outcome. Much of the data was categorical in nature and lacked a clear binary or ordinal scale. The covariates were adjusted to ensure each could be utilized in the model fitting process. Multiple random samples were drawn from the data so that prediction models were fitted multiple times and tested against student records not used in the model fitting process. Predicted outcomes were obtained from the fitted models applied to the withheld test data. The predicted outcomes were compared to the actual outcomes by determining the number of correct and incorrect STEM predictions. The logistic regression models are much stronger than originally anticipated and provide good or better predictive accuracy. These models are potentially useful in their own right for future educational research.

Survival analysis was chosen as another statistical technique to be linked with logistic regression analysis in creating the integrated models. Survival analysis was employed to gain additional information from the NELS dataset by examining if and when students failed to "survive" on the track to earning a STEM degree and instead experienced a different educational outcome. Analyzing the different survival times for the students and the factors that potentially affected the times provided a way to use additional data about the students' educational process to see if better predictions could be obtained by knowing differences in the STEM track survival times.

ROC Curve analysis was the third technique chosen to analyze how responsive the prediction accuracy was to slight changes in the models' settings. This technique visually depicted the percentage of correct STEM predictions vs. the percentage of incorrect STEM predictions for different model settings.

The integrated model was constructed by using the estimated probability of a STEM outcome from the logistic regression analysis and its predicted outcome for each student as input variables for survival analysis and applying the third technique to examine the correct vs. incorrect predictions that resulted. One integrated approach linked each analytical technique in series while a second also linked the logistic regression in parallel so that final predictions of the students' earning a STEM degree depended upon agreement between both techniques. Different integrated models were fitted and tested using the same samples created for the logistic regression analysis.

2.0 LITERATURE REVIEW

2.1 INTRODUCTION

There are areas relating to this work in which significant prior research has been accomplished. These include the acquisition of STEM degrees, the disparities in educational achievement between different population subgroups, and the identification of variables significant in the educational acquisition process. In addition, quantitative methods employed in the conduct of this research are based upon work done in other areas particularly Survival Analysis and Receiver Operating Characteristics (ROC) Curve Analysis. The background for these techniques is also described as part of the literature review. Both of these methodologies have been used extensively in other areas such as medical research, but their application to educational outcome modeling is newer.

2.2 THE STEM PROBLEM

As discussed earlier, one approach to increase the number of American students that pursue STEM degrees in college is to widen the pool of students that consider studying STEM. There are numerous theories for the disparity in academic achievement between various segments of the American population. Differences exist in the preparation of children for elementary school

between the races/ethnicities. Caucasian children tend to start elementary school with a larger spoken vocabulary[13] and more exposure to reading. Socioeconomic status may affect the access to resources that assist children[14] in learning. Parents of greater economic means can afford to select housing in areas with higher quality schools, send their children to private schools with superior academic records, and pay for private tutoring. Differences in family structure including parenting practices and educational involvement can affect children's emotional support[15] for learning. Schools vary in the quality of education they deliver, and many of the schools with primarily minority or poor populations[16] are of low quality. Differences in expectations by teachers may affect teaching and ultimately how students perform[17]. Another theory suggests minority children may feel "threatened" in test situations by fear that they will not do well, thereby, confirming negative stereotypes about their group's academic skills[18]. In effect, the negative stereotype may cause test anxiety that leads to lower performance on tests.

The gaps in performance that exist at the elementary school level tend to expand as children progress through secondary school[19]. Acquiring a college degree in a STEM subject requires strong skills in science and mathematics. Students who lack strong self-confidence in their skills in these topics are less likely to choose a STEM major in college[20].

These disparities in academic performance between the races/ethnicities are generally referred to as the "achievement gap." Different approaches are being pursued to solve the achievement gap problem and promote overall higher performance by students of all segments of the population. These include increasing the rigor of the academic curriculum[21] [22], reducing the student/teacher ratio[23], increasing school budgets[24], offering tutoring programs[25], creating mentoring programs[26], instituting regular standardized testing of school achievement[27],

increasing teacher performance standards[28], providing vouchers for other schools[29], creating charter schools[30], and experimenting with single-sex classrooms[31] [32].

2.3 DATABASES

General research into students' college educational experience has been extensively conducted. Among the datasets developed for this purpose are the Cooperative Institutional Research Program (CIRP) [33], the Your First College Year survey (YFCY), the National Longitudinal Study of the High School Class of 1972 (NLS-72)[34], the High School and Beyond Study (HS&B)[35], the National Education Longitudinal Study of 1988 (NELS:88), and the ongoing Educational Longitudinal Study of 2002 (ELS:2002)[36].

CIRP is operated by the Higher Education Research Institute (HERI). Since 1966, CIRP has conducted surveys of incoming college freshmen students collecting data including their high school grades, SAT/ACT scores, attitudes, behaviors, goals, and intended major. More than 11 million students have been surveyed across about 1,800 community colleges, four-year colleges, and universities. The CIRP data contains self-reported data about the students' high school experiences from the perspective of an incoming college freshman. It does not contain independently verified data about high school performance or measures of students' attitudes, behaviors, and experiences at time points during the period of high school.

Institutions that agree to participate in the CIRP survey are automatically entitled to receive the data obtained from incoming freshmen for their school. Institutions are then able to use the data to analyze trends in their incoming students from year to year. A researcher may gain access to another university's CIRP data by contacting the other school to request that it be

provided. If the request is approved, the other institution provides its data directly. HERI has conducted a set of follow up CIRP surveys which track the educational outcomes of students originally surveyed as freshmen. In addition, HERI now offers institutions the option to also conduct surveys of students as they finish their freshmen year and again in their senior year to find out how their attitudes, behaviors, etc. developed since the incoming freshman survey. The Your First College Year survey (YFCY) was initiated in 2000. The College Senior survey (CSS) was redesigned from a more general college student survey to serve as an exit survey. The data collected continues to focus on attitudes, behaviors, experiences, career aspirations, and cultural aspects. Students are asked to report their average grade level, the highest degree expected to be completed soon, and any intentions they have to pursue a graduate degree. This allows institutions to study trends in how students progressed from their freshman to senior years.

The National Education Longitudinal Study of 1988 (NELS:88) study consisted of collecting demographic, attitudinal, experiential, educational, and vocational data about a representative cohort of American students at specific stages in their scholastic progression. The goal of the study was to be able to draw conclusions about the factors that could affect the student's progression and achievement by 2000. Academic performance was validated by obtaining transcripts from post-secondary school attended and by conducting cognitive learning tests in three waves of data collection during high school. Parents, teachers, and school administrators were also invited to complete surveys of questions regarding specific students participating in the study. In contrast to the CIRP survey, NELS:88 data was collected at periodic intervals during the high school years, during the likely midpoint of college, and after most students had completed their post-secondary education.

The NELS:88 dataset is very impressive in terms of the scope of the data collected and the degree of detail obtained. Portions of the dataset are available to the public for analysis and review. Obtaining access to the full dataset including sensitive restricted access variables such as academic performance measures, standardized test scores, college transcript records, and identifiable details is limited to researchers that complete an extensive set of application requirements. Researchers seeking access to the restricted data must supply a rigorous explanation of the proposed research, obtain their institution's Institutional Review Board (IRB) approval, provide a security plan to safeguard the data, and sign notarized affidavits agreeing to keep the data confidential.

The NELS:88 study was part of a series of high school longitudinal studies developed by the National Center for Education Statistics (NCES) including the National Longitudinal Study of the High School Class of 1972 (NLS-72), the High School and Beyond Study (HS&B), and the Educational Longitudinal Study of 2002 (ELS:2002). The ELS:2002 study began with students as high school sophomores in 2002 and continued with the first follow up in 2004 when most of the students were seniors. A second follow up was conducted in 2006 as many students were completing their second year of post-secondary education. One to two more follow ups may be conducted in 2010 and 2012 to determine the overall life outcomes of the students. While the ELS:2002 dataset contains the most current longitudinal data for high school students, its incomplete status made it less suitable for this research than the NELS:88 dataset. In addition, the ELS:2002 study began with students in 10th grade as opposed to the 8th grade of the NELS:88 study. The potential for a successful intervention to encourage students towards STEM is higher at an earlier point in their educational career prior to their assignment to a general mathematics path vs. a college-preparatory mathematics path[37] for the final years of high

school. ELS:2002 would be appropriate for ongoing studies of students' educational choices and outcomes.

2.4 PRIOR EDUCATION RESEARCH

2.4.1 CIRP-based Studies

The CIRP data has been utilized by many researchers to study students in U.S. colleges. Sax[38] studied students that achieved a bachelor's degree in a STEM* subject to determine the likelihood they would go on to pursue a scientific research career. She explored differences by gender in the students' persistence in a scientific research career. Persistence in Sax's research was defined as students who achieved a bachelor's degree in a STEM major continuing their STEM education until they earned a STEM graduate degree. It should be noted that other researchers have defined persistence differently to suit the educational outcomes they were examining. Examples of these include persisting through high school until achieving a diploma, persisting through college to earn a bachelor's degree, or persisting in a particular major to earn a degree in that subject.

The study reviewed prior research into persistence predictors including science/math preparation, undergraduate experience, post-graduate experience, parental influences, personal confidence in skills, and balancing family vs. career. Initial freshman data was collected from

* The term SME is often used by researchers to refer to science, math, or engineering. The acronym STEM will be used in this document to refer to these areas of study.

12,000 students as part of CIRP and follow-up data was collected four and nine years later on the cohort of 2,563 that achieved a science, math, or engineering bachelor's degree.

Sax employed stepwise linear regression to identify the student profiles and undergraduate measures that best predicted pursuing a STEM graduate degree. The CIRP data was used to obtain specific independent variables reflecting the students' goals, reasons for choosing their college, activity times per week, intended major, intended career, and family educational/career history. There were differences between the areas of STEM studied in determining which students were most likely to pursue a STEM graduate degree. Engineering and the physical sciences were more likely to produce students interested in graduate school than mathematics/computer sciences or biological sciences. Gender differences were observed in the graduate school tendencies of students in the physical sciences and mathematics/computer science. Female students in these fields were less likely than male students to pursue a graduate degree. The desire to make a theoretical contribution to science was the most valuable predictor of females' persistence towards a STEM graduate degree. Another positive predictor for females was having a mother that was a college educator or a research scientist.

A study by Smyth and McArdle[39] used CIRP data to obtain students' average high school grades, SAT scores, and intended major. This was combined with the students' college transcripts data and college attributes to explore racial/ethnic and gender differences in students achieving a STEM degree from selective institutions. The social sciences were deliberately excluded from the STEM degree category for this analysis. A series of hierarchical linear models (HLM) were used to estimate the effects of student oriented variables based upon the institutional characteristics. The authors modeled the log-transformed odds of graduating with a STEM degree to avoid serious violations of standard regression assumptions such as normality

and linearity. The results indicated minority students were less likely than Caucasian students to achieve a STEM degree, and much less likely than Asian students to do so. Male students were more likely to graduate with a STEM degree than female students. Quantitative measures of student achievement such as SAT math scores and high school average grades were significant in predicting STEM graduation. Among the authors' conclusions was the notion that students would be advantaged by attending a school where the mean student measures in math and high school grades were comparable to their individual measures. The risk of not achieving a STEM degree increased when the student chose an academically challenging college in which his/her scores were in a low percentile.

Pascarella, Smart, Ethington, and Nettles[40] explored racial and gender differences in the impact of institution on social and academic self-concept of students. CIRP data from 487 institutions were combined with follow up data collected in surveys about nine years later. The CIRP data were used to obtain variables measuring pre-college self-concept, expectations, goals, and academic achievement for a sample of 4,597 students at 379 four-year institutions. The self-concept information was obtained from students' self-ratings of their academic and social qualities. Gender, race, degree aspirations, and socioeconomic status were also obtained from the CIRP dataset. Multiple regression analysis was utilized to test the hypotheses of gender/racial effects on the students' self-concept. The data was broken into four groups by gender (male/female) and race (black/white) for separate analysis of each cohort. College social and academic experiences had significant effects on the students' self-concept. The factors that affected academic and social self-concept were generally alike for race and gender. Pre-college social self-concept positively influenced social leadership/participation in college. Similarly, pre-college academic self-concept positively influenced college academic accomplishment.

Nicholls, Wolfe, Besterfield-Sacre, Shuman, and Larpkiattaworn[41] created a methodology for rapidly analyzing CIRP data to identify variables that predicted interest in a STEM major. Majors in Science, Technology, Engineering, Math, or the professional health fields were classified as STEM. All other college majors including "Undecided" were classified as "Non-STEM" in this analysis. Basic statistical tests were applied to the CIRP dataset variables to look for statistically significant differences between STEM and Non-STEM students in population subgroups divided by educational institution, gender, and race/ethnicity. Variables that consistently revealed significant differences between STEM and Non-STEM students across multiple subgroups were considered more valuable than variables which showed significant differences across fewer subgroups. The analysis of two universities' CIRP freshmen surveys found that 22 of 216 variables consistently exhibited significant differences between STEM and Non-STEM students across the multiple subgroups. The methodology identified several variables as the most consistently valuable predictors of STEM interest including high school GPA, SAT scores, self-rating of mathematics ability, self-rating of academic ability, and commitment to studying STEM. These findings were consistent with results reported by other researchers suggesting that the methodology developed was successful and could be applied to a general subject of interest. The approach allowed the voluminous CIRP database to be efficiently analyzed for variables that predicted an event or characteristic of interest.

Leslie, McClure, and Oaxaca[42] developed separate binomial logit, multinomial logit, and ordered logit models using data from CIRP and the National Longitudinal Survey of Youth (NLSY) [43] to predict achieving an engineering or science degree. Maximum likelihood estimators were employed. These methods were chosen since the CIRP and NLSY data are virtually all categorical rather than continuous in scale. The data was analyzed separately in

groups by gender (male/female) and race/ethnicity (black/white/Hispanic) where the volume of data was sufficient. The authors concluded that self-concept and self-efficacy were significant in explaining the lower participation by women and minorities in STEM studies. Caucasian males were the most likely to rate their science and math preparation as above average. The self-efficacy gained from this preparation indicated they were also more likely than Caucasian females, Hispanic females, and African-Americans of both genders to pursue a STEM major. Hispanic males showed even more positive effects from self-efficacy in math and science in choosing a STEM major. Females that possessed a lower self-concept and sense of control were less likely to select a STEM major. Caucasian females were far more likely to graduate with a STEM bachelors or masters degree if they achieved a B or higher grade point average (GPA) in their undergraduate years.

Leslie, et al. also found that having a parent employed in a STEM profession positively influenced the choice of a STEM major. The effects were largest for Hispanic males, African-American males, and Caucasian males, in that order. The effect was greater for male students than female. However, the mother's education level was one family-related variable for which the positive effect was greatest on female students' achievement of a STEM degree. Having lower interest in marrying was positively related to choosing a STEM major while popularity was negatively related to selecting STEM. Students that identified STEM as their likely career were more likely to major in and be employed within STEM. The effect was larger for males than females.

Astin[44] studied CIRP data for a group of 36,581 students across numerous institutions and found that the academic achievement in high school and quantitative tests such as the SAT accurately projected students' academic accomplishments in college.

2.4.2 Longitudinal Data-based Studies

Research by Zhang, Anderson, Ohland, and Thorndyke[45] obtained results that were similar to those of Astin. Zhang, et al. reported high school GPA and math SAT scores were significant predictors. Their research involved 87,167 engineering students at nine institutions that matriculated between 1987 and the summer of 1996. The data was obtained from the Southeastern University and College Coalition for Engineering Education (SUCCEED). The students were followed over time to determine graduation rates and estimate the time-to-graduation using multivariate logistic regression. Gender, ethnicity, verbal SAT scores, and citizenship were also significant predictors of graduation at some of the nine colleges and universities.

The attitudes of freshmen college students are another area that has been explored to explain why some students persist in engineering while others do not. Besterfield-Sacre, Atman, and Shuman[46] developed a methodology for assessing the attitudes, self-confidence, and expectations of freshmen engineering students. The goal was to determine if student attitudes could accurately predict academic performance and persistence in engineering such that targeted intervention programs could positively influence both. The analysis revealed that students who voluntarily left engineering while performing acceptably often had lower interest in engineering and were ranked highly in high school. These students were academically strong enough to be successful in engineering but were less personally motivated to pursue it. It was theorized that these students may have responded to family influence in selecting engineering initially and later chose to switch majors to pursue interests they personally found more appealing. In contrast, students who changed from an engineering major after performing poorly tended to have high

expectations of engineering and may have been drawn to it by anticipated financial benefits from employment after graduation.

A subsequent study by Besterfield-Sacre, Moreno, Shuman, and Atman[47] built upon this analysis by examining differences in attitudes among freshmen engineering students across 17 institutions by ethnicity and gender. The Pittsburgh Freshman Engineering Attitude Survey (PFEAS) was administered at the start of the freshman year and again at the end of the first semester or first academic year. Since the assumption of normality did not hold for much of the attitudinal data non-parametric comparison tests including the Mann-Whitney were used to detect differences between cohorts. Female students tended to have less confidence in their general knowledge of engineering and their ability to achieve success in this field. Male students tended to rate their study skills lower and their problem-solving skills higher than their female counterparts. There were also significant differences in attitudes between majority students and their African-American, Asian Pacific, and Hispanic counterparts. Interestingly, the changes in attitude between pre and post surveys varied across the institutions. There were differences in direction observed by gender and institution suggesting that the academic experience of the school attended may affect the students' self-concept and attitudes.

Larpkiattaworn, Muogboh, Besterfield-Sacre, Shuman, and Wolfe[48] examined issues associated with employing statistical analysis to assess and evaluate engineering education initiatives. The article illustrated several techniques for managing Type I errors when conducting multiple comparisons and provided guidance in selecting from among the techniques. The classic Bonferroni method, Scheffe's procedure, Tukey's test, and the sequentially rejective Bonferroni procedure (SRBP) were among the methods discussed. Two example cases were presented including scenarios under which a particular technique was best suited. The paper also

presented an argument for employing tree diagrams instead of classification tables when evaluating the predictive power of a model whose recommendations were not always followed. The researchers developed a model using logistic regression to predict whether freshmen engineering students would be more likely to succeed if they took a class in pre-calculus before calculus. Five separate models were created by randomly selecting samples from the same dataset and splitting them into portions for model fitting and model validation. The final prediction was the result of agreement between three or more of the five logistic regression models. Advisors were encouraged to consider the model's recommendation prior to making their own. Thus a classification table that compared predicted success/failure with actual success/failure did not accurately reflect actual results from ignoring the model's prediction. The tree diagram more accurately measured the model's effectiveness.

Adelman[49] analyzed longitudinal data over a 13-year time span for a set of students starting in 1982 during their sophomore year. The dataset employed was the NCES' High School and Beyond/Sophomore Cohort (HS&B/So). His findings included a direct relationship between the pattern of advanced technical courses in high school and choosing engineering as a college major. The students with a greater propensity to major in engineering were the students with higher quantitative standardized test scores and greater academic achievement. Students with lower test scores and lesser overall academic records were less likely to select engineering as a major. This pattern was also observed when examining the records of female students. There were comparatively fewer female than male students pursuing engineering degrees. However, those females majoring in engineering also tended to have higher academic achievement.

Hintze and Silberglitt[50] examined oral reading curriculum-based measurement (R-CBM) scores of students in 1st through 3rd grade to determine how well they predicted the students that would pass a high stakes academic capability test, the Minnesota Comprehensive Assessment (MCA), which was taken at the end of the 3rd grade. The R-CBM was tested eight times from the winter semester of grade 1 through the spring semester of grade 3 by having 1,766 students read carefully structured passages to demonstrate their level of fluency. Each of the R-CBM scores across the three year period was used as a predictor in a series of single variable models to project whether the students would pass or fail the MCA test. In addition, each of the 8 longitudinal R-CBM scores was used to predict the result of the successive R-CBM assessment. Logistic regression, receiver operating characteristics (ROC) curves, and discriminative analysis were used to create the predictions in order to determine if one method performed better than the others. The sensitivity, specificity, positive predictive power, and negative predictive power were assessed for each independent variable at a set "cut score" used as the dividing line between predicting the student's outcome on the MCA test.

Hintze and Silberglitt reported results indicating that the R-CBM scores were good predictors of MCA performance, and the later R-CBM scores gathered in 3rd grade were better predictors of the MCA test results than those collected earlier in the study period. Also, the predictions of the R-CBM score for a given semester based on the R-CBM scores from the preceding semester were very strong. All three statistical techniques produced cut scores that offered acceptable accuracy, but the authors indicated that the logistic regression models created with a single R-CBM predictor variable were more parsimonious than the models created with discriminative analysis. The specificity and negative predictive power of the models appeared to be stronger than the sensitivity or positive predictive power. The authors suggested that using R-

CBM scores would allow educators to identify students that had skill deficits such that an intervention program could be offered to them. However, there was no discussion of how the evaluation measures at a single cut score could be used to adjust the models to produce greater accuracy. There was also no indication that separate random samples of the 1,766 students were used to fit and test the models.

In 2008 Brasier[51] utilized the NELS:88 dataset to examine the effects of parental involvement on children's aspirations to complete college. Logistic regression models were constructed to predict how far in school students expected to get based on their parents' level of involvement in terms of discussing school events, topics studied, and selection of courses. The models also included variables that controlled for race/ethnicity, gender, maternal expectations, paternal expectations, socioeconomic status quartile, and the students' academic capabilities. The models were fitted with records from 9,707 students that participated in the 1988 and 1990 waves of data collection and provided complete responses to the questions that were used to obtain the variables. The college aspiration variables were recoded to have a binary scale with 0 indicating a low level and 1 indicating a high level of aspiration to complete a college degree. The covariates were recoded to have a binary or ordinal scale as well. Principle factor analysis (PFA) was employed to distill the three measures of parental involvement into a single covariate. A probability threshold of 0.50 was used to divide the estimated probability of high aspiration into a prediction of high vs. low aspiration. Separate logistic regression models were constructed for the 8^{th} grade and the 10^{th} grade using only the covariates collected in 1988 and 1990, respectively. Two models were created for each of the grade levels with one including parental involvement as a factor and one without this variable. This research design allowed the analyst to determine if parental involvement significantly contributed to the prediction of college

aspiration in each grade and if the contribution of parental involvement varied between the grades. The observed vs. predicted aspiration levels were compared and an overall percentage of correct high and low aspiration level predictions was calculated for each model. The area under the ROC curve for each of the four logistic regression models was among the different statistics used to assess their strength.

Brasier found that the models strongly predicted college aspiration levels and that parental involvement was a significant predictor in both grades. However, the contribution of parental involvement to the students' aspiration level was significantly stronger in the 8th grade model than in the 10th model. Male students were less likely to have high college aspirations than their female counterparts in both grades. Being African-American was found to be a significant predictor of having high aspirations in both grades while being Hispanic was significant only in the 8th grade and being Asian was found to be significant only in the 10th grade. Higher academic ability, socioeconomic status, maternal expectations, and paternal expectations were significant predictors of higher aspiration levels in both grades. These results led to a conclusion that parental involvement and children's college aspirations had a dynamic relationship subject to change as the children developed and that encouragement from parents was very important in students deciding to pursue post-secondary education.

The main differences between the methodology developed in this dissertation and the analysis by Brasier are the focus on college aspirations vs. final educational outcomes and the objective of significance testing vs. outcome prediction. Brasier was exploring a portion of the students' educational trajectory towards college in order to determine if parental involvement was a significant predictor of college aspirations. The logistic regression models were developed using the entire set of 9,707 students in the sample developed by Brasier as opposed to selecting

multiple random sub-samples of the records for repeated model fitting and testing. Brasier's goal was to determine the extent to which parental involvement could predict student aspirations to finish college. Predicting the level of college aspirations was a means to evaluate the models' findings of significance rather than the ultimate objective of creating the models. The set of potential covariates used by Brasier was carefully chosen and meticulously tested for inclusion in the modeling process, but it was much smaller than that chosen for this dissertation.

Huang, Taddese, and Walter[52] analyzed the factors that affect female and minority students' pursuit, persistence, and completion of postsecondary science and engineering degrees. The researchers analyzed two datasets to test groups of variables for their relative importance in the continuing technical education gaps exhibited by female and minority students. NELS:88 data from the base year of 1988 through the third follow-up in 1994 was analyzed to study variables that affected female and minority students through high school and entry into a technical field of study. Data from the Beginning Postsecondary Student Longitudinal Study (BPS) covering a five-year period was analyzed to study variables that affected students during the pursuit of a technical degree. This paper was particularly informative for the purposes of this research since it focused on the issue of persistence in acquiring an engineering or science degree in college.

Huang, et al. drew a sample from the NELS data which included students that did not attend college since this outcome represented a portion of the achievement gaps and not including these students could have led to underestimation of the gaps. Descriptive analysis was utilized first to decide which potential variables showed an association with selecting a technical major. The results of this analysis suggested that female students experienced primarily psychological disadvantages in pursuing a technical field of study while minority students were

hampered by lesser educational opportunities as well as psychological disadvantages. The psychological disadvantage stemmed from comparatively fewer female/minority students wanting to study technical subjects in college.

Multivariable logistic regression analysis was used by Huang, et al. to determine the relative importance of the factors identified with descriptive analysis. The dependent variable was the enrollment in science or engineering in college of a student from the NELS dataset. Among the specific independent variables tested were the students' personal interest in science, goals to work in a technical career, participation in gifted/advanced programs, total credits in science and math, total advanced credits in science and math, teachers' major/minor in science or math, and the schools' science credits requirement. Other school characteristics such as minority enrollment percentage, private vs. public school, and urban vs. rural setting were considered but not included since they were deemed to have a lesser impact on selection of a college major.

Many of the findings by Huang, et al. were consistent with prior research in that students with greater academic skills particularly in science and mathematics were more likely to pursue a college major in a technical field. Students that had a greater personal interest or motivation to study a technical subject were more likely to major in one. Having a supportive family environment and high parental expectations were also significant factors. Once these factors were controlled, the racial/ethnic and gender gaps in pursuit of a technical degree narrowed. The study of persistence and achievement of a postsecondary degree indicated that the racial/ethnic gaps tended to reappear as minority students experienced more difficulty in completing the technical degree. This was not found in female students who tended to outperform male students in completing their studies. So although female students did not select technical majors in comparable numbers to male students, they performed well once in the program.

The Huang, et al. study summarized the history of the gender and racial/ethnic gaps in science and engineering fields including progress made in narrowing the gaps. Is also outlined prior research and highlighted numerous factors that have been found significant in the past. The factors fell into three broad classes: general family environment and support for the student; student characteristics including attitudes, personal goals, and academic capability; and institution characteristics including financial aid, special programs to promote entry/retention, and precollege coursework. These classes were the starting point of the multivariate models used to analyze the NELS and BPS datasets.

The Huang, et al. family environment and support class included variables that record parents' aspirations for their children's education when the children are in 8th and 12th grade. The responses included completing high school, attending a vocational/trade/business school, attending a 2-4 year college, finishing a 2/4-5 year college program, or earning an advanced degree. Other variables explored how parents supported their children's interest in science or math by taking them to museums, saving money for college expenses, and discussing education and post-secondary studies with their children.

The Huang, et al. student actions class of variables measured the aspirations of students, what programs they participated in, the coursework they studied, the strategies they developed to advance learning, and how they performed in math and science. The aspirations were measured by asking students how far they expected to advance in their education and whether or not they expected to be working in a technical subject area at the age of 30.

The program participation and coursework were measured by asking students if they had ever been in a special program for college preparation, advanced placement, or high ability. This question was asked in the 8th, 10th, and 12th grade surveys. Transcript data was collected that

indicated the number of credits taken in subjects such as math, science, computer science programming/data programming, English, foreign language, and social studies. Remedial course-taking was explored to further evaluate the students' capabilities in the core subject areas. Students were also asked in 8th and 10th grade if they had participated in any special math/science enrichment programs.

The attitudinal and learning strategy variables were provided by survey questions that asked students why they took math and science classes, whether or not they did their homework in math and science, whether or not they adopted certain learning strategies, and how they would assess their confidence in their level of science/math education. Performance in math and science was assessed through standardized proficiency tests that the students surveyed took in 8th, 10th, and 12th grade.

The Huang, et al. institution characteristics class of variables examined the environment in which the students were educated. It included questions about the availability of computer labs, advanced placement or college-level math/science classes, and the graduation requirements of math/science coursework. The credentials of the teachers in the high school were assessed by asking the surveyed schools to identify what percentage of their math/science teachers majored or minored in the subject they teach.

The main difference between the methodology developed in this dissertation and the analysis by Huang, Taddese, and Walter is the exploration of persistence using NELS:88 data as opposed to BPS data. The multivariable logistic regression employed by Huang, Taddese, and Walter to identify significant predictors of a technical major was explored in this dissertation as the current standard of practice. However, the integrated methodology employs survival analysis to examine additional data in order to predict the probability of an individual student with a

given vector of variable values departing the STEM track and not "surviving" to achieve a STEM degree. This is a more powerful analytical technique that builds upon the results of the logistic regression analysis of significant predictors of persistence.

In summary, the prior education research indicates that certain variables consistently predict statistically significant differences between students who pursue a STEM degree and those who do not. Among the most consistent variables are gender, race/ethnicity, average high school grades, SAT math scores, personal interest in STEM subjects, and students' self-confidence in their mathematics skills. Other variables which were found to be significant in some of the studies include having a desire to make a theoretical contribution to science; having a parent employed in a STEM career, attending a school where the student's average high school grades and SAT math scores were comparable to student body averages; and possessing a higher self-concept and sense of control. Additional variables that were significant in a few of the studies were socioeconomic status, SAT verbal scores, and interest in social activities.

2.5 PRIOR RESEARCH UTILIZING SURVIVAL ANALYSIS

Since education is a process that requires many years to ultimately achieve a goal, researchers have sought means to acquire and analyze longitudinal data measuring the same students at successive points in time. Survival analysis techniques have been applied to longitudinal data in order to identify factors predictive of students ultimately experiencing a general event of interest.

Mensch and Kandel[53] utilized the National Longitudinal Survey of Youth (NLSY) dataset to analyze the impact of drug involvement upon dropping out of high school. NCES began NLSY in 1979 with surveys of 12,684 people aged 14-21 and continued with annual

surveys until 1994. NLSY data from the 1984 follow up survey when the students were aged 19-27 was utilized in the analysis. Survival analysis techniques were used to create hazard models of the risk of dropping out or achieving a GED as a function of various independent variables. These independent variables included demographic, behavioral, experiential, and personal attribute characteristics. The models were discrete-time and assumed the risk of event occurrence was constant within a single year time period. Logistic regression was used to estimate the discrete-time hazard models.

The results indicated that the risk of dropping out increased with drug use. The earlier in life drug use began the greater the risk of dropping out. The same was found with girls for sexual intercourse and pregnancy, and the effect was even stronger than that of early drug use. For both genders, use of more disfavored substances was more strongly associated with dropping out. For example, use of illicit drugs had a greater association than use of alcohol or cigarettes. Among illicit drugs, use of marijuana had a weaker association with dropping out than use of other presumably harder drugs. Lesser parental education, broken family status, minority group membership, lower academic capability, lower self esteem, and having an external locus of control indicated a greater risk of dropping out. Of all the variables, academic capability was the strongest predictor of dropping out. With respect to the "risk" of acquiring a GED, the strongest predictors were academic capability and parental education. The analysis of female students found two other variables to be significant in predicting GED acquisition: race and self-esteem. When the models controlled for academic capability and parental education, nonwhite females were more likely to get a GED than white females. Females with low self-esteem were less likely to achieve a GED than those with higher self-esteem.

Civian[54] used survival analysis techniques to explore the time to complete a doctorate at the Harvard University Graduate School of Education (HGSE). Data was collected about 625 full-time students who matriculated at HGSE between Fall 1982 and Fall 1988. Proportional hazards models were constructed to look for significant differences in groups that varied by factors including gender, race/ethnicity, citizenship, year cohort, doctoral program, and estimated academic ability. Significant differences were identified with foreign students in two of the three programs completing their doctorates sooner than American students. Caucasian students required more time to complete their degrees than non-Caucasian students less than 30 years of age did. Older non-Caucasian students took longer than Caucasians to graduate. Students with higher achievement test scores completed their degrees a bit faster than the students with lower test scores.

Willett and Singer[55] stated that educational researchers should employ survival analysis techniques in order to study topics such as student persistence and teacher attrition. The article maintained that one of the best reasons to apply survival analysis is that standard statistical techniques require knowledge of when the event occurred (the outcome) for each sample member. This is a standard unlikely to be met in studying event times. Regardless of the length of the study, it is probable that some sample members will not experience the event of interest prior to the end of data collection. For example, a study of student drop out behavior may follow students from the 8th grade through the time at which they should have graduated high school. Those students who graduated without dropping out did not experience the event of interest so their event times are referred to as "censored." A student who transferred prior to graduation or dropping out also has a censored event time because the competing risk of transferring occurred first. Survival analysis methods can take the censored event times into account when examining

the probability of the event of interest occurring by a set time. This permits an analyst to estimate when the probability of the event happening is at its greatest.

Willett and Singer analyzed teacher attrition data with survival analysis methods. Their findings indicated the greatest risk of leaving the profession is in the first few years, and that throughout a sample over 12 years the risk of leaving was higher at every point for female teachers. Statistical models were used to estimate the discrete-time hazard function for teachers. A logit transformation approach was used to test various models with several predictor variables including gender, age when hired, a cross-product term of gender and age at hire, and salary. The first three variables tested were examples of time-invariant measures that remained constant throughout the sample period. Salary is an example of a time-variant variable since salary levels changed over the sample period. Time-varying variables offer a more insightful measurement of the sensitivity of the hazard or survival functions to changing conditions. In the case of the teachers, this allowed the analysts to explore the effects of a variable whose impact on the hazard function changed with time. These more complex models offered a way to pinpoint when a category of teachers is most apt to consider leaving the profession and potentially to intervene positively. Traditional statistical methods are less effective in modeling the effects of time-varying predictors.

Willett and Singer also examined the applicability of survival analysis on student dropout tendencies by developing hazard models to identify when students are most likely to leave school. Survival analysis was favored for its ability to model different competing risks including graduation, dropping out, changing schools, being expelled, stopping out before returning, etc. throughout students' educational careers. The findings of a grade specific profile of dropping out indicated that the rate of dropping out is below 1% for grades K-7th, rises to a peak in 10th

grade, and then drops in 11th and 12th grade. This suggests that anti-dropout programs are needed most as students approach 10th grade.

Vegas, Murnane, & Willett[56] examined the conditional probability of high school students continuing their education and entering the teaching profession to determine if race/ethnicity, gender, or academic skills were significant predictors. Longitudinal data was obtained from the HS&B survey that followed sophomore students from 1980 through the year 1992. A series of four hierarchical samples were drawn from the dataset starting with 10,584 sophomores in 1980. Subsequent samples were drawn from all members of the preceding sample that matched the criteria for continuing on the path to teaching. These samples were used to determine the conditional probability of a student graduating from high school, entering college, graduating from college, and entering the teaching profession. The percentages of males and females of different racial/ethnic categories that achieved each step were compared to see if some groups were less likely to proceed than others. Logistic regression models were constructed to determine if academic skills were significant predictors of the different racial/ethnic and gender groups completing each step.

The findings of Vegas, et al. indicated that academic skills in tenth grade explained most of the differences in high school graduation probability among the racial/ethnic groups. Female non-majority students were more likely to matriculate than their male counterparts. Conversely, male majority students were more likely to matriculate than their female counterparts. Minority students with good academic skills tended to enroll in college after graduating from high school. Unfortunately, many minority students in the study that graduated high school lacked strong academic skills. The results indicated a dramatic variation in the probability of graduating from college. Asian-Americans followed by Caucasians had the highest rate of graduation. The rates

for Hispanics, African-Americans, and Native Americans were far lower and declined in that order. However, the graduation rates were very similar across the racial/ethnic groups when considering students of the same academic skill level. This suggested the graduation rate differentials would have been greatly reduced if minority students were better prepared for college. Native Americans, African-Americans, and Hispanics, respectively had the largest percentages of college graduates entering the teaching profession. Asian Americans had the lowest percentage. Across all the racial/ethnic groups, females were far more likely to start their career in teaching than their male counterparts.

The prior education research indicates that the use of survival analysis techniques can be quite powerful in modeling educational event occurrences. The ability to test time-varying predictors as well as time invariant predictors is a particularly valuable benefit of applying survival analysis techniques. The research to date has employed single statistical techniques or a series of nonintegrated single techniques to explore these complex problems. This limits the degree of insight that can be obtained and the potential for decision-making about intervention methods. Ideally, analysis should be able to pinpoint the most critical time to initiate educational interventions as well as the set of predictors that describe which students would benefit most.

2.6 RECEIVER OPERATING CHARACTERISTICS CURVE ANALYSIS

Developing a model to predict between a STEM outcome vs. another outcome for a given student involves using data to discriminate between the two potential results. A valuable tool in assessing the accuracy of the discrimination is Receiver Operating Characteristics (ROC) curve[57] analysis. ROC curve analysis was developed as a concept in signal detection theory during

World War II where radar operators examined radar signals to detect oncoming Japanese aircraft and distinguish such readings from "noise" in the signal. The goal was to increase the accuracy of predictions and decrease the likelihood of false alarms or missed detections. The prediction accuracy is a tradeoff between sensitivity and specificity. Sensitivity is the probability of correctly identifying a signal while specificity is the probability of correctly identifying system "noise." In terms of STEM prediction the sensitivity is the probability of correctly classifying a student as having a STEM outcome. The specificity is the probability of correctly classifying a student as having a "Not-STEM" outcome.

Classifying a student outcome as STEM vs. Not-STEM is based upon the value of a prediction threshold. Consider a threshold value between [0, 1] where a prediction represents the probability of a STEM outcome. The threshold value or "cutpoint" determines which of two outcomes the model predicts. If the cutpoint is set to 0.5 then records for which the model estimates a probability of a STEM outcome ≥ 0.5 will be classified as a STEM prediction.

Records for which the model estimates a probability of a STEM outcome < 0.5 will be classified as a Not-STEM prediction. If the cutpoint is set to a low value, then the model will predict more students to have a STEM outcome. As a result more true STEM students will be correctly predicted to have a STEM outcome but correspondingly, more true Not-STEM students will be incorrectly predicted to have a STEM outcome. If the cutpoint is set to a high value, then fewer students will reach that value and be predicted to have a STEM outcome. Thus fewer true STEM students will be correctly predicted to have a STEM outcome and correspondingly fewer true Not-STEM students will be incorrectly predicted to have a STEM outcome. The choice of the cutpoint value determines the results in discriminating between the two potential outcomes.

While this type of analysis was developed in the refinement of radar signal detection it has been adapted in other fields to evaluate how well models discriminate between potential outcomes. It has wide usage in medical research to evaluate the diagnostic value of medical tests[58, 59], determine the therapeutic value of treatments, and to make decisions in interpreting radiology images[60, 61]. A medical test may result in concluding that a disease is present (a "positive" test result) or that it is not present (a "negative" test result). Ideally, a diagnostic test should accurately detect when a disease is present and accurately indicate when it is not. False positive test results lead to unwarranted concern and potentially unnecessary treatment while false negative test results may lead to adverse health results as a condition goes untreated.

Discriminating between two outcomes leads to four possible results. The test could classify a result as positive or negative. The classification could be correct or incorrect. The four possible results are correct disease detection (true positive), incorrect disease detection (false positive), correct healthy status (true negative), and incorrect healthy status (false negative). The sensitivity of a diagnostic test measures its ability to identify the presence of disease and the specificity of the test measures its ability to identify the absence of disease. In this context, sensitivity is the probability of a true positive while specificity is the probability of a true negative. The cutpoint used in the outcome discrimination determines the sensitivity vs. (1 - specificity) for the diagnostic test. The two measures are directly associated for a given cutpoint value. This association means there is a tradeoff between achieving good sensitivity and good specificity in outcome discrimination.

ROC curves visually depict the tradeoff by plotting sensitivity vs. (1 – specificity) for a range of cutpoint values. Plotting (1 – specificity) on the horizontal axis and sensitivity on the vertical axis produces a curve line. Ideally, the ROC curve should resemble a vertical line at a

low value of (1 – specificity) which transitions to a horizontal line over the remaining values of (1 – specificity). ROC curves with a steep gradient mean that a very high value of sensitivity with a correspondingly low value of (1 – specificity) is attainable. They indicate that a cutpoint can be chosen which offers a high probability of true positive detection and correspondingly low probability of false positive prediction. If an ROC curve contained the point (0, 1) it would indicate the model could be set to a cutpoint that would provide perfect predictive ability. Figure 2.1 illustrates an ROC curve with a steep gradient. A cutpoint for the model can be chosen to discriminate between outcomes with 80% sensitivity and a corresponding (1 – specificity) value of approximately 15%. This curve indicates the model could correctly classify 80% of the true STEM students with just 15% of the true Not-STEM students incorrectly predicted to have STEM outcomes.

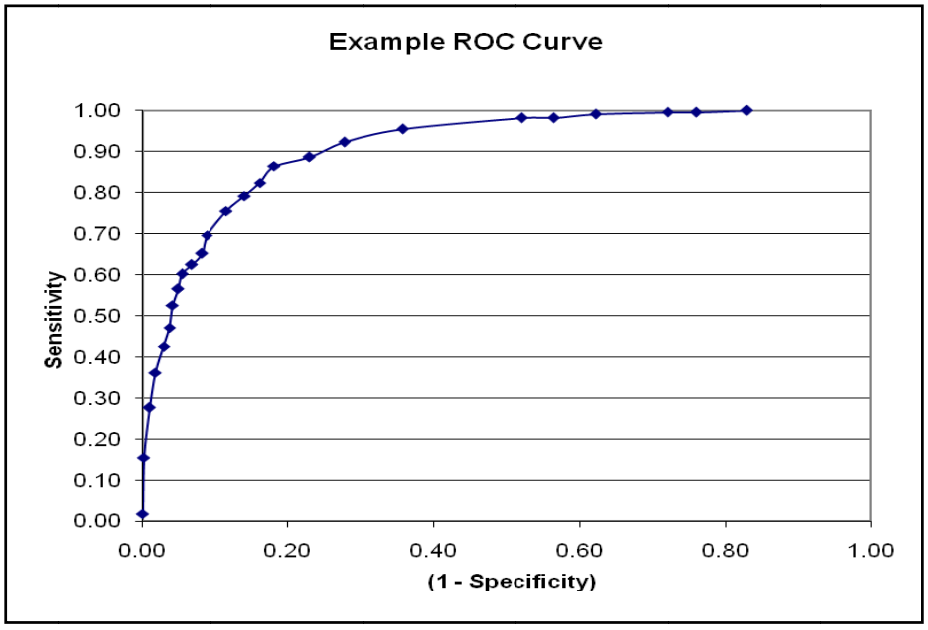

Figure 2.1 Example of an ROC curve

ROC curve analysis allows prediction models to be evaluated by examining the resulting ROC curve for a wide range of cutpoints. Models which result in an ideally shaped ROC curve have better ability to discriminate between two potential outcomes than those with a flatter ROC curve. A shallow ROC curve that resembles a 45 degree line between the axes implies that the model has negligible discrimination value. Such a model is as likely to predict a true positive as a false positive and has no useful predictive ability. A visual estimate of the model's predictive ability may be gained by comparing the ROC curve to a 45 degree line and determining how much space lies between the two curves.

The area under the ROC curve ("AUC" or "*c*") provides an estimate of the model's predictive ability. The sensitivity and (1 – specificity) range from 0 to 1 so the ROC curve is plotted within the unit square formed by the points (0, 0), (0, 1), (1, 0), and (1, 1). Therefore, the area under the ROC curve is a portion of the unit's square area of 1, and *c* is a value between 0 and 1. A model with a high value for AUC is judged to better discriminate between the outcomes. Hosmer and Lemeshow[62] indicate that a value for AUC of 0.5 indicates that the model is of little use since it is as likely to correctly predict a binary outcome as flipping a fair coin. A result of $0.7 \leq AUC < 0.8$ represents "acceptable" ability to discriminate between potential outcomes. A result of $0.8 \leq AUC < 0.9$ represents "excellent" ability to discriminate between potential outcomes. A result of $AUC \geq 0.9$ represents "outstanding" ability to discriminate between potential outcomes.

The ROC curve can be used to improve a predictive model by selecting a cutpoint for outcome discrimination that provides a good tradeoff between sensitivity and specificity. The selection of a preferred cutpoint to use in discriminating between outcomes is based on the objectives of the analyst in developing the model. If the goal is to optimize sensitivity and

specificity then both are plotted against the range of cutpoint values and the cutpoint value at the intersection of the curves is selected. If the goal is to maximize the correct prediction of the outcome of interest, then the cutpoint can be chosen without regard to the probability of incorrect predictions. If the goal is to maximize the correct prediction of the outcome of interest subject to a constraint, then the cutpoint may be chosen to achieve the highest sensitivity probability without violating the constraint. For example, the constraint may be budgetary. If a cost is associated with false positive predictions, then the cutpoint is limited by the probability of a false positive applied to the population of interest.

3.0 DATA SOURCE REVIEW

3.1 INTRODUCTION

The literature review revealed that there are two major sources of large size educational datasets, the longitudinal studies by NCES and the CIRP surveys of college students gathered across the nation by HERI. The NELS 88:00 dataset and the CIRP survey of incoming freshmen are of particular interest. The following sections outline each dataset to explain the data offered, what types of variables are provided, the time period covered, and how it has been utilized in this research. The CIRP dataset provides the capability of examining more students at specific colleges around the nation, but offers a less rich assortment of variables. It covers a narrower period of time in the development of the students. Its main application for this research was to provide insight into the selection of potential variables to be tested for significance in predicting educational outcomes.

The NELS dataset is so extensive in size and scope that understanding what it offers requires serious study. Section 3.3 describes the design used in the data collection; explains how the sample was developed and refined over time; defines the coding scheme used in naming variables; and indicates how researchers gain access to the full dataset. Section 3.4 explores the reasoning that led to the selection of NELS as the primary dataset used in this. Section 3.5 briefly covers testing of the statistical software selected for this analysis to ensure it could be

applied to this dataset. The testing was done to confirm that the software code generated by the interactive user interface provided with the NELS dataset to import and format the data could be used with modest edits rather than writing entirely new code.

3.2 CIRP DATA REVIEW

The selection of promising variables to explore started by examining CIRP variables that had been valuable in prior studies of STEM students. The results of other researchers as well as prior personal experience in analyzing the CIRP data were utilized. The results of analyzing CIRP data[63] for two universities' incoming freshmen found 22 variables were consistent in predicting between STEM and Non-STEM majors for at least 5 of 7 gender/ethnic sub-groupings of the students. The sub-groups included Caucasian students at both schools, African-American students at both schools, female students at both schools, and Hispanic/Latino students at one of the schools. Among the most consistent CIRP variables that predicted an intention to major in STEM were the students' SAT Math scores, self-rating of math ability, personal goal to contribute to Science, self-rating of computer skills, self-rating of academic ability, high school grade point average, time spent per week playing video/computer games, and decision to attend college to obtain special training for a specific career. Other variables were found to be consistent in predicting an intention to major in a Non-STEM topic. These variables included the students having a goal to influence social values; anticipating potential changes of major and/or career; choosing a college based upon its size or social activities; intending to participate in student government; intending to study abroad; having a goal to influence politics, having a goal to be a community leader and/or be in a position of administrative responsibility; wanting

to create artistic work; wanting to understand foreign countries and cultures; and spending a larger portion of discretionary time attending parties.

These findings suggested the most promising NELS:88 variables to explore in initial modeling were those that measured academic ability, personal interests, personal attitudes, and future career goals. The NELS:88 dataset includes numerous variables that assess mathematics abilities, prior performance in science classes, attitudes about math/science, and overall high school grade achievement. The dataset includes SAT scores, indications of future career intentions, and measures of interest in various academic/social activities.

3.3 NELS DATA REVIEW

The NELS data was collected in five waves starting with the base year (BY) study in 1988 designed to assess the students' situation and academic strength in 8th grade before they actually entered high school[64]. Specialized schools including those for students with disabilities, vocational schools, Department of Defense Dependents' schools, and Bureau of Indian Affairs schools were excluded from the study. Home schooled or privately tutored students were also excluded as well as students that had dropped out prior to the 8th grade. The students in the survey were gathered in a stratified national sample of 1,052 schools teaching 8th grade. Both public and private schools were included and clusters of students were studied at the same school. Schools identified students that were considered unable to fully participate in the study due to severe disabilities or a lack of English language proficiency. These students represented 5.3% of the potential 8th grade sample and were classified as ineligible for the study. Initially, 26,432 students were selected for the sample and 24,599 ultimately participated in the base year.

Cognitive tests were conducted in the base year to assess the students' strengths in reading comprehension, mathematics, science, and social studies. Questionnaires were also administered to the students, parents, teachers, and school administrators. These instruments were used to gather information about the students' including their academic performance, socioeconomic status, family structure, home environment, future educational/vocational plans, high school environment, personal impressions, parental impressions, teacher impressions, post-secondary educational experiences, post-school employment, etc.

Follow up data collection was done in successive waves in 1990, 1992, 1994, and 2000. These were classified as the first follow up (F1), second follow up (F2), third follow up (F3), and fourth follow up (F4), respectively. Cognitive tests were administered again in 1990 and 1992. Various questionnaires were utilized to measure the students' continuing development. Even school drop-outs were surveyed to learn if they had obtained a GED or other certification and the reasons they left school. For those who remained in school, NELS collected data from the students' transcripts in high school and college.

3.3.1 F1 Sample Size

The first follow up[65] started with the BY sample size of 26,432. A total of 96 BY students were excluded from F1 because they had moved out of the U.S. or died. These students were classified as "out of scope." The 348 students that had dropped out of school between the BY and F1 were automatically retained in the sample to maximize the dropout subgroup for study.

It was determined that the students had spread out to a vastly increased number of schools (3,736) by 1990. In addition, the schools attended by another 221 students could not be determined and 10 students were then being home-schooled. These cases were handled by

creating individual "schools" for the students in question. This increased the total number of potential schools to 3,967. To reduce administrative and data collection costs the remaining BY in school sample size of 25,988 was reduced in F1 (1990) to 21,474 students through proportional sub-sampling that decreased the number of schools participating in NELS. Students attending a school with ten or more of the BY sample members attending were automatically selected for inclusion. Students attending a school with less than ten BY sample members were included with a lower probability. The probability of inclusion increased with the number of BY sample members in the school. The sampling design was based on an algorithm that balanced the budgetary costs against the desire to retain the maximum number of students. This resulted in a dramatic reduction in the number of schools being surveyed from 3,736 to 1,500.

Another 1,229 students were added to the sample in a process referred to as "freshening." This was done as required in later years to maintain the sample's representative quality for the U.S. population of sophomores (1990) and seniors (1992). A second sub-sampling was then undertaken for students in two groups: (1) 1,991 transfer students that were no longer in the school identified in the first sub-sampling; and (2) 742 potential dropouts that had not been available during the first follow up survey days at their schools. The transfer students were sampled based on a 20% probability and the potential dropouts at a 50% probability. There were 386 transfer students and 357 potential dropouts that were retained in the sample. Of the 357 potential dropouts, 75 were ultimately found to be true dropouts. Another 7 BY students were excluded from the sample since they were later found to have been selected in error. This brought the F1 potential sample size to 20,706.

In addition, 340 students originally classified as Base Year Ineligible were reclassified as eligible and included in the potential sample for a total potential sample of 21,019. The sample

was further reduced by excluding 27 students due to sampling errors or out of scope status for a total potential sample of 21,026. The final number of students that participated in F1 was 19,264 including 18,221 students in school and 1,043 dropouts.

3.3.2 F2 Sample Size

The second follow up[66] in 1992 began with a series of design goals. These included having a representative probability sample of the 1991-1992 school year's senior enrollment; keeping the maximum number of Asian, Hispanic, and American-Indian students from the F1 sample; keeping all of the dropped out students in the sample; retaining all F1 nonrespondents in the sample to minimize the potential for nonresponse bias; and limiting the sample to students at 1,500 schools to minimize the costs of gathering teacher reviews, administrator reviews, and transcript data. Some of these goals proved to be contradictory. Excluding 1,564 "known" dropouts the number of schools attended by the eligible sample of students was 3,224. Attempts to whittle down the number of schools would have greatly limited the ability to retain a disproportionate number of the racial/ethnic minority students. The study organizers compromised by including the same potential sample of 20,747 students that were sought in the first follow up in 1990. However, the contextual data including administrator surveys, teacher surveys, and transcripts were obtained only for students at 1,500 schools[67]. The 1,500 schools were chosen by selecting all of the schools that had at least four F1 sample members enrolled (1,030), a random sample of the schools that had just one F1 sample member (321 of 1,008 schools), a random sample of the schools with two F1 sample members (104 of 160 schools), and a random sample of the schools that had three F1 sample members (45 of 60 schools).

The F2 sample was then "freshened" to ensure that it represented a valid probability sample of senior students. Students with severe mental, physical, or linguistic disabilities to participation were still excluded. Students that could not meaningfully complete the cognitive tests or survey questionnaires in English but could in Spanish-language versions were considered to be eligible for the survey. There were 366 students initially added in the freshening process. Of these 288 were found to be eligible for the cohort and 266 were deemed eligible for F2. Of these students, 22 were later ruled out as ineligible with 1 excluded for language issues, 8 excluded for mental or physical disabilities, and 13 excluded due to moving out of the country. A net total of 244 students were ultimately added to the potential F2 sample through freshening. As a result, the final sample size was 20,923 with 18,209 members in school and 2,714 dropouts[68]. The actual number of students that participated in F2 was 19,264. This figure included 16,842 students in school and 2,378 dropouts.

3.3.3 F3 Sample Size

The third follow up[69] in 1994 was based upon the potential sample from the second follow up in 1992. The sample was broken into 18 categories based on the sample members' race, socioeconomic status, test scores, attendance at a private school in prior follow ups, nonresponse pattern, dropout status, freshening status, and eligibility to participate. Table 3.1 summarizes the categories, lists the selection probability for each member in a category, and indicates the number of members in the group ultimately selected for inclusion in the F3 sample.

Ultimately, this set of students provided a sample size of 15,964. A further 89 students were excluded, and the final sample of 15,875 students was selected for the third follow up. Of this sample, 14,915 sample members[70] responded by completing the surveys.

Table 3.1 NELS:88/94 Sampling Groups' Selection for F3 Sample

Group	Description	Selection Probability	Number in F3 Sample
0	Excluded (Ineligible or out of scope in 1992; freshened students that dropped out prior to survey in the year freshened; or BY dropouts.)	0.00	0
1	Nonresponders (never completed a prior survey)	0.15	43
2	Poor Responders (didn't complete either F2 survey or survey in first eligible round)	0.25	596
3	Ever Dropped Out	1.00	2351
4	Ineligible prior to1992 (due to mental/physical disability or language barrier)	0.90	191
5	Private School in 1988	0.80	2,387
6	Private School in 1990 or 1992	0.90	98
7	Hispanic	1.00	1,466
8	Asian or Pacific Islander	1.00	874
9	Native American	1.00	132
10	Black – top quartile in cognitive tests	1.00	79
11	Black – other test scores	0.90	1,114
12	White – lowest Socioeconomic Status quartile	1.00	1,295
13	White – highest Socioeconomic Status quartile	0.60	1,522
14	White – middle Socioeconomic Status quartile	0.80	3,810
15	Freshened in F1 (1990)	0.30	1
16	Freshened in F2 (1992)	0.30	2
17	Other	0.40	3
Total			15,964

3.3.4 F4 Sample Size

The last follow up in the study was complicated by the fact that in the six years since the last follow up many sample members had left their previous addresses and migrated to other geographic areas[71]. Finding sample members to conduct the surveys was both critical and difficult. As with the third follow up a subsample was sought to ensure sufficient representation

of the different subpopulations while keeping data collection costs within acceptable limits. The potential sample members were grouped into strata based on their response history up to and including F3 and a set of domains of interest. The probability of including a stratum in the sample was assigned based on previous response rates and the reaction to inclusion. Eligible members who had refused in a hostile manner to participate were pursued at a much reduced subsampling rate of 0.05 to 0.15. Poor respondents who proved difficult to trace were also assigned a lower subsampling rate of 0.30 to 0.35. Eligible members who were easier to trace were included at a higher probability (0.60 to 1.00) even if they had a history of poor response. The subsampling rates were chosen by balancing two competing goals: minimize unequal weighting effects and sampling variances associated with different strata while minimizing survey costs.

The subsampling identified 15,236 subjects for the F4 sample of which 14,900 responded in F3 and 336 were nonrespondents. Of the 336 nonrespondents 14 were hostile refusals. A total of 12,144 sample subjects actually participated in F4.

3.3.5 Validity of F4 Sample vs. BY Sample

Since the final sample of 12,144 students that participated in the fourth follow up was notably reduced from the initial base year sample of 24,599 a question arose regarding the validity of the final sample in representing the initial sample. Since the goal was to determine if educational outcomes could be better predicted with an integrated model rather than one using a single statistical technique, the representativeness of the sample was not critical. The data was not intended to provide a probabilistic sample of the national student population from 1988 – 2000, and the final sample used in this research was not used to draw conclusions about the STEM vs.

other outcomes in the national population. The final sample was large enough to reflect each potential educational outcome and permit model development/evaluation.

3.3.6 Quantity of NELS:88/2000 Data

A second area of concern was the quantity of data available for the final sample. Each of the 12,144 students was included in the final sample because they were selected for and responded to the final wave of data collection. This final wave potentially included students that were not eligible for the base year, were added to the sample in freshening during F1 or F2, and/or did not respond during one of the previous follow ups. Given these facts it was important to ascertain exactly how many of the 12,144 students participated in all five waves of data collection from the base year to the fourth follow up. This was determined by examining the "universe" variables created to show the sample member status at various points in time during the study. F4UNIV1 indicates the sample members' status in each of the five data collection waves from the base year to the fourth follow up. F4UNI2A shows how the sample members entered the study including base year eligible, base year ineligible, first follow up freshening, or second follow up freshening. F4UNI2B is the base year status of the sample members including eligible, ineligible, or not applicable (freshened later). F4UNI2C is the status of the sample members in the first follow up. F4UNI2D is the status of the sample members in the second follow up. F4UNI2E is the status of the sample members in the third follow up.

The first universe variable, F4UNIV1, contains a series of alphanumeric codes that describe the status of a sample member at each point during the data collection. Table 3.2 contains the codes used and their meaning.

Table 3.2 F4UNIV1 Codes and Definitions for Sample Members' Status in Each Wave

Data Collection Wave	Potential Codes	Code Definition
Base Year (BY)	BYE	Base Year Eligible
	BYI	Base Year Ineligible
	BNA	Base Year Not Applicable
First Follow Up (F1)	F1A	1st Follow up, In-school, in grade
	F1B	1st Follow up, In-school, out-of-grade
	F1D	1st Follow up, Dropout
	F1I	1st Follow up, Ineligible
	F1X	1st Follow up, Out of Scope
	F1?	1st Follow up, Status Unknown
	F1FA	1st Follow up, Freshened, In school, in grade
	F1FI	1st Follow up, Freshened, Ineligible
	F1F?	1st Follow up, Freshened, Status Unknown
	1NA	1st Follow up Not Applicable
Second Follow Up (F2)	F2A	2nd Follow up, In-school, in grade
	F2B	2nd Follow up, In-school, out-of-grade
	F2D	2nd Follow up, Dropout
	F2?	2nd Follow up, Status Unknown
	F2FA	2nd Follow up, Freshened, In school, in grade
Third Follow Up (F3)	F3H	3rd Follow up, Received HS diploma
	F3G	3rd Follow up, Received GED/HS equivalent
	F3P	3rd Follow up, Pursuing GED/HS diploma
	F3N	3rd Follow up, Not pursuing GED/HS diploma
	F3?	3rd Follow up, Status Unknown
Fourth Follow Up (F4)	F4Q	4th Follow up Respondent

The sample members that were surveyed in each of the five data collection waves were those coded as BYE, F1A/F1B/F1D, F2A/F2B/F2D, and F3H/F3G/F3P/F3N. Since the NELS:88/2000 data only included those sample members that were surveyed in the fourth follow up, the final code was only F4Q. The full set of status codes are shown in Table A.1 in Appendix A. Table A.2 in Appendix A contains the frequency of all observed combinations of the codes above in the NELS:88/2000 data and how they were flagged as having participated in all five rounds or ever having been in drop out status. Of the 12,144 sample members from the fourth follow up a total of 11,328 participated in all five waves of data collection. This is more

than 93% of the final sample. Of these sample members 1,488 indicated that they had at one time dropped out of high school.

3.4 SELECTION OF NELS DATA FOR THE DISSERTATION

The NELS dataset offers a very large volume of data including numerous demographic, attitudinal, performance, and outcome variables. The extensive volume, its longitudinal nature, and the wealth of prior educational research utilizing NELS made the dataset very attractive for this purpose.

Among the disadvantages of using the NELS:88 data was that the study was designed and data collected by other researchers whose purposes were different from the purpose of this research. Complete records were not available for each of the over 25,000 students that initially participated. The final sample contained just 12,144 students that were available and willing to be surveyed in final wave of data collection. This represented a loss of over half the initial sample for reasons that could not be studied.

Among the advantages of using the NELS:88 data was that the purpose of the study was to permit education researchers to examine then-current federal education policies and develop new policies. Analysis was anticipated on three levels: crosswave, cross-sectional, and cross-cohort. Examining the data to explore the relative significance of factors in affecting students' leaving "the STEM track" was a legitimate application for this data. The extensive nature of the data permitted the study of many factors that could not be captured in a narrower study.

Of the longitudinal high school/college/career studies available from NCES, the NELS:88 data was the most recent study that was complete. As previously discussed, data from

the Educational Longitudinal Study of 2002 was more current, but the study was not sufficiently advanced in time for this research. Prior NCES longitudinal studies were complete, but they reflected much earlier trends in education.

Most of the NELS data is publicly available free of charge to researchers. Transcript data, standardized test scores, and other more sensitive information are available only through an application to use the restricted dataset rather than the free public use dataset. An application to obtain the full dataset for this research was approved in 2003.

3.5 TESTING SOFTWARE FOR ANALYSIS OF NELS DATA

The ability to access the NELS data and perform analysis upon it was tested by using the accompanying interface software to export raw data for a set of variables in two ways. First, specific variables were identified and exported to a file readable by SAS™. The interface software produced a SAS program file that extracted the data for these variables from the dataset on the NELS:88 CD-ROM, performed standard descriptive statistics such as frequency and percentages, and created tables with this information. Second, the same variables were selected and exported to a MS Access-readable file. The interface software extracted the raw data for the selected variables and exported them as a comma separated text file. The text file was then imported manually to a MS Access database.

The variables chosen for this test were SEX, RACE, F2SEX, F3SEX, F3RACE, F4SEX, and F4RACE2. These are categorical responses to questions about gender and race obtained at various time points throughout the NELS study. The timing of the variables' collection is shown in Table 3.3.

Table 3.3 Description of Variables Tested for Analytical Method

Variable	Description	Timing
SEX	Composite gender of the student. This variable is obtained from student questionnaire, school roster, interpretation of first name, or imputed randomly (in that order).	Base Year (1988)
RACE	Composite race of the student. Derived from a question about race on the student questionnaire.	Base Year (1988)
F2SEX	Composite gender of the student. This variable is based on the first follow-up (F1SEX) composite variable and supplemented by the second follow up survey. If the data is missing in F1SEX and the second follow up the variable is obtained by imputing it from the students' first names.	2nd Follow up (1992)
F3SEX	Composite gender. This variable is equal to F2SEX unless a correction was noted in 1994.	3rd Follow up (1994)
F3RACE	Composite race. This variable is equal to F2RACE unless a correction was noted in 1994.	3rd Follow up (1994)
F4SEX	Gender of the student, derived from F2SEX	4th Follow up (2000)
F4RACE2	New definition of race – primary choice which offers greater detail about Hispanic/Latino ethnicity.	4th Follow up (2000)

The SAS program was edited to insert code that calculated the mean and standard deviation of each variable selected. Queries were written in the Access database to obtain the same statistics for these variables from the raw data. The calculated statistics were not individually meaningful in this analysis since all the variable values were categorical. These statistics were calculated solely to test whether or not the results obtained by analysis with SAS and Access were identical. The results of this analysis for each variable are provided in Table 3.4.

Table 3.4 Descriptive Statistics for Test Variables from SAS and Access

Variable	Value (Code)	Freq.	SAS Results		Access Results	
			Mean	Std. Dev.	Mean	Std. Dev.
SEX	Male (1) Female (2) Legitimate skip/not in wave (9)	5349 6035 760	2.0	1.9	2.0	1.9
RACE	Asian/Pacific Islander (1) Hispanic (2) Black Not Hispanic (3) White Not Hispanic (4) Amer Ind/AK Native (5) Missing (8) Legitimate skip/not in wave (9)	764 1444 1041 7908 117 110 760	3.8	1.7	3.8	1.7
F2SEX	Male (1) Female (2)	5782 6362	1.6	0.5	1.6	0.5
F3SEX	Male (1) Female (2) Legitimate skip/F3 nonresp. (9)	5710 6341 93	1.4	1.05	1.4	1.05
F3RACE	Asian/Pacific Islander (1) Hispanic (2) Black Not Hispanic (3) White Not Hispanic (4) Native American (5) Missing (6) Legitimate skip/ F3 nonresp. (9)	851 1610 1155 8264 161 10 93	3.3	1.5	3.3	1.5
F4SEX	Male (1) Female (2)	5782 6362	1.6	0.5	1.6	0.5
F4RACE2	Amer Ind or AK Native (1) Asian/Pacific Islander (2) Black Not Hispanic (3) White Not Hispanic (4) Hispanic or Latino (5) Missing (9)	131 712 1120 8203 1687 291	3.6	2.1	3.6	2.1

The results of the SAS program applied to the overall dataset and the Access queries applied to raw data imported from the overall dataset were identical. This supported the conclusion that editing the interface software-produced SAS code to analyze the data stored on the NELS CD-ROM produced the same results as performing the analysis on raw data extracted

from the CD-ROM. Based on this finding, a decision was made to proceed by selecting variables of interest from the NELS dataset, using the interface software to generate the initial SAS code to access the NELS variables, and editing the code to conduct the statistical modeling.

3.6 SELECTING AND PREPARING NELS VARIABLES FOR MODELING

As part of experimenting with the NELS data several multivariate linear regression models were tested and confirmed to be inadequate in predicting the students that obtained a STEM degree versus some other outcome. This outcome was expected since nearly all of the NELS data is categorical leading to violations in the standard assumptions relied upon in linear regression modeling. Logistic regression models were then developed to model the students' educational outcome as a binary (STEM or other) result.

The first attempts to create logistic regression models predicting a STEM vs. Not-STEM educational outcome using all of the potential variables for the students' three waves of high school data (BY, F1, and F2) were unsuccessful in producing strong logistic regression models. The volume of data was so great that it overwhelmed the software's modeling attempts. A strategic decision was made to focus mainly on base year (BY) variables obtained when the students were in 8^{th} grade since significant models at this stage would allow prediction at an earlier point in the students' educational career. In addition to the BY variables, a set of standardized test scores (SAT and ACT) obtained in the second follow-up (F2) were added as potential variables in the models. Even with the focus mainly on the BY period the set of potential variables was extensive. There were nearly 1,400 BY variables.

Difficulties were also encountered in modeling the outcomes based on so many variables with categorical values. Most of the NELS:88 variables were categorical with numeric values representing different potential responses to the survey questions. For example, the variable measuring a student's overall reading proficiency quartile from the cognitive test in the base year, "BY2XRQ," had seven potential values. These values are members of the set [1, 2, 3, 4, 6, 8, 9] and represented "Quartile 1 Low," "Quartile 2," "Quartile 3," "Quartile 4 High," "Legitimate Skip/Not in wave," "Missing," and "Test Not Completed," respectively. There were 1,180 records out of the original 12,144 that had a value other than 1 through 4 for the BY2XRQ variable. Since the potential values were not purely ordinal the variable's utility for model-fitting was hampered.

Other categorical variables had dichotomous responses but were not truly binary since the potential values were [1, 2] as opposed to [0, 1]. For example, the variable for students' sex, "F4SEX," was originally coded as 1 for male and 2 for female. Still other variables possessed a purely nominal set of potential values. The variable "F4Race2" had potential values within the set [1, 2, 3, 4, 5, -9] representing American Indian/Alaska Native, Asian/Pacific Islander, Black - not Hispanic, White – not Hispanic, Hispanic/Latino, or Missing, respectively.

The variables measuring standardized test scores for the SAT and ACT were categorical but resembled a series of mostly ordinal integer values. For example, the variable "F2RACTE" provided a student's ACT English score and included integer values of 6 through 36, 98, and 99. The first range of values represented an actual point score on the English portion of the ACT test while "98" indicated "Missing Data," and "99" represented "Legitimate Skip/Not in wave."

The difficulties caused by the categorical nature of the data were addressed by reviewing each variable's potential values and creating companion recoded variables that conveyed the

information in a strictly ordinal or binary fashion. This required a process of examining the potential values and recoding non-ordinal values such as those for "Missing," "Legitimate Skip," etc. to a value of "0" indicating that no useful information was provided by that variable for that individual record. Other responses were grouped to create stronger delineations between answers. An example of this was a recoded variable for the father's highest educational level that categorized the answer as either "College and above" or "No college degree."

Categorical variables with dichotomous potential values were recoded as needed to make them truly binary. Thus, a binary recoded version of the variable F4SEX variable, "F4SEXrb," was created with 0 for male and 1 for female. In such a case, the reference or base case was set to 0. Similarly, the nominal values for the variable measuring students' race, "F4Race2" was recoded into a set of binary dummy variables in which the base case was Caucasian with a value of 0. For these dummy variables "F4RACE2AI, " "F4RACE2As," "F4RACE2Bl," and "F4RACE2Hi" the value was 1 if the student's race was American Indian/Pacific Islander, Asian, African-American, or Hispanic, respectively. Thus Caucasian students were represented by having each of these four dummy variables equal to 0.

The complexity of reviewing each BY variable to develop a recoded version was deemed too time-consuming. Another strategic decision was made to limit the set of potential variables to a more manageable size. Prior research findings in the literature and experience gained in working with the dataset were used to select the smaller set of variables to be tested. A set of 66 variables out of the over 7,000 available was chosen. These variables reflected aspects of students for which prior educational research had found significant differences existed between outcomes[72]. These variables included basic demographic measures of sex, race, socioeconomic status, and family structure. Performance variables indicating standardized test results, NELS

cognitive testing measures, subject competency ratings, and average grades were also included. Several attitudinal/behavioral variables were also selected. These included measures of student and parental attitudes about education, individual subjects, degree aspirations, and student capabilities. Behavioral variables examined how students spent time on homework, social activities, television watching, etc. Once the recoding process was completed including creation of several dummy variables a set of 76 potential predictors was available for model development. The set of variables chosen for modeling purposes is listed in Table A.3 of Appendix A.

4.0 DEFINING "STEM"

4.1 INTRODUCTION

In order to test the integrated model's ability to predict between educational outcomes for different students it was first necessary to classify those outcomes. In the course of the literature review it became clear that there was a lack of a consistent scheme for classifying college majors into STEM vs. other outcomes. Developing a logical process for classifying majors became necessary and was an unanticipated benefit of the research. The process started by considering the aspects of different majors that would or would not qualify them to be considered "STEM." From there each major recognized by the NELS:88 dataset was given an initial categorization based on their general aspects. The classifications were compared to those made by prior education researchers, and adjusted if a persuasive case had been made. Then statistical tests were conducted to determine if a logistic regression model could significantly predict which students would select between two categories of majors.

4.2 CLASSIFYING COLLEGE MAJORS AS STEM

A critical aspect of this analysis is the ability to categorize students as having earned a STEM degree vs. another outcome based on their response to questions about their college major and

degrees earned. The categorization of a major as STEM or not is based upon the verbal name of the major, the body of coursework that is presumed to accompany that major, and the anticipated career application of a degree in that field.

The decision of which majors should be considered part of STEM is not consistently agreed upon. There is no universally accepted list of majors that can be classified as STEM. Prior research of persistence in various degree programs often does not make it clear how STEM has been defined. Many of the papers cited in the literature review have referred only to majors in "Science," "Engineering," "Mathematics," or some combination of these fields. In few cases was a comprehensive list of what was or was not included in the majors being studied provided. The literature is not sufficiently detailed to provide an agreed upon definition of STEM.

4.3 PRIOR RESEARCH CLASSIFYING COLLEGE MAJORS

Some prior researchers have assembled a set of majors that were classified as STEM or a subset of STEM based on their individual research interests and their interpretation of the relevant academic literature. For example, Adelman[73] created a list of majors that were classified as Science or Engineering for his 1998 analysis of the paths taken by students in their undergraduate careers in Engineering. One of his key points in separating engineering from Science was that the practice of Engineering involved working closely with clients to satisfy their expectations. This involved far more social interaction and required greater awareness of customer service issues than the practice of a bench science. He indicated that it was important to understand that Science and Engineering students can be quite different.

Seymour and Hewitt[74] defined a set of majors as Science/Mathematics/Engineering (SME) in their research. For the purposes of this analysis SME and STEM are synonymous. The broadly defined majors considered to be STEM were Mathematics/Statistics, Physical Sciences, Biological sciences, Engineering, and Agriculture. The broadly defined Non-STEM majors were Computer Science, Health, Business, Education, All Humanities & Fine Arts, Other Non-Technical, and Undecided. Table B.1 in Appendix B lists the fields of study included in these majors by Seymour & Hewitt.

Smyth[75] classified certain majors as SME for his analysis of ethnic differences in graduation from selective colleges with a science degree. This built upon earlier work in logistic regression analysis in graduation trends by race/ethnicity at selective colleges by Smyth and McArdle[76]. Smyth analyzed data obtained from a set of colleges within the College and Beyond (C&B) database[77]. This database was developed by the Andrew W. Mellon Foundation (AMF)[78] and includes data from 34 colleges. Smyth worked with a subset of the C&B colleges[79] that were defined in his analysis as selective in their acceptance of students and possessing greater academic prestige. Smyth obtained data from the Cooperative Institutional Research Program (CIRP) for 24 of the universities within the C&B database. The other universities within the C&B database did not have corresponding CIRP data available. Two of the CIRP variables concerning the students' intended majors and how they rated the importance of contributing to the body of scientific knowledge were utilized in examining differences in SME graduation between Cacuasian and African-American students. The majors that Smyth categorized as SME (STEM) for the purposes of his research are listed in Table B.2 of Appendix B.

Smyth referred to research by Astin and Astin[80] and Hilton, Hsia, Solorzano, and Benton[81] in his decision to exclude the social sciences and psychology from the STEM category. Prior research had found that very few of the students that declared an intention to major in these two areas switched to a major generally considered as STEM despite approximately half leaving their original majors. This suggested that the students who selected the social sciences and psychology as a major tended to have less academic interests in the standard STEM majors than with other potential majors. The accuracy of regression analyses that included majors in the social sciences to predict STEM graduation was much poorer than regression analyses excluding these majors. When the social sciences and psychology were excluded from the STEM category, the predictive ability of the regression models was more accurate.

The National Center for Education Statistics (NCES) collects data about the educational progress of students and makes several longitudinal datasets available to researchers. However, NCES has not created an "official" definition of which majors constitute STEM. Its function[82] is to collect and disseminate statistics about education rather than to specifically classify portions of education for research purposes.

The National Science Foundation (NSF) has created an official list of which majors it considers Science, Engineering, and Health related. NSF issues annual figures of degrees earned at the bachelors, masters, and doctoral levels by major, gender, and citizenship within the United States[83] through statistical reports available on its website. NSF classifies majors as falling within the Science, Engineering, or Health fields. An extensive taxonomy of fields of study has been developed by NSF to aid in the classification process. The broad categories included under the Science heading are Agricultural sciences, Biological sciences, Computer sciences, Earth/atmospheric/ocean sciences, Mathematical sciences, Physical sciences, Psychology, and

Social sciences. The Social sciences include fields such as Economics, Political science, Sociology, Linguistics, Anthropology, Archeology, Criminology, and Geography. The Engineering category includes Aerospace engineering, Chemical engineering, Civil engineering, Electrical engineering, Industrial engineering, Mechanical engineering, Metallurgical/materials engineering, and a general Other set for smaller Engineering disciplines. The Health fields include Medicine, Dentistry, Veterinary medicine, Health systems/service administration, Nursing, Pharmacy, and Rehabilitation/therapeutic services.

4.4 EXPANSIVE VS. NARROW DEFINITION OF STEM

There were several potential approaches to take in defining STEM within the context of this dissertation. The first approach was to define STEM very narrowly along the lines used by Smyth and Seymour and Hewitt. This approach limited STEM to the "hard" Sciences, the Engineering majors, and Mathematics while specifically excluding the Health fields, virtually all Technology majors, and the "soft" Sciences such as Psychology and Social Sciences. Under this narrow definition, the "hard" Sciences, the Engineering majors, and Mathematics would be classified as STEM majors and every other college major would be classified as "Non-STEM."

A second approach was to develop a more expansive definition of STEM to reflect that many other majors involve significant quantitative coursework in order to prepare for demanding careers applying scientific knowledge, mathematical skills, and independent judgment. Under an expanded definition of STEM many of the Health fields such as Medicine, Dentistry, advanced Nursing, etc. were included. Psychology and Social Sciences were included as well since techniques such as statistical analysis are often used to evaluate data in these fields. Many of the

Business majors would also be eligible since Business Administration, Business Finance, and Marketing may each involve significant analytical coursework to prepare students for applying quantitative techniques in making business decisions.

One argument for adopting an expansive definition of STEM is that many students in majors other than Science, Technology, Engineering, and Mathematics take extensive coursework in Science and Mathematics/Statistics so they can apply their knowledge in later careers. For example, an Accounting major requires a good knowledge of math. Sophisticated financial models for predicting the results of investments are developed by people with degrees in Business Finance. Researchers in the fields of Psychology and Sociology use complex statistical models to evaluate hypotheses. Medical doctors, Dentists, and Veterinarians take extensive coursework in topics such as biology, chemistry, and math in order to apply this knowledge in treating their patients. Nurses also require a strong background in biological science and chemistry. Medical professionals need strong skills to apply their technical, quantitative knowledge with independent judgment. If these majors require extensive coursework in and apply the techniques of majors traditionally classified as STEM, there is a case to be made for considering them STEM majors.

A counterargument in favor of not accepting an expanded definition of STEM is that the prediction of STEM vs. another outcome is enhanced by having sharp divisions between the outcomes. By "fencing off" a narrow definition of STEM that is easy to enforce in the classification of student records, the accuracy of the predictive models may be improved. Including records from students with a degree in other fields may blur those divisions and weaken the predictive value of the model as was reported by Smyth[84].

Part of the challenge in determining which of the other potential STEM majors should be included as STEM lies in evaluating the degree obtained. A bachelor's degree in Business Finance at a highly selective, prestigious university may involve significantly more demanding quantitative coursework than at a smaller university with more modest goals for its graduates. Even among highly selective, prestigious universities the body of quantitative coursework may vary making it harder to evaluate the suitability of a major for inclusion as STEM.

Another argument against the more expanded definition of STEM lies in the way the bachelor's degree is applied. Engineers, scientists, and mathematicians learn quantitative material in order to apply it in a very creative way. It is not enough for them to merely apply a difficult quantitative technique; they must understand the technique intimately so that it can be adapted independently as the situation warrants. In contrast, an accountant uses math more as an off-the-shelf tool. A medical doctor treating patients is applying his or her knowledge of biology, chemistry, and mathematical facts to prescribe an accepted standard treatment. A medical data entry clerk would need to understand medical terminology and be able to use data entry technology to record medical findings, but the clerk would not be applying technical skills in an independent fashion. It is indisputable that psychologists, medical doctors, and financial analysts may do groundbreaking research in their field, but this is not the focus of the vast majority of professionals in these fields. Therefore, including degrees earned in these fields as STEM outcomes may cloud the analysis of differences between STEM and Non-STEM students. The blurring of differences would correspondingly degrade the ability to predict persistence in STEM.

A third approach to defining STEM was to compromise between the narrow and expansive definitions. Majors in the "hard" Sciences, Engineering, and Mathematics were

categorized as STEM. Majors other than the "hard" Sciences, Engineering, and Mathematics that require extensive quantitative coursework were placed in a third category referred to as "STEM-Related." The remaining majors were categorized as Non-STEM.

The advantage of this third approach was that it offered a way to reflect the advantages of both the narrow and expansive definitions of STEM. The value of retaining the STEM-Related category as a separate outcome could be objectively determined by statistical analysis of potential significant differences between the students in each of the three categories. If a predictive model could discriminate between the potential outcomes with acceptable accuracy then there would be merit in keeping three categories. If there was insufficient accuracy in discrimination between two of the categories that would suggest the compromise STEM-Related category was not useful. Prior research has clearly indicated that significant statistical differences exist between STEM and Non-STEM majors[85], so it would be informative to test whether the proposed STEM-Related category was significantly different from STEM and/or Non-STEM.

If analysis found the STEM-Related category to be significantly different from one of the main two categories and not significantly different from the other this would suggest the majors within it could be grouped with the latter category. Then additional statistical tests would be appropriate to identify any significant differences between STEM + STEM-Related vs. Non-STEM or STEM vs. Non-STEM + STEM-Related. If analysis found the STEM-Related category to be significantly different from both of the main two categories then it would remain a separate grouping.

The third approach was more complicated, but it offered the prospect of creating a definition of STEM that could be logically tested and evaluated for future applications.

Therefore, this approach to defining STEM was the one chosen for this research. A proposed definition of STEM is as follows.

> "STEM is a path of study that involves significant coursework in advanced Science, Technology, Engineering, or Mathematics such that successful students acquire a comprehensive understanding of these subjects in order to extend and create knowledge. A STEM career requires extensive quantitative skills that can be utilized creatively and with a high degree of independent authority."

4.5 CATEGORIZING COLLEGE MAJORS

4.5.1 Selecting Majors for the Three Categories

As previously discussed, the number of STEM degree-holders produced by American universities directly affects the nation's competitive ability in the international marketplace. Producing more graduates with technical expertise to creatively apply their knowledge of Science, Engineering, and/or Mathematics is critical to the nation. Since all the prior research into technical education subjects is consistent in classifying the "hard" Sciences, Engineering, and Mathematics as "STEM," this narrow set of majors was automatically placed within the STEM category for this analysis.

The Non-STEM category was applied to majors that clearly did not require extensive coursework in quantitative, technical subjects. This category included the Fine Arts, English, and Other Humanities.

The STEM-Related category contains those majors that involve extensive quantitative coursework and represent a potential "gray" area between STEM and Non-STEM. This included the Health professions (medicine, dentistry, veterinary, pharmacy, nursing, and clinical therapies), Agriculture, Forestry, Social Sciences, Psychology, Business (Accounting, Business Administration, Finance, Marketing, and Management), and technical fields such as Computer Programming.

Table 4.1 lists the categorization of college majors within the NELS dataset by STEM, STEM-Related, and Non-STEM and compares these with the conclusions of Seymour & Hewitt, Smyth, and the National Science Foundation (NSF) classification of majors. It should be noted that the categorization as STEM or Non-STEM shown for Seymour & Hewitt and Smyth are based on interpretations of their STEM major classifications in prior published works. The categorizations shown for NSF are based on interpretations of their annual classification of programs into Science, Engineering, or other fields of study.

Table 4.1 Comparison of STEM vs. Non-STEM Major Classification by Researcher

NELS 88/00 Dataset F4EMJ1D var = "Major/field of study code - 1"	Gillian Nicholls			Seymour & Hewitt		Fred Smyth		Natl. Sci. Foundation	
	STEM	STEM-Related	Non-STEM	STEM	Non-STEM	STEM	Non-STEM	STEM	Non-STEM
Agriculture		Y		Y			Y	Y	
Agricultural science		Y		Y			Y	Y	
Natural resources		Y			Y		Y	Y	
Forestry		Y		Y			Y	Y	
Architecture			Y		Y		Y		Y
American civilization			Y		Y		Y	Y	
Area studies			Y		Y		Y	Y	
African-American studies			Y		Y		Y	Y	
Ethnic studies-not Black/area studies			Y		Y		Y	Y	
Accounting		Y			Y		Y		Y
Business-finance		Y			Y		Y		Y
Business-business/management systems		Y			Y		Y		Y
Business-management/administration			Y		Y		Y		Y
Business-secretarial			Y		Y		Y		Y
Business-business support			Y		Y		Y		Y
Business-marketing/distribution			Y		Y		Y		Y
Journalism			Y		Y		Y		Y
Communications			Y		Y		Y		Y
Communication technology		Y			Y		Y		Y
Computer programming		Y			Y	Y		Y	
Data processing technology		Y			Y	Y		Y	
Computer and information sciences	Y				Y	Y		Y	
Consumer services-cosmetology			Y		Y		Y		Y
Consumer services-mortuary			Y		Y		Y		Y
Education-early childhood			Y		Y		Y		Y
Education-elementary			Y		Y		Y		Y
Education-secondary			Y		Y		Y		Y
Education-special			Y		Y		Y		Y

Table 4.1 (continued).

NELS 88/00 Dataset F4EMJ1D var = "Major/field of study code - 1"	Gillian Nicholls			Seymour & Hewitt		Fred Smyth		Natl. Sci. Foundation	
	STEM	STEM-Related	Non-STEM	STEM	Non-STEM	STEM	Non-STEM	STEM	Non-STEM
Education-physical education			Y		Y		Y		Y
Education-other			Y		Y		Y		Y
Engineering-electrical	Y			Y		Y		Y	
Engineering-chemical	Y			Y		Y		Y	
Engineering-civil	Y			Y		Y		Y	
Engineering-mechanical	Y			Y		Y		Y	
Engineering-all other	Y			Y		Y		Y	
Engineering technology		Y		Y		Y		Y	
Spanish			Y		Y		Y		Y
Foreign language-non-European			Y		Y		Y		Y
Foreign language-European (not Spanish)			Y		Y		Y		Y
Health/allied-dental/medical technology		Y			Y		Y		Y
Health/allied-Therapy and mental health			Y		Y		Y		Y
Health/physical education/recreation			Y		Y		Y		Y
Nursing-nurse assisting			Y		Y		Y		Y
Health/allied-general and other			Y		Y		Y		Y
Nursing-nursing, post-RN		Y			Y		Y		Y
Health-audiology		Y			Y		Y		Y
Health-clinical health science		Y			Y		Y		Y
Health-dentistry		Y			Y	Y			Y
Health-medicine		Y			Y	Y			Y
Health-veterinary medicine		Y			Y	Y			Y
Nursing-registered nurse		Y			Y		Y		Y
Health-health/hospital Administration			Y		Y		Y		Y
Health-public health			Y		Y		Y		Y
Health-preparatory programs			Y		Y		Y		Y
Health-dietetics		Y			Y		Y		Y
Textiles			Y		Y		Y	Y	

Table 4.1 (continued).

NELS 88/00 Dataset F4EMJ1D var = "Major/field of study code - 1"	Gillian Nicholls			Seymour & Hewitt		Fred Smyth		Natl. Sci. Foundation	
	STEM	STEM-Related	Non-STEM	STEM	Non-STEM	STEM	Non-STEM	STEM	Non-STEM
Home economics-all other			Y		Y		Y		Y
Health-chiropractic			Y		Y		Y		Y
Health-pharmacy		Y			Y		Y		Y
Health-optometry			Y		Y		Y		Y
Vocational home economics-child care			Y		Y		Y		Y
Vocational home economics-other			Y		Y		Y		Y
Law-paralegal (includes pre-law)			Y		Y		Y		Y
Law			Y		Y		Y		Y
Letters-American/English literature			Y		Y		Y		Y
Letters-creative/technical writing			Y		Y		Y		Y
Letters-all other			Y		Y		Y		Y
Liberal studies			Y		Y		Y		Y
Library/archival science			Y		Y		Y		Y
Biological science-zoology	Y			Y		Y		Y	
Biological science-botany	Y			Y		Y		Y	
Biological science-biochemistry	Y			Y		Y		Y	
Biological science-all other	Y			Y		Y		Y	
Mathematics-statistics	Y			Y		Y		Y	
Mathematics-not statistics	Y			Y		Y		Y	
Military sciences			Y		Y		Y		Y
Women's studies			Y		Y		Y	Y	
Interdisciplinary-environmental studies		Y			Y		Y		Y
Interdisciplinary-biopsychology		Y			Y		Y		Y
Interdisciplinary-integrated science		Y			Y		Y		Y
Interdisciplinary-all other			Y		Y		Y		Y
Leisure studies			Y		Y		Y		Y
Basic/personal skills			Y		Y		Y		Y
Philosophy			Y		Y		Y		Y

Table 4.1 (continued).

NELS 88/00 Dataset F4EMJ1D var = "Major/field of study code - 1"	Gillian Nicholls			Seymour & Hewitt		Fred Smyth		Natl. Sci. Foundation	
	STEM	STEM-Related	Non-STEM	STEM	Non-STEM	STEM	Non-STEM	STEM	Non-STEM
Religious studies			Y		Y		Y		Y
Clinical pastoral care			Y		Y		Y		Y
Physical sciences-chemistry	Y			Y		Y		Y	
Physical sciences-earth science	Y			Y		Y		Y	
Physical sciences-physics	Y			Y		Y		Y	
Physical sci-not chemistry/physics/earth	Y			Y		Y		Y	
Psychology		Y			Y		Y	Y	
Protective services			Y		Y		Y		Y
Social work			Y		Y		Y		Y
Public administration-not social work			Y		Y		Y	Y	
Anthropology/archaeology		Y			Y		Y	Y	
Economics			Y		Y		Y		Y
Geography			Y		Y		Y	Y	
History			Y		Y		Y		Y
Sociology		Y			Y		Y	Y	
Political science			Y		Y		Y	Y	
International relations			Y		Y		Y	Y	
City planning			Y		Y		Y	Y	
Industrial arts-construction			Y		Y		Y		Y
Mechanics-transportation			Y		Y		Y		Y
Industrial arts-electronics		Y			Y		Y		Y
Mechanics-all other			Y		Y		Y		Y
Arts-commercial art			Y		Y		Y		Y
Precision production			Y		Y		Y		Y
Transportation-air			Y		Y		Y		Y
Transportation-not air			Y		Y		Y		Y
Arts-design			Y		Y		Y		Y
Arts-speech/drama			Y		Y		Y		Y

Table 4.1 (continued).

NELS 88/00 Dataset F4EMJ1D var = "Major/field of study code - 1"	Gillian Nicholls			Seymour & Hewitt		Fred Smyth		Natl. Sci. Foundation	
	STEM	STEM-Related	Non-STEM	STEM	Non-STEM	STEM	Non-STEM	STEM	Non-STEM
Arts-film arts			Y		Y		Y		Y
Arts-music			Y		Y		Y		Y
Arts-visual/performing/fine			Y		Y		Y		Y
Arts-crafts, folk art, artisanry			Y		Y		Y		Y
No major			Y		Y		Y		Y
{Don't know}			Y		Y		Y		Y
{Refused}			Y		Y		Y		Y
{Legitimate skip}			Y		Y		Y		Y
{Uncodeable}			Y		Y		Y		Y
{Not reached-partial/abbrev interview};			Y		Y		Y		Y

4.5.2 Creation of Additional Categories

The three initial student outcome categories of STEM for those who earned at least a bachelor's degree in a STEM topic, STEM-Related for those who earned at least a bachelor's degree in a STEM-Related topic, and Non-STEM for those who earned at least a bachelor's degree in a Non-STEM topic applied to all students who completed a college degree. These categories excluded any student who did not attend college, did not complete a degree, or achieved a less than four year degree. These outcomes were also of interest since they represented a departure from the STEM degree track at some point during the students' educational careers. The model was initially conceptualized to predict between STEM and Non-STEM outcomes where STEM was defined narrowly and all other outcomes would be grouped together as Non-STEM. This interpretation of STEM vs. Non-STEM as well as that of modeling STEM vs. Non-STEM solely based on students earning a college degree was the modeling approach generally taken in prior educational research.

Thought was given to grouping the students with no degree or a less than 4-yr degree into a non-bachelor's degree category, but the reasoning that led to the creation of the STEM-Related category ultimately led to a decision to keep the data more finely divided. Better predictive accuracy is generally obtained when the differences between the two groups being compared are distinct. It became clear that including other outcomes in the Non-STEM category might be misleading. Having a group whose membership included a diverse group of students earning other college degrees, less than four year degrees, or no degree at all could lead to a model with less predictive accuracy. Excluding the students who earned two-year degrees or none at all would preclude analysis of students who departed the STEM track at earlier points in the

educational process. Instead, it was decided to sort the students into a set of categories that would consider the four other potential educational outcomes besides STEM

Two additional categories were created as "Sub 4-yr Degree" for those who earned a college degree no higher than an Associates and "No Degree" for those who did not seek or did not complete a college degree. These five categories result in ten individual models discriminating between the pairs of outcomes. While STEM vs. STEM-Related, STEM vs. Non-STEM, STEM vs. Sub 4-yr Degree, and STEM vs. No Degree were of greater interest initially, the other models provided some interesting insights into the relationships between outcomes.

In addition, combinations of the five individual educational outcomes were created to allow for an even wider set of comparisons. These included combining the students who earned a STEM-Related or Non-STEM degree into the "Other Degree" category for a STEM vs. Other Degree model. Grouping the Sub 4-Yr Degree and No Degree outcomes into a "NonDegree" outcome allowed for comparisons with STEM and a Degree category comprising all students achieving a four year college degree. Another model was created by grouping the four outcomes other than STEM together into "All Else" to predict STEM vs. All Else.

4.5.3 Preparation of the Dataset

A binary dummy variable, "All5," was created to identify which of the 12,144 student records related to a student that had participated in all five waves of data collection. This was determined from the five "universe" variables that as outlined in Section 3.3 communicated the status of each student during the BY and four follow-up waves of data collection. There were 11,328 students classified as responding in all five waves of data collection.

Once the definition of STEM for the purposes of this research had been created and the reasoning behind the other categories had been established, the dataset had to be adjusted accordingly. Each record in the dataset needed to be assigned to one or more of the categories. This was done by creating new binary dummy variables that acted as yes/no flags in classifying the students into the five separate educational outcome categories and the various combination outcome categories. The first and second degrees earned (F4EDGR1 and F4EDGR2) and the majors associated with those degrees (F4EMJ1D and F4EMJ2D) were examined. The categorical values stored for the two majors were compared to ranges of values for the five separate educational outcome categories. The two majors were classified as STEM, STEM-Related, Non-STEM, Sub 4-Yr Degree, or No Degree. Each degree was classified as having achieved less than a bachelor's degree or having achieved at least a bachelor's degree. This resulted in a set of 10 binary dummy variables set to 0 or 1 depending on whether majors and degrees were "positive" for the circumstances described.

A total of five additional binary dummy variables were created to classify the students' overall as having a STEM, STEM-Related, Non-STEM, Sub 4-Yr Degree, or No Degree outcome. The values were set by creating programming code that examined the variables for the first two majors and associated degrees. If either major was in a STEM topic and the associated college degree was at least a bachelor's, the student was categorized as STEM. If at least a bachelor's degree was earned in a major other than STEM, and at least one of them was in a STEM-Related topic, the student was categorized as STEM-Related. If the student earned at least a bachelor's degree in a subject other than STEM or STEM-Related then the outcome was categorized as Non-STEM. If an associate's degree or other less than 4-year degree was earned the outcome was classified as Sub 4-yr Degree. If the student earned no college degree, the

outcome was classified as No Degree. Through setting this series of five binary dummy variables to 1 for affirmative conditions, each student was classified as belonging to one of the five separated educational outcome categories. Additional binary dummy variables were created to categorize records by the other groupings including 4 Year Degree, Non-4 Year Degree, Other Degree, and All Else. The final outcome of the classification was as follows in Table 4.2.

Table 4.2 Numbers of Students Classified by Group

Category	Number of Students	Included in Other Combination Category			
		All Else	4 Yr Degree	Non-4 Yr Degree	Other Degree
STEM	738	No	Yes	No	No
STEM-Related	1,077	Yes	Yes	No	Yes
Non-STEM	2,084	Yes	Yes	No	Yes
Sub 4-Yr Degree	1,732	Yes	No	Yes	No
No Degree	5,697	Yes	No	Yes	No
Total	11,328	10,590	3,899	7,429	3,161

Once the records were categorized in this manner it became possible to model these outcomes as dependent variables resulting from a set of demographic, academic, attitudinal, and experiential covariates. The various categories were grouped into pairs so that models were fitted to predict students as having a STEM vs. STEM-Related, STEM vs. All Else, STEM vs. Non-STEM, STEM vs. Other Degree, 4 Yr Degree vs. Non-4 Yr Degree, etc. outcome.

5.0 PHILOSOPHY OF THE INTEGRATED MODELING PROCESS

5.1 INTRODUCTION

The purpose of the model is to identify the significant quantitative and qualitative factors of the students' secondary education and to utilize them to accurately project the students' ultimate post-secondary educational outcome. The model uses this set of variables as inputs to an integrated series of statistical methodologies. The model's output is a predicted probability that a student with a given multivariate vector remains on track to complete his/her post-secondary educational process with a STEM degree given that he/she has "survived" on the track past a specific point in time. This permits the identification of those students for whom an intervention could be beneficial in terms of prompting them to consider a STEM degree. The model's value is assessed by comparing how accurately the predicted outcomes for a group of students matches the actual results and how this compares to the results obtained with standard logistic regression.

5.2 METHODOLOGY

The model combines logistic regression, survival analysis, Receiver Operating Characteristics (ROC) curve analysis, and sensitivity analysis. The initial model phase is logistic regression analysis where the set of variables is tested to determine which variables are the most

consistently significant predictors of educational outcomes. The results of the logistic regression portion serve as inputs to the survival analysis portion along with two additional variables: the actual educational outcomes of the students and the time point at which they left the path of achieving a STEM degree (if this occurred). The output of the survival analysis phase is an estimated probability of a student with a given set of input variables remaining on the STEM path past a specific time point. The probability estimates serve as inputs to the ROC curve methodology which visually depicts the tradeoff between correct predictions of a STEM path departure and correct predictions of not departing STEM for various cutoff values of the departure probability. The last portion of the model is a sensitivity analysis module that guides the selection of the probability cutoff point to meet the analyst's policy goals. Sensitivity in this context refers to the model's responsiveness to a small change in the input rather than to the probability of a correct STEM prediction. The sensitivity analysis indicates the mix of STEM and Not STEM students that would be reached by a structured intervention program if the prediction cutoff point is altered. It is envisioned that the potential intervention would be designed to increase the students' interest in STEM and positively affect their likelihood of getting a STEM degree. A diagram of the integrated model is shown in Figure 5.1.

The model in Figure 5.1 enables identification of students who have the potential to achieve a STEM degree but might not seek to pursue one without encouragement or assistance. Integrating survival analysis provides the ability to specifically identify the time points in the educational process at which the probability of leaving the STEM track changes the most and which variables are significant predictors of the probability. Knowing when students are likely to drop out of pursuing a STEM degree and which significant variables predict this enables policymakers to test strategies to address the loss of potential STEM degree-holders.

Combinations of changes to significant quantitative and qualitative factors in high school education can be explored with the model. This provides a means to determine the sensitivity of the probability of STEM track departure to changes in variable values.

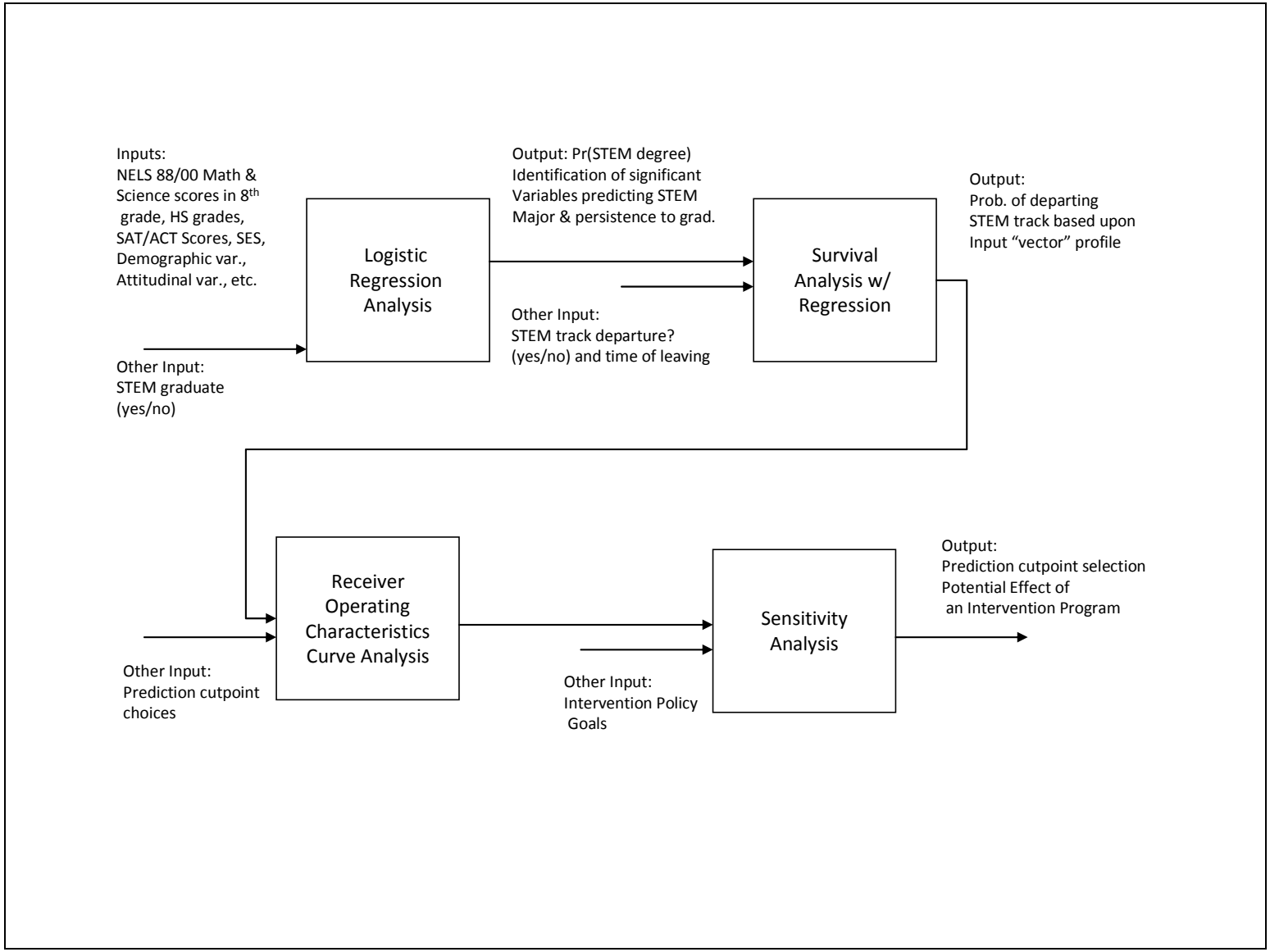

Figure 5.1 Integrated Sequential Statistical Technique Model

5.3 RESEARCH HYPOTHESIS

The research hypothesis was that the integrated approach would produce more accurate results in terms of predicting which students get a STEM degree or not than a single standard statistical technique such as logistic regression. The null hypothesis was that the integrated model would not prove to be a statistically significant improvement over the logistic regression approach alone. In the process of testing this hypothesis there were a number of research theories that were also explored.

5.4 RESEARCH QUESTIONS EXAMINED

The research questions which arose through the development of the integrated model included the applicability of survival analysis techniques to predicting post-secondary educational achievement; the feasibility of integrating multiple statistical techniques for analyzing this type of complex problem; the determination as to whether some students had an inherently high probability of departing the STEM track while still in junior high school; the negative or positive impact of significant variables; the presence of key time points in the educational process when the probability of STEM track departure changed abruptly; and the ability to identify a narrow time window in which an intervention could benefit a large number of students.

Question 1: Can multiple statistical techniques such as logistic regression analysis, ROC curve analysis, and survival analysis be successfully integrated and applied to a complex problem such as the achievement of a STEM degree?

This question was examined by fitting the models and determining if the predictive results of the models achieved acceptable accuracy.

Question 2: Can a set of variables that were measured for a group of students as they progress through high school and beyond be shown to affect the probability that a given student fails to "survive" to achieve a bachelor's degree in Science, Technology, Engineering, or Mathematics (STEM)?

These variables included:

- measures of academic skill in math, science, and English
- measures of academic self-confidence such as self-ratings of ability in different subjects
- measures of personal attitudes such as the student's school/career aspirations
- measures of academic focus such as hours per week on homework by subject, how many hours the student works per week, how many more advanced classes the student has taken, and what discussions the student has had with others regarding school
- measures of demography, socioeconomic status (SES), family structure, and language

This question was explored by determining if the variables which were found to be significant in prior research were statistically significant predictors in the logistic regression and integrated models developed in this dissertation.

Question 3: Will Survival Analysis of the NELS:88 data reveal that the probability of a student achieving a STEM degree differ over time for students in different outcome groups?

This question was studied by plotting the hazard curves over time for the students in different departure type classes to examine how the probability of a departing the STEM track at a given time varied by type and whether the different probability curve plots overlapped or diverged at key time points.

Question 4: Are there key time points in the educational process where distinct decreases or slight increases in the probability of achieving a STEM degree occur as students developed academically? If so, are these key time points at which students were most likely to depart the STEM track sufficiently common for different student profiles that they could suggest the timing for delivery of pro-STEM intervention? It is possible the analysis could reveal that many students had a relatively low probability of surviving to achieve a STEM degree as they were measured in 10^{th} grade due to lack of interest in the material or very poor academic preparation. It is also possible that individual students may have had a relatively low probability of surviving to achieve a STEM degree prior to 8^{th} grade, but the study could not assess that since data collection started with the base year in 8^{th} grade.

This question was evaluated by examining the hazard curves to identify any points in time at which the slopes of the curves changed dramatically.

5.5 EVALUATION OF THE LOGISTIC AND INTEGRATED MODELS

The evaluation plan for the models involved applying the models created with the "fit" data to the 30% of the records withheld as the "test" data to determine how accurate the predictions of the test data were when compared to the known outcomes. The logistic regression model was further tested by repeating the process multiple times in a cross-validation for two of the outcome pair models. The initial data for construction of the logistic regression model was randomly selected multiple times with different models built each time. The evaluation of the retained data sets was conducted and the results of the logistic regression models were compared to one another repetitively. For example, when predicting between a STEM degree and all other possible educational outcomes ("All Else") there were 11 separate samples drawn to create and test the logistic regression model. The same was done when predicting between a STEM and STEM-related educational outcome. This was done to determine if the modeling was sensitive to either the diversity of the All Else group or the lack of diversity between the STEM and STEM-Related groups.

Once the stability of the STEM vs. All Else logistic regression model was established it was compared to the integrated model. The results obtained for the test data from each model were examined to determine if the number of correct predictions was acceptable. The last step was to evaluate how the accuracy of the integrated model compared to that of the logistic regression model. As the standard technique, the logistic regression model predictions were treated as the "expected" results and the integrated model predictions were treated as the "observed" results. Classification tables that quantified the number of correct and incorrect predictions were prepared for each model and compared with chi-square testing to determine if statistically significant differences were found. The number of true and false STEM predictions

by random sample for each model were then compared and analyzed with *t*-tests. Again, the null hypothesis was that the integrated model offered no significant advantage over the standard logistic regression model.

6.0 THE MODEL

6.1 INTRODUCTION

Two separate models were designed: one to predict outcomes using a single standard statistical method (logistic regression) and one to predict outcomes using an integrated model with logistic regression results input into a survival analysis model. The logistic regression model was created first to set a standard of predictive accuracy for comparison to the overall integrated model. The survival analysis module inputs were the variables found to be significant by prior research, the logistic regression model's estimated probability of a STEM outcome for each student, the manner in which students departed the STEM track, and the time at which the students left STEM. The output of the survival analysis module was the probability of a given student remaining on the STEM track past a specific point in time given that they had survived to that point. The survival analysis probability of departure is an input to the receiver operating characteristics (ROC) curve analysis to depict the tradeoff between correct and incorrect predictions. The last module is the sensitivity analysis which assesses the effects on the correct and incorrect predictions if the prediction probability threshold is altered. It also explores the number of students that would be targeted for an intervention program to either encourage students to consider STEM and/or assist them in strengthening their academic capabilities based upon the threshold probability cutoff.

6.2 LOGISTIC REGRESSION ANALYSIS MODULE

Once created and tested, the logistic regression model became a module within the integrated model. Separate multivariate logistic regression models were constructed to predict between the various pairs of educational outcomes. The models were fitted using SAS with a large set of potential variables and the Stepwise selection method with an alpha (α) error level threshold of 0.05 to enter or leave. This variable selection method ensured that SAS constructed the model by first identifying the most useful predictor with a chi-square *p*-value of 0.05 or less. Once this variable had been entered into the model, SAS continued choosing potential variables in the same fashion with the provision that if a variable's entry caused a prior entrant's individual *p*-value to increase above 0.05 it was automatically removed from the model. SAS stopped after considering all the potential variables for inclusion or reaching a user-defined limitation on the number of potential cycles. Using the Stepwise selection method to test potential variables for inclusion in the model was far more efficient in considering a large set of variables. It would have been impractical to test such a large group of potential variables by constructing separate models with different fixed combinations of variables. The set of variables considered for model-fitting was chosen after considering the results of prior education research. Factors that previous research had found to be significant were compared to the NELS dataset to identify comparable variables.

Stepwise multivariate logistic regression was used to create models predicting a particular outcome between two possibilities such as STEM vs. All Else, STEM vs. STEM-Related, STEM vs. Non-STEM, STEM vs. Other Degree, Degree vs. Non-Degree, etc. The models were fitted with randomly selected sub-samples of students that were constructed to proportionally represent their numbers in the entire sample with those outcomes. Thus, a model

to predict between STEM and STEM-Related was fitted using records randomly drawn from the sets of STEM and STEM-Related students. The sub-samples were stratified by the outcome of interest with 70% of the total records from each stratum randomly selected for the model fitting. For example, in modeling STEM vs. All Else the total number of student records was 11,328 of which 738 obtained a STEM degree. These records were stratified by STEM = 1 for STEM students and STEM = 0 for all other outcomes. Of the 7,931 records approximately 70% of the STEM (517 records) and All Else students (7,414 records) were randomly selected for the model fitting.

Subsequently, each of the logistic regression models were validated by taking the model developed and applying it to predicting the outcome for the remaining 30% of the records from each stratum. Thus in the STEM vs. All Else case, the records of the remaining 221 STEM students and 3,176 of the All Else students were used to validate the model created with the original 70% of the total records. A Receiver Operating Characteristics (ROC) curve was then created to measure the impact on correct/incorrect predictions of STEM based on the cutpoint in the probability of STEM estimated. These curves plotted the probability of a correct STEM prediction (sensitivity) vs. the probability of an incorrect STEM prediction (1 – specificity).

The random sub-samples by stratum were created by a procedure within SAS that utilizes a random number generator. The procedure uses a "seed" number in combination with the random number generator to select sample members. The seed number is provided with the SAS output so that the user can replicate the sample by deliberately opting to use that seed number again. In this research, SAS was permitted to randomly choose a seed number for each of the two-outcome models fitted. However, once the seed number was created for an individual outcome pair model, it was used again in the validation process as well as subsequent model

fitting exercises for the same pair of outcomes. There were several reasons for this. First, in order to properly validate a model, it was essential that the records used to create the model were not included in the sub-sample chosen to test it. So if a seed number was used to create the sub-sample for model fitting, the same seed number was used to identify the records not previously used for the fitting process in order to validate it. Second, SAS was allowed to randomly create the seed numbers for each two-outcome pair to avoid unconsciously prejudicing the results by manually selecting seed numbers in advance. In addition, using the same seed number for later modeling of an individual two-outcome pair allowed the user to directly compare the predictive accuracy of the different models. For example, using the same seed number to generate model fitting/validating samples of STEM vs. All Else made it possible to directly compare the accuracy in predicting STEM when additional variables were included in the potential set of predictors.

In two cases that will be discussed in greater detail later, multiple seed numbers were generated to create different random sub-samples to fit models for the same two-outcome pair. This was done to determine the sensitivity of particular models to the sub-samples that were chosen. 11 different seed numbers were used to determine the consistency of significant predictor variables across the models fitted with the 11 random sub-samples.

The initial models were created using all of the 76 potential predictor variables developed through the sifting and recoding process. This was a much larger set of potential predictors than utilized in prior educational research modeling. While prior research has examined large sets of variables for significant differences between outcomes, the prior predictive models have generally used a much smaller set of predictors. In some of these prior attempts at predictive

modeling interaction terms have been considered, but these were limited to potential interactions of sex and race/ethnicity.

The potential set of two-way interaction terms for the 76 recoded variables was very large. Even after excluding meaningless interactions between related dummy variables such as F2RACE2As and F2RACE2Hi, there were still over 2,800 potential terms. Including all these terms as potential predictor variables would have been impractical. Instead, a sensible way of exploring the potential impact of interactions without overwhelming the models was sought. The compromise was to first create models using the set of 76 variables, identify which were significant predictors, and then to test new models including interaction terms for which both halves were variables found to be significant in the first round of model fitting. For example, if an initial model found that F4SEXrb and F2RACE2As were significant then the interaction term between F4SEXrb and F2RACE2As was tested in the model considering potential two-way interaction terms. Depending on the number of significant variables in the first round, the modeling with interactions considered included between 145 – 800 additional predictor variables.

After examining models including the potential recoded BY variables and standardized test scores from subsections of the SAT and ACT tests, it was decided to explore models using just the BY variables. The reason for doing so was to determine if acceptable predictive accuracy could be obtained using just the BY variables from 8th grade and ignoring the standardized test scores obtained in the second follow up (generally 12th grade). If acceptable accuracy could be achieved with these models it would bode well for identifying potential STEM students at a much earlier point in their educational careers when the probability of a successful intervention to encourage STEM would be higher. These "BY-only" models also featured

selected two-way interaction term variables for which both halves were BY variables previously found to be significant for a specific outcome pair.

For the sake of academic curiosity, additional models were considered for some of the other two-pair outcomes. Once a set of models had been created using the Stepwise selection method for BY plus standardized test score variables; BY variables, standardized test score variables, and selected interaction terms; and BY variables with selected BY interaction terms the models were re-fit using the fixed selection method with just the variables previously found to be significant for the relevant models. For example, if a model using stepwise selection had identified F4SEX, F2RACE2As, and overall BY mathematics proficiency as the only significant predictors, the model was then refit using each of those variables and only those variables.

6.3 SURVIVAL ANALYSIS MODULE

There were two alternative approaches in defining the characteristics of the survival analysis module. One alternative in designing the survival analysis module was to examine the survival of students on the STEM track. In this design alternative, the event of interest was the point at which students depart the STEM track. This departure could have occurred at any point along the educational progression that began in 1988 with the 8^{th} grade and continued until the study concluded in 2000. The hazard function in this sense was the conditional failure rate or the "approximate" probability that a student with a given profile at a point in time departed the STEM track in the next moment. What was measured was the time to failure with departing the STEM track considered failure.

There are a number of ways in which students departed the STEM track. For example, students departed the track by dropping out of high school, by not continuing on to college, by dropping out of college, by switching out of a STEM major, by graduating college with a degree in a subject other than STEM, or by pursuing a college degree without completing it by the study's end. Other theoretical departures included dying or declining to participate further in the study, but these departure types were eliminated by the design of the F4 wave of data collection. Students were purged from the study if they died, were not selected for further sample inclusion, could not be located for the fourth follow up, or declined to participate in the fourth follow up. The 12,144 records in the NELS:88/2000 dataset reflected all students that were chosen for inclusion in F4 and responded to the survey. Students who actually achieved a STEM degree never experienced the event of interest because they did not depart the STEM track.

In Survival Analysis if an event of interest is known to have taken place within certain time periods as opposed to an exact point in time, then the data is referred to as "censored." Censoring[86] is categorized by the relationship of the time period in which the event occurred and the time period of the data collection. If a subject does not experience the event of interest prior to the study's end date, then the data is said to be "right censored." The term means that if the event of interest occurred, it had to have happened after the study's data collection ended. This was the case for students that were still pursuing a STEM degree at the time the study ended. Since the NELS study ended in 2000, the records for students who had not yet experienced the event of dropping out of STEM were right censored as of December 31, 2000. Another class of right censoring is "competing risks" censoring in which some subjects experience the event of interest for different reasons. For example, students departed the STEM track by never graduating from high school or by graduating from college with a major other than STEM.

These two departure scenarios were competing risks and each was of interest. "Random censoring" is a special case of competing risks in which a student experiences a competing risk that precludes further participation. A student that graduated with a STEM degree was classified as randomly right censored because the competing "risk" of graduating with a STEM degree made it impossible for the person to later experience the event of interest: departing the STEM track.

Another censoring category is "interval censoring." This is utilized when the event of interest is known to have occurred within a fixed time interval, but the exact time cannot be precisely determined. This was the case for students who dropped out of high school or college by a certain point but for whom the exact departure date was unknown. The NELS study was designed to try to elicit information to gauge when students dropped out of high school or stopped attending college; however, it was not always possible to obtain this information. For some of the records, the student's educational status changed within a time interval and the exact time was not determinable.

A second alternative was to frame the analysis to examine the survival of students without a STEM degree. Under this design the event of interest was the point at which students acquired a STEM degree. Although this might have been easier in terms of identifying the occurrence of the event of interest and when it happened, there were a relatively small number of students that experienced this outcome. A total of 738 students out of the 11,328 classified as having responded in all five waves of the data collection were classified as having earned a STEM degree. A model framed this way would have 738 instances of the event of interest occurring and 10,590 cases of the data being censored at the study's conclusion. This model formulation had no way to distinguish between the various groups of students who were

censored since technically, even a person who failed to finish high school by the study's conclusion could have gone on later in life to earn an engineering degree. Losing the ability to distinguish between how students departed the track would have weakened the predictive utility of the model and sacrificed the chance to study students' decisions to leave STEM.

After considering both alternatives, a decision was made to model the data in terms of time to STEM track departure. This appeared to be the best way to take advantage of the dataset's size, broad knowledge about the subjects, and range of potential competing risks.

6.3.1 Classification of Students

The status of the students throughout the study was determined by examining several sets of variables. The variables F4UNIV1 and F4UNI2A through F4UNI2E indicate the overall status of the student's participation in the study as it was known during the fourth follow-up (F4). This subset of variables was referred to by the NELS:88 designers as the "universe" variables since they described the students' status through the study. Other sets of variables were used to establish departures from the STEM track, estimate the event occurrence time, and identify the types of departure or censoring that occurred. The description of individual variables used for classification purposes is found in Table A.4 of Appendix A. The methods for analyzing the variables to determine how to classify the records are discussed in Appendix C. Examples of the SAS code developed for the classification process are provided in Appendix D.

The classification of students by STEM track departure type utilized the post secondary educational transcript (PETS) dataset as well as the main restricted dataset for the NELS dataset that was utilized for the logistic regression analysis. The PETS data was contained on the N0T CD-ROM for the NELS dataset while the main restricted data was contained on the N0R CD-

ROM. The PETS data contained additional variables related to the timing and nature of educational events from high school through college attendance that were necessary for obtaining the STEM departure time event data needed for survival analysis. The use of this data allowed the student outcomes to be more precisely identified so that students earlier classified as having No Degree could be further sorted by whether they had graduated high school or ever attended college.

The process of sorting the students by STEM track departure type resulted in shifting some students between categories as degrees that were reported were in some cases not confirmed by the transcript data. In several other cases, students that had not reported four year degrees were found to have earned them. Each student's record was tested by SAS code designed to consult both the original dataset and the PETS dataset to ascertain the final educational outcome. If students reported degrees but did not have valid dates of graduation in one dataset, but they had valid dates in the second dataset they were categorized as having earned a degree. If the information from the two datasets conflicted in a way that could not be resolved, the records were excluded from further analysis. This was the case for 200 of the 11,328 records so the data used for the survival analysis module contained 11,128 records. Logistic regression models of STEM vs. All Else were fitted with the original classification of the dataset and the revised classification using the PETS variables applied to all 11,328 records. The results of the models were very similar with no significant effect on the model from changing the classification scheme of the records.

The final numbers of students in each category were very similar between the original and revised classification schemes. Some of the records did switch from one category to another with a net change in the numbers that was small. For example, 3 students that had been

classified as STEM for the initial logistic regression analysis were excluded from the revised dataset classification because their STEM degree could not be verified from the PETS data. Meanwhile, 1 student that was originally classified as No Degree was found to have multiple STEM degrees based on the PETS data and was recognized as STEM for the survival analysis. The net change in STEM students was a decrease of 2 students from 738 to 736. Of the 200 excluded records 177 were previously No Degree students, 17 were previously Sub 4 Yr Degree students, 3 were previously classified as Non-STEM students, and the remaining 3 were formerly considered STEM students as discussed above.

6.3.2 STEM Track Departure Types

6.3.2.1 Drop Out of High School

A series of variables that record drop out status and history allowed the analyst to get a sense of when a student departed the STEM track by dropping out of high school. These include the periodic drop out status variables F1DOSTAT, F2DOSTAT, and F2F1DOST as well as variables that measured whether a student reported having ever dropped out of high school in the past (F2EVDOST and F3EVDOST). Note that the first two characters of these variable names refer to the wave of data collection in which they were obtained: F1 for first follow up (1990), F2 for second follow up (1992), and F3 for third follow up (1994). The variables F1D7MNTH and F1D7YEAR indicate the last date the student reported having attended school as of the first follow up. The variables F2D6M and F2D6Y indicate the date the student reported having attended high school as of the second follow up. These variables were used to estimate the date at which the student dropped out of high school. If the student was reported to have dropped out but no date was given, the date was imputed to be the start of 1991 and what would probably

have been 11th grade. The reasoning for this decision was that it lay within the first and second follow ups and prior research[87] indicates that the risk of dropping out rises sharply to peak in the 10th grade and then decline in the 11th grade with a further decline in the 12th grade. Thus imputing a drop out date in the middle of the 11th grade was deemed a sensible way of accounting for the missing data. Of the 567 students classified as high school dropouts there were 122 (21.52%) students for whom the departure time was imputed.

It was decided that students who left the STEM track by dropping out of high school permanently would be treated as remaining on the track until their final drop out date. This meant that if a student dropped out more than once only the final date was utilized. The design of this educational model has the potential for repeated events of interest. Unlike a survival analysis model in which the event of interest is the subject's death, the event of dropping out of school could occur more than once. A student could drop out, return to school, drop out again, and either return or not at a later point.

There are different ways of handling repeated events[88]. One approach would be to ignore the repeated nature and to use either the first or final occurrence as the sole time the event occurred. Another approach would be to create two records for these students and treat each drop out as a separate instance. In this case the time to event could be measured as the time from the study's start to the time of dropping out in each occurrence or the second instance could be assigned a start time corresponding to the student's estimated return to school date. However, this approach could present problems since creating separate records for student drop outs with multiple returns would introduce correlation into the dataset.

While this phenomenon is undoubtedly interesting, exploring it within this model would have greatly complicated the analysis. Examination of the 11,328 students that participated in all

five waves revealed that 84 students dropped out more than once. Of the students that dropped out more than once 47 students returned after the second instance of dropping out. This represents only 0.0041% of the total sample. Given the small population size experiencing this repeated event, the additional model complexity was deemed unwarranted. It should be noted that the repeated drop outs could have resulted from many reasons including illness, family emergencies, etc. An examination of the repeated drop out phenomenon would be better suited to a study designed to focus exclusively on high school students and their reasons for dropping out vs. staying in school. In this analysis, repeated events were handled by using the final dropout as the time the student departed the STEM track.

6.3.2.2 Conclude Education at High School

In this scenario, the student completes high school by earning a diploma, passing a General Educational Development (GED) test, or achieving some alternative certificate; and decides not to pursue a college education. The NELS dataset contains variables that indicate high school completion, type of completion, date of completion, and whether the student reported ever pursuing post-secondary education or not. If the student did not pursue post-secondary education then the STEM track departure was deemed to have occurred at the point of graduation from high school. Two variables were utilized in determining the high school graduation date. The main N0R restricted dataset provided the variable F4HSGRDT to capture this date and the PETS N0T dataset provided the PETSHSDT date. In most cases the dates were identical although F4HSGRDT was formatted as YYYYMM and PETSHSDT was formatted as YYYY.mm. The "mm" portion of the format was a decimal version of the month. For the sake of consistency, the PETSHSDT variable was used to determine the students' high school graduation date. There

were 2,369 students that were identified as having graduated high school and not pursued a college degree.

6.3.2.3 Drop Out of College

This type of departure was verified through variables indicating college attendance during a specific year and when the student left college. As with the high school drop out scenario, there could be repeated instances of this event. Students might have left one college and transferred to another. They might have dropped out of college for a period and then returned.

This was addressed by examining the variables for post-secondary attendance and breaking these cases into sub-outcome categories. If a degree was ultimately earned the student was not classified as a college dropout. If a student's records indicated some post-secondary education at one or more institutions without ever earning a degree, then the last date of college attendance was considered the point at which the student departed the STEM track. The PSEND variable from the transcript dataset indicated the latest date at which the student was enrolled in a post-secondary educational institution. This final date of college attendance was considered the departure time. There were 2,164 students that were determined to have graduated high school and attended a post-secondary institution without achieving any degree.

6.3.2.4 Incomplete College Degree

Students in this category were still enrolled in college when the study ended and had not yet earned any degree. These students may have taken longer than normal to start a college degree or they may have taken a much longer than normal time once in college. An example would be a student who attended college part time and was not able to complete the degree prior to the study's end. Another would be a student who sought employment after high school graduation

and then began attending college several years later. This scenario was established through the college attendance, degree earned, major declared, and graduation date variables. In addition, the dataset offers a variable that indicates whether a student still in college at the time of the study's conclusion was expected to complete the degree within a year. The censoring date was the end of the study. The study ended on December 31, 2000 so the departure date was set to 2001.00. This was a case of "right censoring." There were 433 students still enrolled in a post-secondary educational institution at the time the study ended who had not previously earned any sort of degree. No distinction was made between the types of degree being sought when the study ended.

6.3.2.5 Graduated College with a Sub 4 Year Degree

Students in this category departed the STEM track by earning a certificate or Associates Degree credential. As with the four year degree categories, the departure time was determined by the date the degree was awarded or the last date the student was enrolled in the post-secondary educational institution. If the date of graduation with the degree was not provided, it was imputed from the last date of enrollment. There were 1,703 students identified as departing the STEM track by earning a less than four year degree.

6.3.2.6 Graduate College with Other 4 Year Degree

Students that completed their education by earning a degree other than STEM were identified by the combination of variables indicating a degree was earned, the major declared for the degree, and the date of graduation. Under this scenario, the students departed the STEM track upon their college graduation with a bachelor's degree other than STEM. For the purposes of the survival analysis module, STEM-Related and Non-STEM degrees were handled identically. There were

3,156 students classified as having departed the STEM track by earning a four year degree in a subject other than STEM.

If a student attempted a STEM degree but switched his or her major prior to graduation, then ideally the departure date would have been the time at which the major changed rather than the date of graduation with a different degree. However, this date was difficult to identify and subject to interval censoring. Therefore, students in this category were classified simply as earning a different degree and the departure time was set to their graduation date.

Note, if a student earned a different degree and then later achieved a STEM degree, then the earlier degree was no longer considered a departure from the STEM track. Instead the student was classified as having a STEM outcome and not having departed the STEM track.

6.3.2.7 Obtain a STEM Degree

If a student completed a college degree in a STEM topic, then there was no departure from the STEM track. This was a random competing risk in which the student could no longer experience the event of interest. This outcome was verified by the variables for degree earned, degree major, and the date of graduation. In this case, the student's record was randomly censored as of the date of graduation with a STEM degree. There were 736 students classified as having earned a STEM degree.

A summary of the different types of STEM track departures and how they were handled in the model is detailed in Table 6.1.

Table 6.1 Determining the Time of Departure by Departure Type

STEM Track Departure Type	Number of Students	Time of STEM Track Departure
Drop Out of High School	567	Final date enrolled in high school or 1991 if missing data
Conclude Education at High School	2,369	High School Graduation date
Drop out of College	2,164	Date of last post-secondary educational enrollment
Incomplete Degree	433	Date of study end: December 31, 2000
Graduate College with a Sub 4 Year Degree	1,703	Date of college graduation
Graduate College with Other 4 Year Degree	3,156	Date of college graduation
Obtain a STEM Degree	736	Date of graduation with STEM degree
Total	11,128	

Table 6.2 compares the number of students in each category used for the logistic regression and survival analysis modules. The categories include those that represent combinations of some of the individual outcomes. For example, the Other Degree category includes the STEM-Related and Non-STEM students; the 4 Year Degree category includes the STEM, STEM-Related, and Non-STEM students; the Non 4 Year Degree category includes the No Degree and Sub 4 Year Degree students; and All Else reflects all the students that did not earn a STEM degree.

Table 6.2 Comparison of Records Sorting between Logistic Regression and Survival Analysis

Category	Logistic Regression Analysis	Survival Analysis
STEM	738	736
STEM-Rel	1,077	1,113
Non-STEM	2,084	2,043
Sub 4 Yr Deg	1,732	1,703
Other Degree	3,161	3,156
Degree (4 yr)	3,899	3,892
Non 4 yr Degree	7,429	7,236
No Degree	5,697	5,533
All Else	10,590	10,392
Excluded records	n/a	200
Total Records	11,328	11,328

6.3.3 Origin Point

The time to the event of interest was determined from the STEM track departure time and the starting or "origin" point in time. Two logical choices for this origin point were the study start time and the students' date of birth. The choice of the origin point affects the estimates of any coefficients in a model as well as its fit. Since the NELS study began in 1988 when the students were in 8th grade and concluded on December 31, 2000 the maximum time to event could range from approximately 12 years using the study start time as the origin to approximately 28 years using the students' individual birth dates.

The argument in favor of using the study start time is that it focused the analysis on the time period during which the data was being collected. Factors which affected the students prior to the study were not determinable and automatically lengthening the time to event by 12-14 years might obscure subtle variations in the data during the study time. More importantly, if we assume that each student is a potential STEM graduate, the opportunity to depart the STEM track

in this analysis does not start until the study's initiation. While a student could theoretically have a particularly weak academic preparation prior to 8th grade such that the risk of departing STEM was elevated prior to the study's start, there is no data from earlier time periods available to quantify this risk.

The argument in favor of using each student's individual birth date as his/her origin point is that students may reach 8th grade at different ages thereby affecting their levels of mental, physical, and emotional maturity. A student's birth may be considered the ultimate starting point of his or her individual educational process. Using the birth date as the origin could help capture the effect of age on the model without adding another potential covariate.

In the absence of a strongly compelling reason for using the study start or the date of birth as the origin time, a decision was made to create two measures of the time to event. Both measures were used to create the initial models and a decision of which to continue using was based upon the strength of the models.

The first time to event variable was based upon the study start time in the Spring semester of 1988. To ensure that this time to event was at least one twelfth of a year, the start time was set to December 1, 1987 by using the decimal date 1987.92 for 1987 and 11/12 months. Similarly, the study conclusion on December 31, 2000 was set to a decimal date value of 2001.00 for 2001 and 0 months. The time to STEM track departure was calculated as the time that elapsed between the study's start and the time at which the student could be determined to have left STEM or the study's conclusion, whichever came first. This time variable was called "Track_Time."

The second measure of the time to STEM departure was created by using the students' birth dates as their individual starting points. The decimal birth date was calculated from the

birth year and birth month as YYYY.0 + (MM – 1)/12. The actual day of the month that a student was born on was not provided so each birth date was set to the beginning of the birth month. This measure of the time to event was referred to as the educational duration and the variable name was "Educ_Dur." The elapsed time to event was measured as the time between the birth month/year and the time to STEM track departure.

6.3.4 Model Selection

There are different options for modeling time to event data using survival analysis. One approach is to use a parametric model[89] that assumes the time to event is distributed according to a particular probability distribution. Then covariates are tested in the model to determine how well they explain variations in survival time. The model can be constructed along the lines of classic linear regression where the survival time (Y) is modeled as function of the natural logarithm of the departure time (X >0). In this scenario the linear model is $Y = \mu + \gamma^{\dagger}\mathbf{Z} + \sigma E$ where $\gamma^{\dagger} = (\gamma_1, \gamma_2, \ldots \gamma_p)$ is a vector of estimated coefficients, $\mathbf{Z}$ is a vector of covariate values, σ is the variance of Y, and E is the error distribution. Different probability distributions may be employed in the model for the error distribution including the standard normal distribution, the logistic distribution, and the extreme value (2-parameter) distribution. The choice of the error distribution results in modeling Y as ln (X) leading to a lognormal, log-logistic, or Weibull regression model respectively. The regression coefficients for the covariates are then estimated using the maximum likelihood method.

Another parametric approach is to create an "accelerated failure-time model"[90] where a baseline survival function is estimated for $\mathbf{Z} = [\mathbf{0}]$ as $S_0(x) = \text{Prob}[Y = \text{survival time} > \ln(x) \mid Z = [\mathbf{0}]\ ] = exp(\mu+\sigma E)$. From there the effect of nonzero covariates changes the survival time by a

factor of $\exp(-\gamma^t \mathbf{Z})$ by "accelerating" or "decelerating" the time to failure based upon the sign of the $\gamma^t \mathbf{Z}$ term. In this scenario the survival time on the STEM track of a student with $\mathbf{Z} = [\mathbf{0}]$ is $S_0(x) = exp(\mu+\sigma E)$ and the survival time of student i with $\mathbf{Z} \neq [\mathbf{0}]$ as $\text{Prob}[X > x \mid \mathbf{Z}] = \text{Prob}[Y > \ln(x) \mid \mathbf{Z}] = S_i(x \mid \mathbf{Z}) = S_0(x \exp(-\gamma^t \mathbf{Z}))$ for all $x > 0$. The factor that accelerates or decelerates the failure term also affects the hazard rate of an individual student. In this context, the hazard rate can be understood as the approximate probability at time x that a student with a given vector of covariates departs the STEM track in the next instant. It is important to note that it is not a true probability since the only limit on $h(x)$ is that it is ≥ 0 and thus mathematically it can be greater than 1.0.

The accelerating factor may be modeled to have a multiplicative effect or an additive effect. If the baseline hazard rate is $h_0(x)$ then an additive model would model the hazard rate for student i with a vector of p covariates $\mathbf{z}$ as $h_i(x \mid \mathbf{z}) = h_0(x) + \sum_{j=1}^{p} z_j(x)\beta_j(x)$. Similarly, a multiplicative model would model the hazard rate as $h_i(x \mid \mathbf{z}) = h_0(x)g(\boldsymbol{\beta}^t \mathbf{z})$ where $g(\boldsymbol{\beta}^t \mathbf{z})$ is a nonnegative function of the covariate vector.

A common multiplicative hazards model is the one proposed by Cox[91] with $g(\boldsymbol{\beta}^t \mathbf{z}) = \exp(\boldsymbol{\beta}^t \mathbf{z})$. The Cox model is often called the Proportional Hazards model[92] since the hazard for student j is a fixed proportion of the hazard for student i with $h_i(x \mid \mathbf{z}_i)/h_j(x \mid \mathbf{z}_j) = h_0(x)\exp(\boldsymbol{\beta}^t \mathbf{z}_i) / h_0(x)\exp(\boldsymbol{\beta}^t \mathbf{z}_j) = \exp[\beta_1(x_{i1} - x_{j1}) + \beta_2(x_{i2} - x_{j2}) + \ldots\ldots + \beta_p(x_{ip} - x_{jp})]$ since the baseline hazard rate cancels out. This model is considered to be a semi-parametric model because the cancellation of the baseline hazard rate means that only the parameters associated with the covariate factors are estimated. One of the advantages of this approach is that there is no need to select a probability distribution to model the survival times since the baseline hazard rate cancels out. Cox suggested a new approach for estimating the parameters called the maximum partial

likelihood method[93]. The expression "Cox Regression" is used to describe the utilization of the Cox model and the maximum partial likelihood estimation method. The Cox model can also be adapted for nonproportional hazards[94].

The choice of which model to employ depends on a number of aspects[95] of the application. SAS has the capacity to fit each of these models using Proc LIFEREG[96] for parametric models and Proc PHREG[97] for Cox models. One important aspect is whether the set of covariates includes any variables that change with time. Such variables are referred to as time-dependent variables. In medical applications variables that measure a patient's health at successive points in time are common and must be considered in modeling survival time. An example of this would be regular tests of a patient's platelet count. Proc PHREG is easily able to handle time-dependent covariates although this can be very complex if there are many of these variables. The size of the dataset and the frequency of tied event times is another aspect in the choice of model. Proc PHREG requires a great deal of computer processing time if the dataset is large and contains many identical event times. If the shape of the hazard function is of interest, then Proc LIFEREG may be more suitable since Proc PHREG does not directly calculate the baseline hazard function. If the data contains any left-censored values then Proc LIFEREG is the better choice since Proc PHREG does not allow for left-censoring. If predicting survival probabilities or event times for the dataset is of interest, Proc LIFEREG is more capable.

Proc LIFETEST[98] was used to estimate the hazard functions for the different departure type sub-populations in the overall sample of 11,128 students with the exception of the 433 students that were still in the college at the time the study ended. The departure times for the students still in school at the time the study ended were all right censored at the same date so the hazard function for that group would not have been informative. The hazard function curves

showed that several of the departure type sub-populations had markedly non-proportional differences between them.

Figure 6.1 indicates the estimated hazard functions by the departure type. Outcome 1 represents high school dropouts, outcome 2 represents high school graduates, outcome 3 represents college dropouts, outcome 4 represents students with sub-4 Year degrees, outcome 5 represents students with 4 year degrees other than STEM, and outcome 6 represents the STEM students. The graph is plotted so that Outcome 1 is the line closest to the vertical axis at its minimum value, Outcome 2 is the line next closest to the vertical axis at its minimum value, etc.

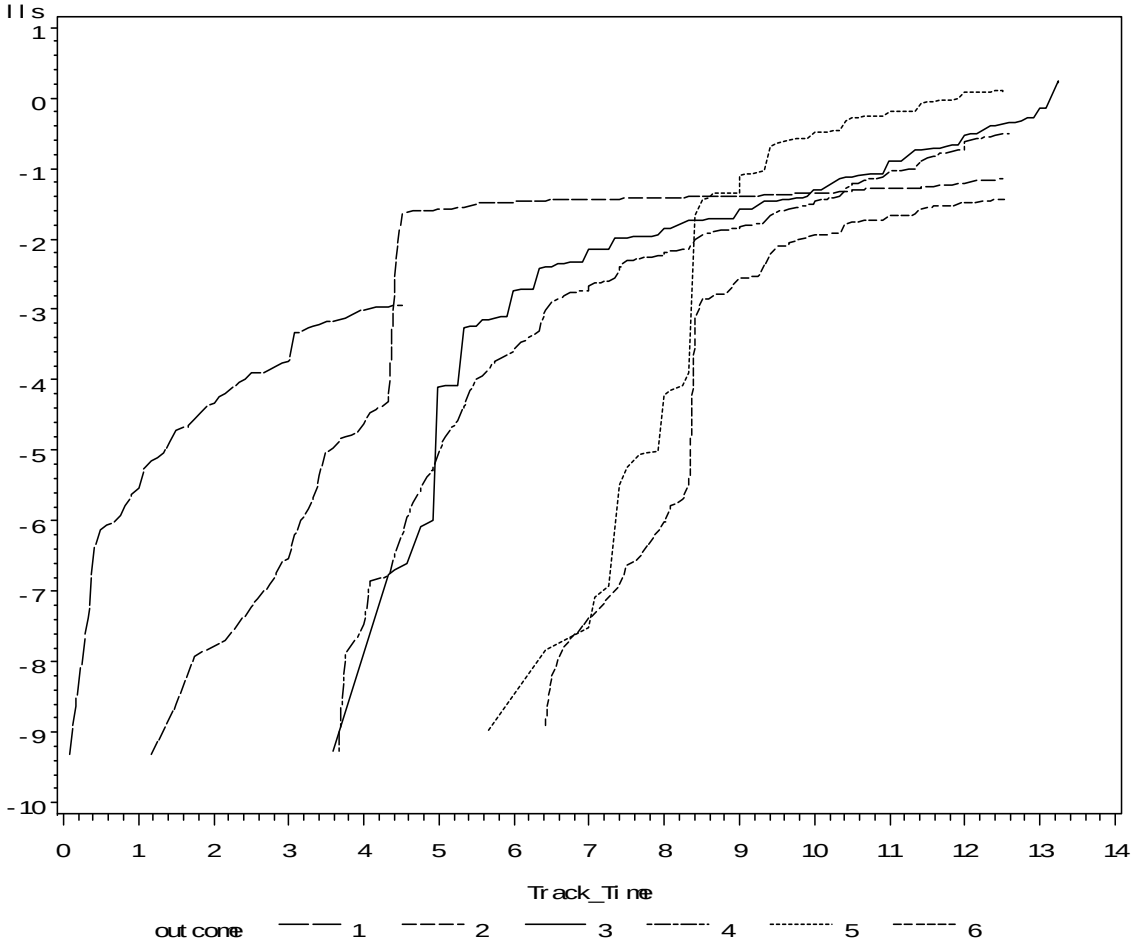

Figure 6.1 Hazard Functions by STEM Track Departure Type

All of the 76 potential covariates were collected in the base year with the exception of the standardized test scores that were obtained in the second follow up. The SAT, ACT, and potentially PSAT scores were obtained from tests the students took between the BY and F2 data collection waves. However, these scores were measured just once by collecting the data from the students during F2. It is possible the students took the tests multiple times, but only one set of scores for each standardized test was reported in the study. If the probability of a student departing the STEM track is estimated at a point in time after F2, these standardized test scores should not be considered time-dependent variables. The PETS transcript data contained enhanced variables for the SAT scores that the NELS:88 study designers created by replacing missing SAT scores with values derived from the students' PSAT scores, if available. There were 378 records across all the departure types that had missing F2RSATM or F2RSATV values for which imputed scores were available. The enhanced SAT scores from the PETS dataset were recoded using the same procedure described in Chapter 4 and substituted for the F2RSATM and F2RSATV variables in subsequent modeling.

Since the model's ability to predict student outcomes was of great importance in evaluating its worth, there were no time-dependent covariates, and there was a desire to compare the hazard functions between departure types, it was decided to use a parametric model for the survival times and to employ Proc LIFEREG to estimate the model parameters.

6.3.5 Fit and Test Sample Creation

The samples for the survival analysis module were created in the same manner used for the logistic regression module. The same 11 random number seeds used for the STEM vs. All Else logistic regression models were used to create samples for the fit and test process. The 11,128

records selected for analysis were first sorted by the students' study assigned ID number. Then a variable, "STEM_Outcome," was created to use in stratifying the population into STEM and All Else outcomes. The STEM_Outcome variable was set to "1" if the STEM track Departure_Type variable = 7 for a STEM degree outcome and "0" for all other departure types. Then the random number seeds were used to create fit data samples by randomly selecting 70% of the data from the STEM and All Else strata. Similarly, the test data samples were created from the 30% of the records not previously selected for the fit data samples. The result of this was 11 randomly chosen samples containing 7,791 records to fit models with and 11 randomly chosen samples containing 3,337 records to test the models developed.

These fit and test samples were not identical to those used for the original logistic regression models since they were created by stratifying the samples using different sorting variables. The original logistic regression models were created using the "STEM" variable and the classification solely by the variables for the first two college degrees and majors. As discussed earlier, this classification was adjusted in order to more precisely gather the time to event data for the survival analysis module. In order to permit direct comparison of the logistic regression model and integrated model prediction results, the logistic regression models for STEM vs. All Else were re-fitted using the new fit and test samples. This ensured that the exact same records were used for the iterations of model fitting and testing under both modeling approaches. The results from each modeling exercise are reviewed in Chapter 7.

6.3.6 Model Fitting

The modeling process began by selecting a probability distribution for the survival times. Proc LIFEREG was used to fit log-normal, log-logistic, gamma, exponential, and Weibull models

using all of the potential covariates as well as subsets that were found to be significant predictors in the logistic regression models. The models were fitted using both time to event measures, Track_Time with the origin at 1987.92 and Educ_Dur with the origin at the students' birth dates. The models were then compared by their likelihood ratio statistics. The log-logistic models consistently provided values with a smaller negative magnitude indicating better models. The standard 2-parameter gamma model provided the next best likelihood ratio statistics. Based on these results the log-logistic probability distribution was chosen for the survival analysis modeling portion of the integrated model. The models using the Track_Time as the time to event had consistently better goodness of fit statistics than those using the Educ_Dur times. Thus a decision was made to proceed with the modeling process using the study start time of 1987.92 as the origin point.

The LIFEREG procedure of SAS was used to build log-logistic models for each of the 11 fit data samples. The first model for each sample was fit using all 76 of the potential covariates plus the estimated probability of a STEM outcome from the logistic regression model. The latter variable, "LRprob_STEM" was an output from the fitted logistic regression model applied to both the fit and test data records for each random sample. Following the conclusions reached in the original logistic regression analysis, interaction terms were not employed in the model fitting.

Covariates whose estimated coefficients were not significantly different than 0 according to a chi-square test at the $\alpha = 0.05$ level were dropped from the model and the modeling was repeated with the subset of previously significant covariates. Subsequent iterations continued until all model covariates were found to be significant and the likelihood ratio statistic confirmed that the global test of model significance was met.

Once a final model for each fit data sample had been created, it was applied directly to the records to estimate the probability of survival on the STEM track past time 7.25 years. This point in time was chosen after examining the estimated hazard functions for the different sub-populations of students. Selecting a time earlier in the study would have resulted in a high probability of continuing to remain on the STEM track in the next instant for most of the students. Even selecting a time after most students had graduated high school and begun college would not have improved the ability to discriminate the STEM vs. All Else student sub-populations since most would still have had a high probability of surviving then.

The graph of the hazard functions suggested that most of the students that departed the STEM track tended to do so by 6.5 years past the origin time of 1987.92. By the early months of year 8 most of the other four year degree students and STEM students had graduated from college. The graph suggested the points at which the probability of survival on the STEM track was the highest for the STEM students and correspondingly lower for the other students was within the window of 6.33 to 8.17 years. The fit data for the original seed was repeatedly modeled to estimate the survival probability beyond time points 6.33, 6.41, 6.67, 7.25, 7.33, 7.41, 7.5, 7.67, 8.0, and 8.17 years. Based on the probability of survival past a given point, the student was predicted to have a STEM outcome for higher vs. lower probability values. Various cutpoints were used. The sensitivity and specificity of the predictions did not vary much, but slight improvements were found using 7.25 years. Thus this was the point in time used to discriminate between the STEM and All Else students based on the probability of remaining on the STEM track.

The final log-logistic model fitted via Proc LIFEREG for the original seed fit data was applied to the original seed test data. The final models for the other ten randomly chosen fit

datasets were applied to their associated test datasets in the same manner. Predicted outcomes were made in response to different cutpoints within the interval (0, 1).

In addition to these predictions, separate predictions were made for each record that reflected the results of both the logistic regression module and the survival analysis module. This approach examined the calculated LRprob_STEM from the logistic regression module and the estimated probability of survival beyond 7.25 years ("Prob"). If LRprob_STEM $\geq$ 0.07 and Prob > 0.5 then the integrated model predicted a STEM outcome. Otherwise the model predicted an All Else outcome. This was done to explore an alternative method of integrating the two modules.

6.4 ROC CURVE ANALYSIS MODULE

The models created for the same two-outcome pair were compared to one another by the area lying underneath the ROC curve associated with the fitted models known as "AUC" or "*c*." The value of AUC permits some comparison between models created for different two-outcome pairs. However, these sorts of comparisons were generally obtained by comparing the ROC curves produced from applying the models fitted with the test data directly to the reserved validation data.

The logistic regression models estimated a probability of a student with a given vector of covariates achieving a STEM outcome. The survival analysis models estimated the probability that a student with a given vector of covariates who had survived on the STEM track to time 7.25 years would remain on the track in the near future. The cutpoints used ranged from 0.01 to 0.99 and represented the dividing line between whether the model predicted a STEM vs. All Else

outcome given an estimated probability of within the range [0, 1]. Setting the cutpoint at a low value of 0.10 meant that most of the records had estimated probabilities above the cutpoint and were predicted to have a STEM outcome. This resulted in high sensitivity - correct identification of most of the STEM students, but a correspondingly low value of specificity – correct identification of most of the All Else students. Conversely, setting the cutpoint at a high value of 0.90 meant that few of the records were predicted as having a STEM outcome and most would be predicted as All Else. Models which had excellent predictive ability resulted in ROC curves with a very steep gradient indicating high probability of correct STEM prediction and a correspondingly low probability of incorrectly predicting All Else students as STEM.

6.5 SENSITIVITY ANALYSIS MODULE

The purpose of this module was to explore the effect of changing the prediction cutpoint on the accuracy of the STEM outcome predictions. The ROC Curve provides a range of cutpoint values for the logistic regression and the survival analysis modules. The visual depiction of the probability of correctly identifying potential STEM students that can be achieved with a corresponding loss in specificity is helpful in understanding how responsive or sensitive the model is to shifting the cutpoint. It should be noted that in this context the word sensitivity is not the same as the probability of a correct STEM prediction plotted on a ROC curve. Sensitivity analysis in this context refers to the effect on the overall application of selecting a specific probability cutpoint value for the model to achieve a particular policy goal or set of goals.

The selection of a preferred cutpoint to use in discriminating between outcomes is based on the objectives of the analyst. If the goal is to optimize sensitivity and specificity then both are

plotted against the range of cutpoint values and the cutpoint value at the intersection of the curves is selected. If the goal is to maximize the correct prediction of the outcome of interest, then the cutpoint can be chosen without regard to the probability of incorrect predictions. If the goal is to identify the largest population of potential STEM students for a proposed intervention program, then the cutpoint may be chosen based on the program budget available.

The results of the sensitivity analysis allow the modules to be "tuned" so that each produces a desired level of sensitivity when they are integrated.

6.6 THE INTEGRATED MODEL

The integrated model consisted of employing logistic regression, survival analysis, ROC curve analysis, and sensitivity analysis to develop predictions of the students' final educational outcome. The logistic regression module was used to estimate each student's probability of a STEM outcome when controlling for covariates. The output of that module, the original set of covariates, the time to STEM track departure, and a variable which indicates whether the departure time was observed or censored become inputs to the survival analysis module. The survival analysis module used the time to event data and the censored or observed status of the event time along with the regression covariates to estimate the probability for each student surviving on the STEM track beyond 7.25 years.

The probability of survival beyond 7.25 years was used to predict the students' final educational outcome with higher values of the probability leading to a STEM prediction and lower values leading to an All Else prediction. A set of predictions was generated based upon different values of the threshold cutpoint for the STEM track survival probability. The set of

predictions was analyzed with the ROC Curve module and the sensitivity was examined to identify the cutpoint producing the best combination of sensitivity and specificity. The best combination is determined by the goals of the analyst and a range of values is provided for selection. Another prediction was generated for each record based on combining the separate predictions of the survival analysis module and the best result from the logistic regression module. The analyst has the final choice of which set of predictions to choose.

6.6.1.1 Integrating the Modules in Series

The first model integrated the methods by linking the logistic regression and survival analysis modules in series by incorporating the output of the logistic regression module as an input covariate to the survival analysis module. The output of the logistic regression module was an estimated probability that a student with a given set of covariates would have a STEM educational outcome. This estimated probability was stored as a variable with the name LRprob_STEM and used an input for the survival analysis module. As with the logistic regression module, predictions were made for a large set of possible cutpoint values.

6.6.1.2 Integrating the Modules in Parallel

The second model built upon the first by using the output of both the series-linked model and the logistic regression module to make a prediction. There were two determining factors for the prediction by the integrated model using logistic regression and survival analysis in parallel. The first factor was having an estimated probability of a STEM outcome greater than or equal to 0.07 from the logistic regression module. The second factor was having an estimated probability of survival greater than or equal to 0.50 given the student had "survived" on the STEM track to

time 7.25 years. If both of these criteria were met then the integrated "combination" model predicted a STEM outcome.

The model integrated in parallel attempts to leverage the strength of the logistic regression module by using it as an input to the survival analysis module as well as obtaining a separate prediction from it that is compared to the prediction from the integrated in series model. Thus, both are used in making STEM predictions in cases where both models agree.

7.0 RESULTS

7.1 LOGISTIC REGRESSION MODEL PREDICTIONS WITH ORIGINAL DATASET CLASSIFICATION

The logistic regression models were built to predict the probability of a specific outcome vs. another one of the possible outcomes. For each of the two-outcome models, the model estimated the probability of the first outcome listed in the model name. In the case of the STEM vs. All Else model, the model estimated the probability of a STEM outcome. A value of 1.0 is the highest possible probability, and if such a value were estimated, the model automatically predicted that particular student would graduate with a STEM degree. A value of 0 is the lowest possible probability, and if the model estimated this value it automatically predicted the student would graduate with an All Else educational outcome. For values between 0 and 1 a decision had to be made about where to draw the line in predicting STEM vs. the alternative outcome. Overall, larger values of the resulting probability were set to predict a STEM outcome and smaller values were set to predict the alternative outcome. For example, if the cutpoint was chosen to be 0.10 then an estimated STEM probability of 0.12 would result in a STEM outcome prediction while an estimated STEM probability of 0.08 would result in an All Else outcome prediction.

The exact cutpoint at which to predict a STEM outcome vs. the other outcome was explored through ROC curves to illustrate the tradeoff between correct STEM predictions and incorrect STEM predictions. The numbers of correct and incorrect predictions of the two outcomes were assessed for different values of the cutpoint in order to choose a value that performed well. The area under the ROC Curve, denoted as AUC, was estimated as a way of assessing the predictive accuracy of the model. As discussed in Section 2.6, the AUC value lies between 0 and 1. Higher values denote better predictive accuracy in discriminating between the two outcomes. AUC values between 0.7 and 0.8 represent acceptable predictive ability. Values between 0.8 and 0.9 indicate excellent predictive ability. An AUC value over 0.9 suggests the model has an outstanding ability to discriminate between the outcomes.

7.1.1 STEM vs. All Else

The logistic regression model for STEM vs. All Else, fitted from the 76 recoded BY and F2 standardized score variables, was significant with an AUC value of 0.848 indicating excellent predictive discrimination. The significant variables were gender; Asian race; African-American race; overall math proficiency; how often parents talked to their child regarding post high school plans; how far the parents expected their child to advance; whether the student intended to attend a private non-religious high school; the highest level of education earned by the student's father; how far the student intended to advance in school; number of hours the student worked per week for pay; the student's ability group for math/science; the student's grades for math/science from 6^{th} to 8^{th} grade; ACT math and reading scores; SAT math and verbal scores; family composition; minority language status; and the student's base year science quartile standing. Figure 7.1 shows that high levels of correct STEM predictions can be achieved with relatively low levels of

incorrect STEM predictions for the STEM vs. All Else model when it is applied to the test data. The significant variables are indicated in the "original seed" column of Table 7.1 along with their estimated coefficients in the model.

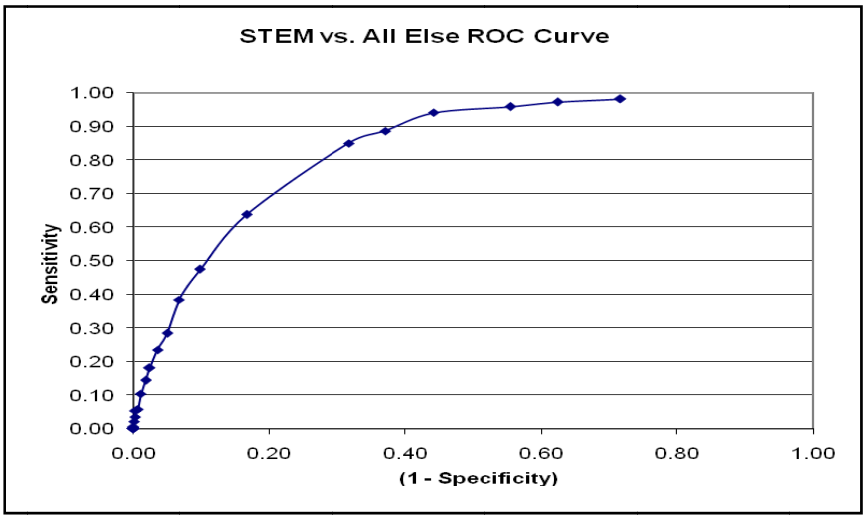

Figure 7.1 Sensitivity vs. (1-Specificity) for STEM vs. All Else model

The ROC curve in Figure 7.1 illustrates the joint levels of correct and incorrect STEM predictions that result when different probability cutpoints are used with the fitted logistic regression model for STEM vs. All Else. For example, the STEM vs. All Else model was able to correctly predict 50% of the STEM outcomes in conjunction with a 10% incorrect STEM prediction when applied to the test data. This means that while it properly predicted half of the STEM students to have a STEM outcome; it also incorrectly predicted 10% of the All Else students to have a STEM outcome. Using a different probability cutpoint, the model was also able to correctly predict 85% of the STEM outcomes in conjunction with a 35% incorrect STEM prediction. The preferred cutpoint depends upon the goals of the analyst. If a very high percentage of correct STEM predictions is sought then the accompanying percentage of incorrect STEM predictions will also be high.

7.1.1.1 Testing Model Stability

In order to assess the stability of the model a series of 10 additional random seeds were used to generate different fit and test data samples for the STEM vs. All Else model. This was done to determine how sensitive the models were to different selections of data used to fit the model. The total sample of 11,328 students contained 738 STEM students and 10,590 with another educational outcome. 70% of these records in each of the two strata were randomly selected a total of 11 times to form 11 separate samples for the model fitting data. Since the resulting STEM category was very small (517) compared to the All Else category (7,414) and the All Else outcome represented a very diverse set of outcomes there was concern that different samples would result in widely differing models. Table 7.1 shows the fitted models for the original seed and the 10 additional seeds used for the STEM vs. All Else analysis. Each of the variables that were found to be significant for at least one model is represented in the table. The estimated coefficients for each variable are provided in the table by seed. Please note that there is a fairly consistent set of significant variables across the 11 models and that their estimated coefficients are similar in value from model to model. This suggests that the overall STEM vs. All Else model is quite stable regardless of the particular set of NELS data used to fit it.

Table 7.1 Coefficients for Logistic Regression Models for STEM vs. All Else

Variable	Description	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Intercept	N/A, constant in logistic regression model equation	-7.4577	-8.3713	-8.0108	-4.9557	-4.6993	-8.9988	-4.9808	-4.7767	-7.674	-5.5652	-5.3724
BY2XMPROro	Overall Math Proficiency	0.1649		0.1345		0.1607	0.1145		0.1367		0.1446	0.1196
BY2XMQro	Mathematics Quartile		0.1831									
BY2XRPROro	Overall Reading Proficiency	-0.1752										
BY2XSQro	Science Quartile	0.1876	0.3205	0.261		0.1765	0.227	0.2976	0.2083	0.2811	0.1788	0.1718
BYFAMINCro	Yearly Family Income				-0.0342						-0.0367	
BYFCOMPr1	Family Composition: Mother & Male Guardian	-0.6366	-0.6494	-0.4777	-0.8264	-0.8187	-0.5121	-0.6047	-0.6338	-0.794	-0.691	-0.5789
BYFCOMPr3	Family Composition: Mother								-0.3537			
BYLMrb	Language Minority Composite				-2.7747	-3.6715		-2.6536	-3.3674		-2.5361	-2.7225
BYP64Bro	Family Rule re How Early/Late Child Watches TV						0.3437					
BYP65Aro	Family Rule About Child Maintaining Grade Avg.	0.2016										0.2627
BYP68ro	How Often Parent Talks To Child re Post H.S. Plans		-0.2262	-0.1439	-0.2234				-0.1872		-0.1797	
BYP76ro	How Far in School Parent Expects Child To Go		0.1087	0.0726	0.0971				0.0914		0.1056	
BYPARMARr1	Parents' Marital Status: Divorced	-0.5334		-0.5576			-0.4732	-0.7292				-0.5997
BYRISKro	# of BY Risk Factors for Dropping Out of School				-0.2334	-0.228					-0.2977	
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	-0.4956	-0.6623			-0.5965		-0.4893	-0.6229	-0.5523		-0.7805
BYS14rPvRel	H.S. Student Plans to Attend: Private Religious			0.3693								
BYS34Arb	Father's Highest Level of Education: College or Not				0.3018		0.3226		0.2427	0.3273		0.2633
BYS34Brb	Mother's Highest Level of Education: College or Not							0.2913				
BYS42Aro	# of Hrs Student Watches TV on Weekdays					-0.0657				-0.0675		
BYS43ro	# of Cigarettes Student Smokes per Day			0.3316		0.3072						
BYS45ro	How Far In School Do You Think You Will Get	0.3117	0.1623	0.2175	0.2585	0.3335	0.2891	0.2972	0.2031	0.213	0.2484	0.2386
BYS46ro	How Sure That You Will Graduate from H.S.	-0.5737						-0.4532				
BYS53ro	# of Hrs Student Works for Pay per Week	-0.1054	-0.1579	-0.1279	-0.1613	-0.1124		-0.1102		-0.1224	-0.1007	
BYS60Aro	Student's Ability Group for Mathematics	0.2222		0.2351	0.2018	0.2374	0.2314	0.1821	0.2035	0.2099	0.2368	0.1952
BYS60Bro	Student's Ability Group for Science	-0.1676		-0.1114	-0.1397	-0.1209	-0.1432		-0.1342	-0.1006	-0.1055	-0.0968

Table 7.1 (continued).

Variable	Description	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
BYS81Bro	Math Grades from Grade 6 Until Now	0.3016	0.3911	0.2879	0.3688	0.3494	0.3561	0.3048	0.2718	0.2951	0.3235	0.3712
BYS81Cro	Science Grades from Grade 6 Until Now	0.2652	0.2978	0.2925	0.2738	0.3477	0.3449	0.2989	0.3593	0.3519	0.3566	0.3457
BYSC13Erc	% of White Non-Hispanic 8th Graders											-0.0053
BYSC16Arc	# of Students in Free Lunch Program											-0.0011
F2RACTCro	ACT (Composite Score)							0.0537				
F2RACTEro	ACT (English Score)			-0.051			-0.0573	-0.0595	-0.0544	-0.0566	-0.0687	
F2RACTMro	ACT (Math)	0.0294	0.071	0.0756	0.0284	0.0683	0.079	0.0785	0.0776	0.078	0.0871	0.0708
F2RACTRro	ACT (Reading)		-0.0449			-0.0439		-0.0489				-0.0446
F2RSATMro	Scholastic Aptitude Test (Mathematics)	0.0041	0.0038	0.0040	0.0046	0.0033	0.0040	0.0043	0.0042	0.0040	0.0040	0.0044
F2RSATVro	Scholastic Aptitude Test (Verbal)	-0.0031	-0.0028	-0.0031	-0.0038	-0.0023	-0.0032	-0.0035	-0.0033	-0.003	-0.0031	-0.0033
F4RACE2rAs	Race of Student: Asian	0.5025		0.6394	0.3604	0.6508	0.5066	0.5625	0.3929	0.4499	0.3975	0.3901
F4RACE2rBl	Race of Student: African-American	0.5273				0.7608	0.6466		0.7341	0.4496		0.535
F4SEXrb	Sex of Student - binary (1 = Female)	-0.6649	-0.7138	-0.6595	-0.6689	-0.7299	-0.7352	-0.5911	-0.5811	-0.7234	-0.5811	-0.7487

The variables that were consistently significant across five or more models of STEM vs. All Else are listed in Table 7.2. The signs of the estimated coefficients were examined to determine if a particular variable had a positive or negative effect on the probability of a student earning a STEM degree. For example, having a higher overall math proficiency rating was a positive influence on achieving a STEM outcome by increasing the estimated probability of graduating with a STEM degree. Having a higher SAT verbal score was a negative influence since it decreased the estimated probability of graduating with a STEM degree. The categorization of these variable effects as positive or negative relates only to whether their contribution to the fitted model increased or decreased the probability of a STEM outcome, respectively.

Table 7.2 Effect of Consistently Significant Predictors of STEM vs. All Else

Variable	Description	Effect on Probability of STEM
Intercept	N/A, constant in logistic regression model equation	Negative
BY2XMPROro	Overall Math Proficiency	Positive
BY2XSQro	Science Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYLMrb	Language Minority Composite	Negative
BYP68ro	How Often Parent Talks To Child re Post H.S. Plans	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYPARMARr1	Parents' Marital Status: Divorced	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS60Aro	Student's Ability Group for Mathematics	Positive
BYS60Bro	Student's Ability Group for Science	Negative
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
F2RACTEro	ACT (English Score)	Negative
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F2RSATVro	Scholastic Aptitude Test (Verbal)	Negative
F4RACE2rAs	Race of Student: Asian	Positive
F4RACE2rBl	Race of Student: African-American	Positive

In addition, the models were tested with different potential sets of predictor variables. The different sets were all drawn from the same group of 76 recoded variables discussed in

Section 3.6. The three sets were Base Year only; Base Year plus F2 standardized scores; and Base Year, F2 standardized scores, and interaction terms for variables found to be significant in models created with the other two sets. The variables that were found to be significant when the interactions were included as potential predictors for the original seed are shown in Table 7.3. Interaction terms are labeled based upon the variables paired in the interaction using the format "Variable_1*Variable_2." This model had an AUC of 0.852 which was slightly larger than the AUC for the model without interaction terms (0.848). Negative coefficients decrease the estimated probability of a STEM outcome while positive coefficients increase it.

Table 7.3 Significant Variables for STEM vs. All Else with Interaction Testing

Parameter	Description	Coefficient Estimate
Intercept	N/A, constant in logistic regression model equation	-8.0700
BY2XMPROro	Overall Math Proficiency	-0.0928
BY2XSQro	Science Quartile	0.1767
BYFCOMPr1	Family Composition: Mother & Male Guardian	-0.6812
BYPARMARr1	Parents' Marital Status: Divorced	-0.5262
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	0.9866
BYS34Arb	Father's Highest Level of Education: College or Not	1.0789
BYS45ro	How Far In School Do You Think You Will Get	0.2798
BYS46ro	How Sure That You Will Graduate from H.S.	-0.5291
BYS60Aro	Student's Ability Group for Mathematics	0.7063
BYS60Bro	Student's Ability Group for Science	-0.1554
BYS81Bro	Math Grades from Grade 6 Until Now	0.2672
BYS81Cro	Science Grades from Grade 6 Until Now	0.2329
F2RACTMro	ACT (Math)	0.0702
F2RSATMro	Scholastic Aptitude Test (Mathematics)	0.0042
F2RSATVro	Scholastic Aptitude Test (Verbal)	-0.0033
F4RACE2rAs	Race of Student: Asian	0.5178
F4RACE2rBl	Race of Student: African-American	0.6482
F4SEXrb	Sex of Student - binary (1 = Female)	-0.6245
BY2XMPROro*BY2XSQro	Overall Math Proficiency * Science Quartile	0.1550
BY2XMPROro*BYS34Arb	Overall Math Proficiency * Father's Highest Level of Education: College or Not	-0.1969
BY2XMPROro*F2RACTMro	Overall Math Proficiency * ACT (Math)	-0.0090
BY2XSQro*BYS14rPvNRel	Science Quartile * H.S. Student Plans to Attend: Private Nonreligious	-0.4306
BY2XSQro*BYS60Aro	Science Quartile * Student's Ability Group for Mathematics	-0.1513
BYS34Arb*F2RACTMro	Father's Highest Level of Education: College or Not * ACT (Math)	-0.0214

The effect of interactions was tested in other models, but they were found to offer little improvement in predictive accuracy and are not presented for the other models.

In order to understand the impact of the cutpoint for these models it is helpful to see how many correct predictions were made based on the dividing line for the prediction. Table 7.4 shows the number of STEM students in the test data sample that were correctly predicted to have a STEM outcome by the STEM vs. All Else models for each seed. There were 221 STEM students in the test data samples and each model correctly predicted between 0 and 219 of them based on the cutpoint chosen.

Table 7.4 Number of STEM Students Out of 221 Correctly Predicted by Cutpoint and Seed

Cutpoint	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
0.010	217	214	215	218	214	215	217	219	215	218	218
0.015	215	210	211	216	212	212	211	219	212	213	215
0.020	212	203	208	214	210	210	211	216	210	203	211
0.030	208	201	201	206	204	205	201	210	198	195	206
0.040	196	191	192	197	194	197	195	201	190	185	194
0.050	188	180	185	186	186	194	189	194	184	175	183
0.100	141	127	145	128	147	145	142	153	138	141	132
0.150	105	99	107	100	111	106	101	114	96	108	99
0.200	85	75	86	71	86	78	75	76	74	82	77
0.250	63	50	72	47	61	59	57	62	56	65	52
0.300	52	38	53	36	46	43	43	51	47	45	40
0.350	40	28	45	26	39	30	30	42	37	36	30
0.400	32	24	35	18	30	23	22	35	29	27	26
0.450	23	19	22	15	22	19	16	23	23	19	18
0.500	13	14	15	11	19	14	12	16	18	12	14
0.550	12	8	11	7	12	9	8	10	12	8	8
0.600	8	5	7	6	5	6	6	6	10	5	5
0.650	5	2	5	4	5	4	4	5	6	3	4
0.700	1	1	3	2	2	4	3	4	4	1	1
0.750	0	1	2	0	0	1	3	1	2	1	1
0.800	0	0	0	0	0	0	2	1	0	1	0
0.850	0	0	0	0	0	0	0	0	0	0	0
0.900	0	0	0	0	0	0	0	0	0	0	0
0.950	0	0	0	0	0	0	0	0	0	0	0
0.990	0	0	0	0	0	0	0	0	0	0	0

As Table 7.4 shows, smaller cutpoint values correctly predicted more of the 221 STEM students across all 11 seeds than larger cutpoint values. However, while smaller cutpoint values

led the model to correctly predict most of the STEM students they also resulted in larger numbers of All Else students being incorrectly predicted as STEM. This can be visually depicted in plotting sensitivity and specificity. Figure 7.2 shows the sensitivity vs. specificity by cutpoint value. Note this is different from a ROC curve which plots sensitivity vs. (1 – specificity). If the goal is to optimize both of these values, then the ideal cutpoint is slightly less than 0.10. This assumes that the costs of correct and incorrect predictions should be balanced. This is a debatable concept that will be discussed in Chapter 8.0 . Another interesting point is how close the prediction was to the actual outcome. For example, if the model incorrectly predicted a STEM outcome for a particular student, was the real outcome a STEM-Related degree, a Non-STEM degree, a Sub-4 Yr Degree, or No Degree? Using the cutpoint of 0.10 for the original seed as an example, there were 531 All Else students that were incorrectly predicted to have a STEM outcome. Of these 24.9% had a STEM-Related outcome, 44.6% had a Non-STEM outcome, 6.4% had a Sub-4 Yr Degree outcome, and 24.7% had a No Degree outcome. So for this instance, 68.9% of the incorrect predictions went on to earn another college degree while 30.1% failed to achieve a four year degree. This suggests that even when the model incorrectly predicted a STEM outcome; it was picking up on characteristics that allowed the students in question to achieve a bachelor's degree. Undoubtedly, some of these students could have achieved a STEM degree had they chosen to pursue such a degree.

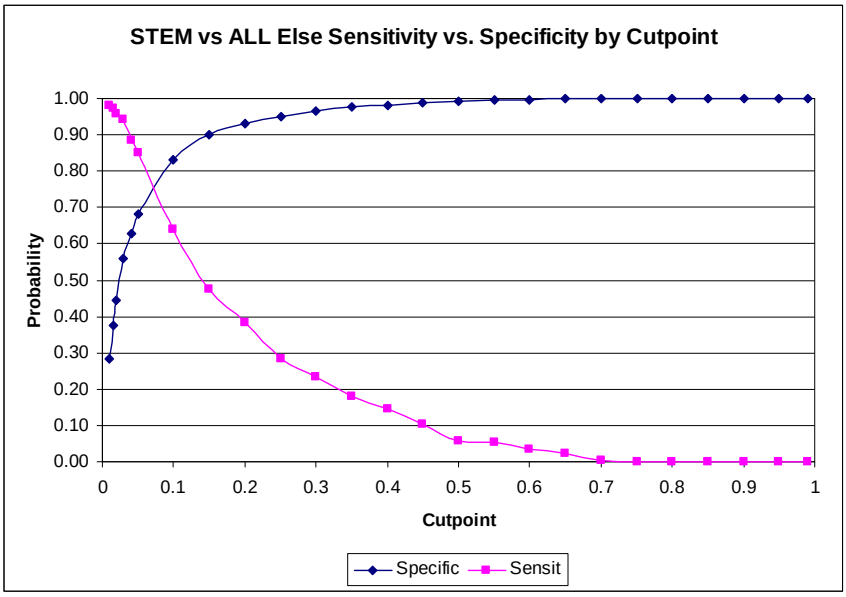

Figure 7.2 Sensitivity vs. Specificity by Cutpoint for STEM vs. All Else

The detailed analyses for the STEM vs. All Else model indicate that it is stable with a set of significant variables that is generally consistent. The students' proficiency in math and science, grades in these subjects, standardized test scores, and expectations of educational attainment are good predictors for this model. The fathers' highest level of educational achievement, parental expectations of the student, family structure, and family support for educational progress are also good predictors. Significant differences were found in race/ethnicity and gender for STEM vs. All Else students. The model was able to discriminate between STEM and All Else students with good predictive ability. The ROC curve indicated that nearly all the potential STEM students can be identified if a 50% error rate on the All Else students is acceptable. This can be adjusted as desired and as budgets allow.

Since the results of the stability testing for this model were encouraging, this level of detail was not pursued with each of the remaining models. Multiple random seeds were again tested with the STEM vs. STEM-Related model, but not with any of the others. The reasoning for this is that if educational outcomes are seen as having an ordered scale with No Degree at one

end and a STEM degree at the other end then the models for STEM vs. No Degree and STEM vs. STEM-Related represent the furthest and closest relationships, respectively. The All Else outcome is the most diverse compared to STEM, and STEM-Related is the least different compared to STEM. If these models are comparatively stable, then the ones "in between" should also be stable.

7.1.2 STEM vs. STEM-Related

The logistic regression model for STEM vs. STEM-Related also found numerous variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.720 indicating the fitted model possessed acceptable ability to discriminate between these two outcomes. The significant variables were gender; Asian race; overall math proficiency; having a family rule about how much time the student could spend watching television; how many hours the student spent watching television on weekends; the number of cigarettes smoked per day; how often parents talked to their child regarding post high school; number of hours the student worked per week for pay; the student's ability group for math/science; the student's grades for math/science from 6^{th} to 8^{th} grade; ACT math; SAT math scores; the student's base year science quartile standing; the percentage of white non-Hispanic 8^{th} graders; and the base salary of a beginning teacher with a B.A. degree at the student's school. The most consistent variables and their associated effect on the model's probability of a STEM outcome are shown in Table 7.5.

Table 7.5 Effect of Consistently Significant Predictors of STEM vs. STEM-Related

Variable	Description	Effect on Probability of STEM
Intercept	N/A, constant in logistic regression model equation	Negative
BY2XMPROro	Overall Math Proficiency	Positive
BY2XRQro	Reading Quartile	Negative
BY2XSQro	Science Quartile	Positive
BYP64Bro	Family Rule re How Early/Late Child Watches TV	Positive
BYP64Cro	Family Rule re How Many Hrs Child Watches TV	Negative
BYP64Dro	Family Rule re How Many Hrs on School Days Child Watches TV	Negative
BYS42Bro	# of Hrs Student Watches TV on Weekends	Negative
BYS43ro	# of Cigarettes Student Smokes per Day	Positive
BYS60Aro	Student's Ability Group for Mathematics	Positive
BYS60Bro	Student's Ability Group for Science	Negative
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
BYSC13Erc	% of White Non-Hispanic 8th Graders	Negative
BYSC19rc	Base Salary for Beginning Teacher w/ B.A.	Positive
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F4RACE2rAs	Race of Student: Asian	Positive
F4SEXrb	Sex of Student - binary (1 = Female)	Negative

Figure 7.3 shows that good levels of correct STEM predictions can be achieved with fair levels of incorrect STEM predictions for the STEM vs. STEM-Related model.

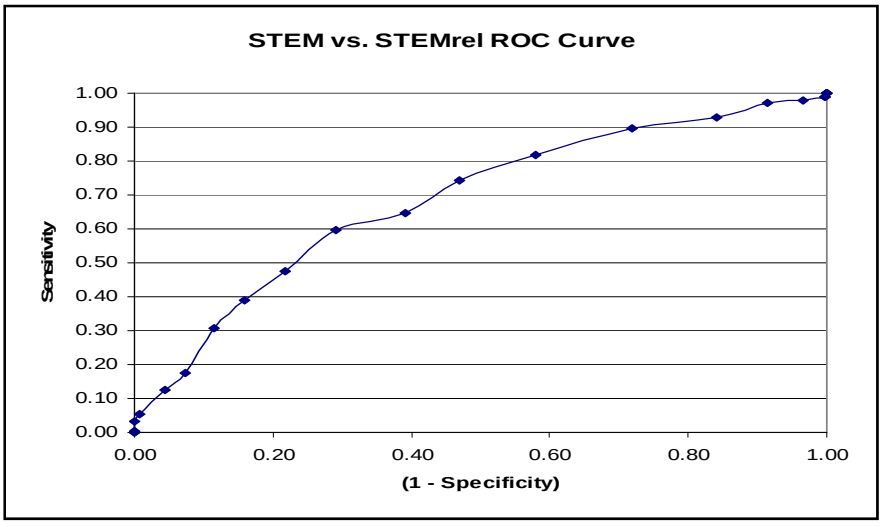

Figure 7.3 Sensitivity vs. (1-Specificity) for STEM vs. STEM-Related model

7.1.3 STEM vs. Non-STEM

The logistic regression model for STEM vs. Non-STEM found several variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.743 indicating the fitted model possessed acceptable ability to discriminate between these two outcomes. Table 7.6 lists the individual significant variables. The table indicates whether each variable's impact on the probability of a STEM outcome was positive (increasing the probability) or negative (decreasing the probability). The significant variables were gender; Asian race; African-American race; how often parents talked to their child regarding post high school plans; how far the parents expected their child to advance; how many hours the student spent doing homework per week; the student's grades for math/science from 6^{th} to 8^{th} grade; whether the student intended to attend a private non-religious high school; ACT math and English scores; SAT math and verbal scores; the student's base year science quartile standing; family composition; the parents' marital status – separated; and the base salary of a beginning teacher with a B.A. degree at the student's school.

Table 7.6 Effect of Significant Predictors of STEM vs. Non-STEM

Variable	Description	Effect on Probability of STEM
Intercept	N/A	Negative
BY2XSQro	Science Quartile	Positive
BYFCOMPr5	Family Composition: Other Relative/Nonrelative	Positive
BYHOMEWKro	# of Hours Spent on Homework per Week	Negative
BYP68ro	How Often Parent Talks To Child re Post H.S. Plans	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYPARMARr3	Parents' Marital Status: Separated	Positive
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
BYSC19rc	Base Salary for Beginning Teacher w/ B.A.	Positive
F2RACTEro	ACT (English Score)	Negative
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F2RSATVro	Scholastic Aptitude Test (Verbal)	Negative
F4RACE2rAs	Race of Student: Asian	Positive
F4RACE2rBl	Race of Student: African-American	Positive
F4SEXrb	Sex of Student - binary (1 = Female)	Negative

Figure 7.4 shows that good levels of correct STEM predictions can be achieved with fair levels of incorrect STEM predictions for the STEM vs. Non-STEM model.

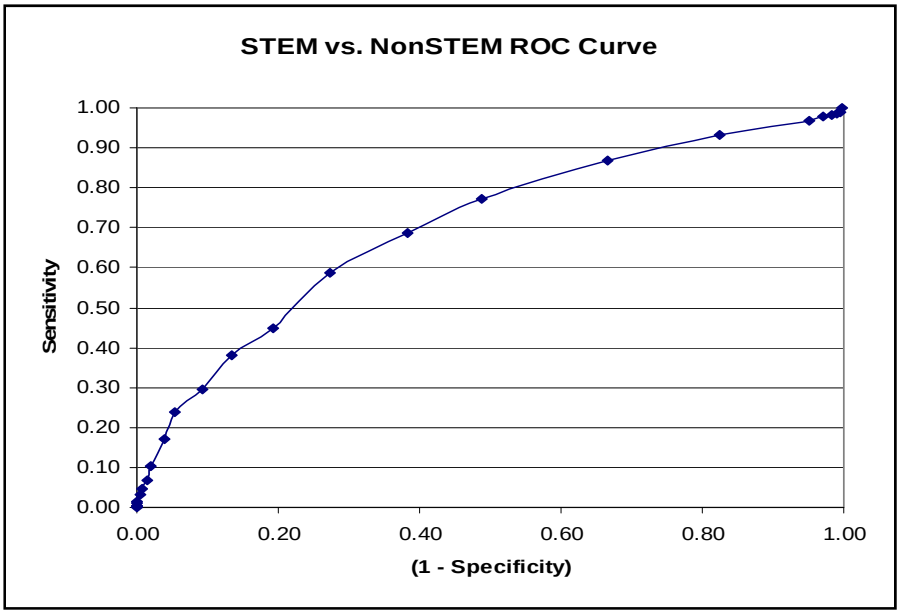

Figure 7.4 Sensitivity vs. (1-Specificity) for STEM vs. Non-STEM model

7.1.4 STEM vs. Sub 4-Yr Degree

The logistic regression model for STEM vs. Sub-4Yr Degree found several variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.924 indicating the fitted model possessed outstanding ability to discriminate between these two outcomes. The significant variables are listed in Table 7.7. They include gender; Asian race; African-American race; limited English proficiency; having a family rule about maintaining grade average; how often the parents helped the student with homework; how far the parents expected their child to advance; the highest level of education earned by the student's father; how far the student intends to advance in school; how sure the student was of graduating high school; how many hours the student spent doing homework per week; how many hours per week the student worked for pay; the number of cigarettes smoked per day; the student's ability group for mathematics; the student's grades for math/science from 6^{th} to 8^{th} grade; whether the student intended to attend a private non-religious high school; ACT math scores; SAT math scores; the student's base year mathematics quartile standing; and the number of students in the free lunch program at the student's school.

Table 7.7 Effect of Significant Predictors of STEM vs. Sub-4Yr Degree

Variable	Description	Effect on Probability of STEM
Intercept	N/A	Negative
BY2XMQro	Mathematics Quartile	Positive
BYHOMEWKro	# of Hours Spent on Homework per Week	Positive
BYLEPrb	Limited English Proficiency Composite	Negative
BYP65Aro	Family Rule About Child Maintaining Grade Avg.	Positive
BYP69ro	How Often Parent Helps Child with Homework	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYS14rPvRel	H.S. Student Plans to Attend: Private Religious	Positive
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS43ro	# of Cigarettes Student Smokes per Day	Positive
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS46ro	How Sure That You Will Graduate from H.S.	Negative
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS60Aro	Student's Ability Group for Mathematics	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
BYSC16Arc	# of Students in Free Lunch Program	Negative
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F4RACE2rAs	Race of Student: Asian	Positive
F4RACE2rBl	Race of Student: African-American	Positive
F4SEXrb	Sex of Student - binary (1 = Female)	Negative

Figure 7.5 shows that very good levels of correct STEM predictions can be achieved with correspondingly low levels of incorrect STEM predictions for the STEM vs. Sub-4Yr Degree model.

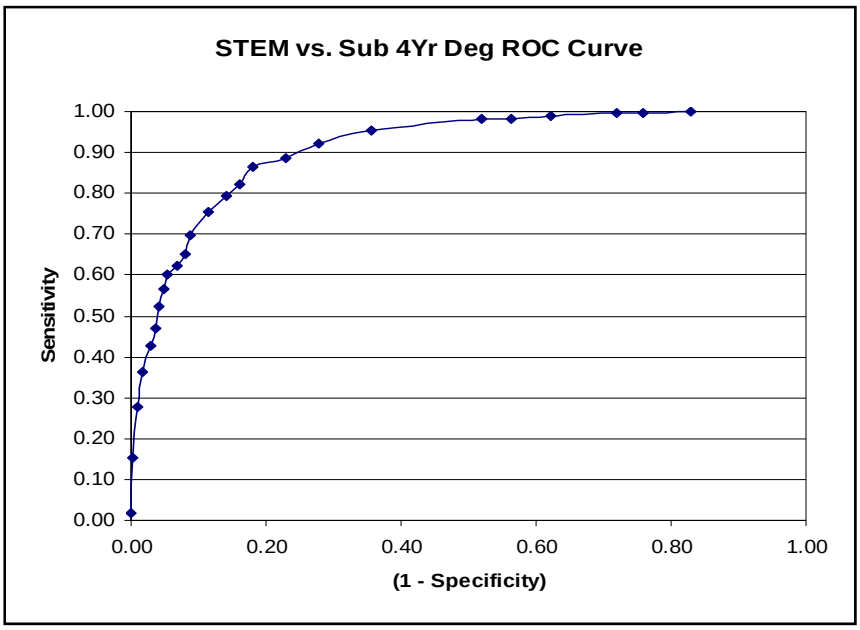

Figure 7.5 Effect of Significant Predictors of STEM vs. Sub-4Yr Degree

7.1.5 STEM vs. No Degree

The logistic regression model for STEM vs. No Degree found several variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.919 indicating the fitted model possessed outstanding ability to discriminate between these two outcomes. Table 7.8 lists the significant predictors. The significant variables were Asian race; language minority status; having a family rule about maintaining grade point average; having family rules about the student's television watching habits; the highest levels of education earned by the student's father and mother; how far the student intends to advance in school; how sure the student was of going further than high school; how many hours per week the student worked for pay; the student's grades for math, science and English from 6^{th} to 8^{th} grade; ACT English and math scores; SAT math and verbal scores; the student's base year mathematics and science quartile standing; family composition; and the number of students in remedial math at the student's school.

Table 7.8 Effect of Significant Predictors of STEM vs. No Degree

Variable	Description	Effect on Probability of STEM
Intercept	N/A	Negative
BY2XMQro	Mathematics Quartile	Positive
BY2XSQro	Science Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYLMrb	Language Minority Composite	Negative
BYP64Bro	Family Rule re How Early/Late Child Watches TV	Negative
BYP64Cro	Family Rule re How Many Hrs Child Watches TV	Positive
BYP64Dro	Family Rule re How Many Hrs on School Days Child Watches TV	Negative
BYP65Aro	Family Rule About Child Maintaining Grade Avg.	Positive
BYS34Aro	Father's Highest Level of Education - ordinal	Positive
BYS34Brb	Mother's Highest Level of Education: College or Not	Positive
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS47ro	How Sure Student Is To Go Further Than H.S.	Negative
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS81Aro	English Grades from Grade 6 Until Now	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
BYSC16Brc	# of Students in Remedial Reading	Negative
F2RACTEro	ACT (English Score)	Negative
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F2RSATVro	Scholastic Aptitude Test (Verbal)	Negative
F4RACE2rAs	Race of Student: Asian	Positive

Figure 7.6 shows that very good levels of correct STEM predictions can be achieved with correspondingly low levels of incorrect STEM predictions for the STEM vs. No Degree model.

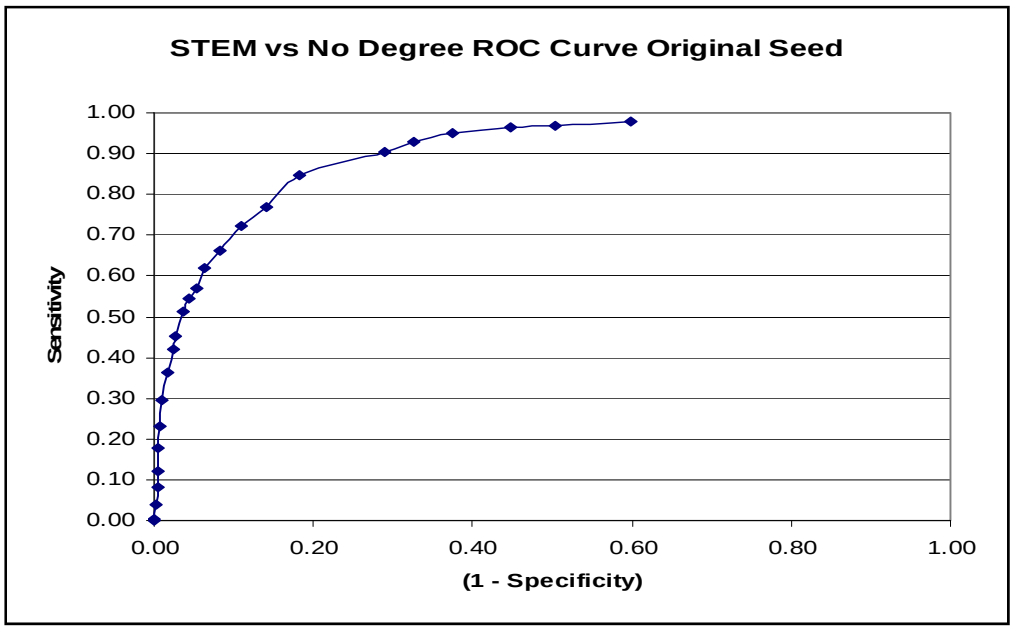

Figure 7.6 Sensitivity vs. (1-Specificity) for STEM vs. No Degree model

7.1.6 STEM vs. Other Degree

The logistic regression model for STEM vs. Other 4 Year Degree was of particular interest in this research since previous education research often defines the opposite of achieving a STEM degree as earning another four year degree. The STEM vs. Other 4 Year Degree and the STEM vs. All Else models are the ones that most closely compare to the modeling done by prior researchers.

The STEM vs. Other 4 Year Degree model found several variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.742 indicating the fitted model possessed acceptable ability to discriminate between the outcome of earning a STEM degree vs. earning another four year degree in a STEM-Related or Non-STEM major. The students earning four year degrees were more similar to one another so it was expected that this model would produce results similar to those of the STEM vs STEM-Related and STEM vs. Non-STEM models discussed earlier. Table 7.9 lists the significant predictors.

The significant variables were gender, Asian race; overall base year math proficiency; family composition; whether the student planned to attend a private non-religious high school; the student's ability group for mathematics; the student's grades for math from 6th to 8th grade; ACT English and math scores; and the student's SAT math and verbal scores.

Table 7.9 Effect of Significant Predictors of STEM vs. Other Degree

Variable	Description	Effect on Probability of STEM
Intercept	N/A	Negative
BY2XMPROro	Overall Math Proficiency	Positive
BY2XSQro	Science Quartile	Positive
BYFCOMPr5	Family Composition: Other Relative/Nonrelative	Positive
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS60Aro	Student's Ability Group for Mathematics	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
F2RACTEro	ACT (English Score)	Negative
F2RACTMro	ACT (Math)	Positive
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F2RSATVro	Scholastic Aptitude Test (Verbal)	Negative
F4RACE2rAs	Race of Student: Asian	Positive
F4SEXrb	Sex of Student - binary (1 = Female)	Negative

Figure 7.7 shows the resulting ROC curve from plotting sensitivity vs. (1 – specificity) for the STEM vs. Other 4 Year Degree model when applied to the test data. The graph shows that acceptable levels of correct STEM predictions can be achieved with fair levels of incorrect Other Degree predictions for the STEM vs. Other 4 Year Degree model. As expected, the ROC curve for this model was similar to those for the STEM vs. STEM-Related and STEM vs. Non-STEM models.

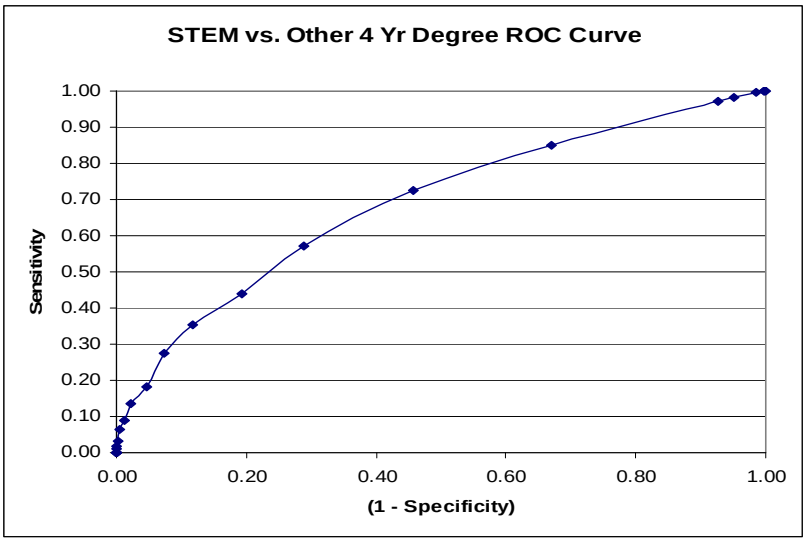

Figure 7.7 Sensitivity vs. (1-Specificity) for STEM vs. Other 4 Year Degree model

7.1.6.1 Analyzing Predictive Accuracy by Cutpoint

In examining the sensitivity of the model vs. its specificity for different cutpoints in discriminating between the outcomes, it appears that the optimization of both is achieved with a probability cutpoint between 0.1 and 0.2. This is illustrated in Figure 7.8. Note that this assumes that the cost of a proposed intervention program would lead to a policy of balancing sensitivity vs. specificity. If the goal were simply to maximize the number of potential STEM students that were reached, then the cutpoint could be chosen to identify the largest group of students the budget allocation would permit.

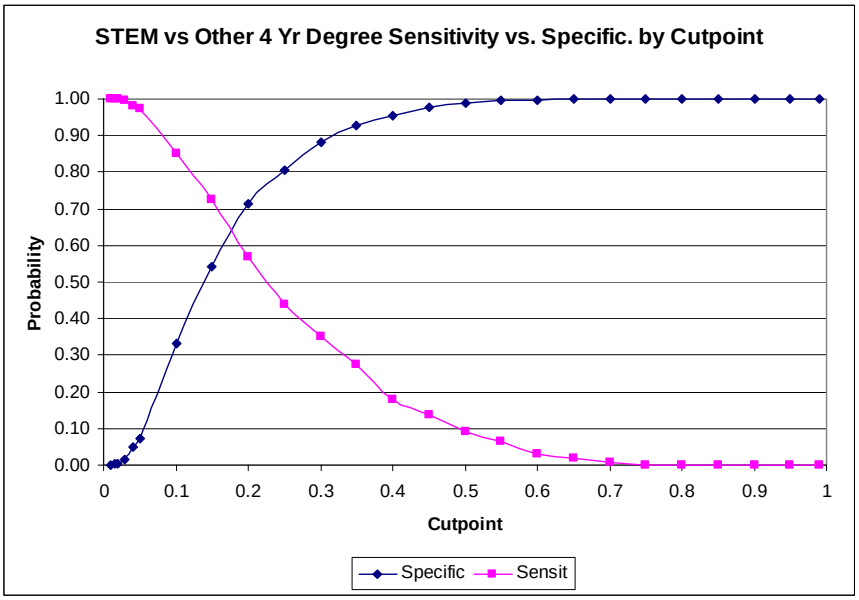

Figure 7.8 Sensitivity vs. Specificity by Cutpoint for STEM vs. Other 4 Year Degree

Exploring the accuracy of the model for these values we can examine the incorrect predictions to determine if the model is more likely to err in the case of one educational outcome or another. A total of 3,899 students out of the 11,328 in the sample earned four year degrees with 738 of them in STEM, 1,077 in STEM-Related, and 2,084 in Non-STEM topics. After 70% of the records were used to fit the model the remaining 30% of the data (1,169 records) was used to test the model. The test data contained 221 students with a STEM outcome and 948 students with an Other Degree outcome.

If a cutpoint of 0.20 is used then of the 1,169 students in the test data, 273 were incorrectly predicted to have a STEM outcome and 126 were correctly predicted to have a STEM outcome. Of these 273 Other Degree students 184 (67.3%) actually achieved a Non-STEM degree while the remaining 89 (32.6%) achieved a STEM-Related degree. These proportions were similar if the cutpoint was set to predict a STEM degree if the estimated probability were greater than or equal to 0.15. In that case, 434 Other Degree students were incorrectly predicted to have a STEM degree while 160 STEM students were correctly predicted to have that outcome.

Of the 434 incorrect STEM predictions 282 (65.0%) actually earned a Non-STEM degree and the remaining 152 (35.0%) achieved a STEM-Related degree.

The breakdown of the incorrect STEM predictions is not unexpected. The 1,077 STEM-Related and 2,084 Non-STEM students in the total sample represent 34.1% and 65.9% of the Other Degree category, respectively. This indicates that the STEM vs. Other Degree model incorrectly predicted the STEM-Related and Non-STEM students to have a STEM outcome at a roughly even rate. So if a student was incorrectly predicted to be STEM, this error was as likely to be made with a STEM-Related student as it was a Non-STEM student. The total number of incorrect STEM predictions closely matched the proportions of STEM-Related and Non-STEM students in the Other Degree sample.

7.1.7 Degree vs. Non-Degree

The logistic regression model for 4 year Degree vs. Non- 4 year Degree found several variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.882 indicating the fitted model possessed excellent ability to discriminate between these two outcomes. Table 7.10 lists the variables found to be significant predictors for this model. The significant variables include history and mathematics quartile; family composition composites; number of hours spent on homework per week; language minority composite; parental expectations for student achievement; frequency of parental discussions with child about future plans; number of risk factors for dropping out; the type of high school the student expected to attend; number of siblings; highest educational level of parents; student expectations of academic advancement; number of hours worked per week; English, math , & science grades; ACT math and science scores; SAT math score; Asian or Hispanic race; gender; and school characteristics of socioeconomic status and language minority percentage. In contrast to most of

the STEM models, sex was associated with a positive effect on the probability of a degree. This means female students were predicated to have a higher probability of achieving a four year degree.

Table 7.10 Effect of Significant Predictors of Degree vs. Non-Degree

Variable	Description	Effect on Probability of Degree
Intercept	N/A	Negative
BY2XHQro	History/Cit/Geog Quartile	Positive
BY2XMQro	Mathematics Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYFCOMPr5	Family Composition: Other Relative/Nonrelative	Negative
BYHOMEWKro	# of Hours Spent on Homework per Week	Positive
BYLMrb	Language Minority Composite	Negative
BYP68ro	How Often Parent Talks To Child re Post H.S. Plans	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYRISKro	# of BY Risk Factors for Dropping Out of School	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Positive
BYS14rPvRel	H.S. Student Plans to Attend: Private Religious	Positive
BYS32ro	Number of Siblings Student Has	Negative
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS34Brb	Mother's Highest Level of Education: College or Not	Positive
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS46ro	How Sure That You Will Graduate from H.S.	Negative
BYS47ro	How Sure Student Is To Go Further Than H.S.	Negative
BYS48Aro	How Far in School the Student's Father Wants Him/Her To Go	Positive
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS81Aro	English Grades from Grade 6 Until Now	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
BYSC16Arc	# of Students in Free Lunch Program	Negative
BYSC16Drc	# of Students in Bilingual Education	Positive
BYSC16Erc	# of Students in English as 2nd Language	Positive
F2RACTMro	ACT (Math)	Positive
F2RACTSro	ACT (Science Reasoning)	Negative
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F4RACE2rAs	Race of Student: Asian	Positive
F4RACE2rHi	Race of Student: Hispanic	Negative
F4SEXrb	Sex of Student - binary (1 = Female)	Positive

Figure 7.9 shows that good levels of correct four year Degree predictions can be achieved with correspondingly modest levels of incorrect Degree predictions for the Degree vs. Non-Degree model.

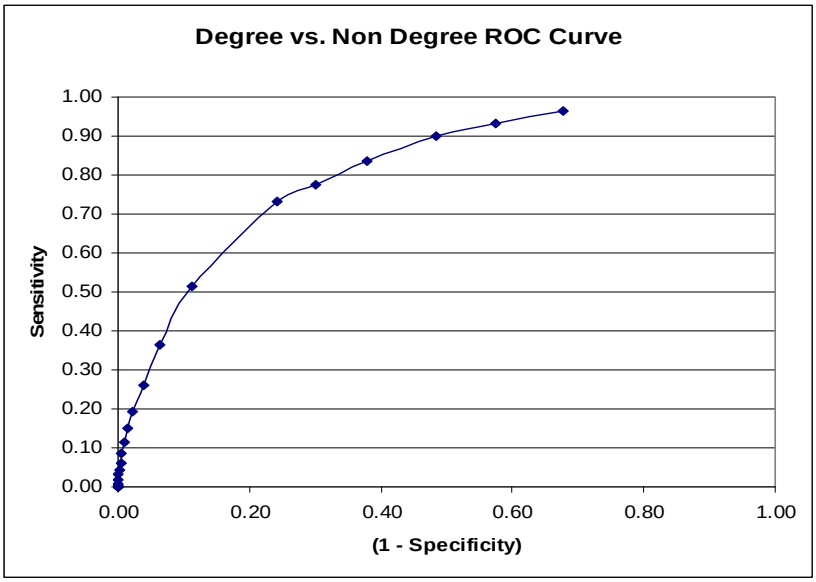

Figure 7.9 Sensitivity vs. (1-Specificity) for Degree vs. Non-Degree model

7.1.8 STEM-Related vs. Non-STEM

The logistic regression model for STEM-Related vs. Non-STEM found few variables to be statistically significant predictive factors. The AUC value for the associated ROC curve was 0.550 indicating the fitted model possessed negligible ability to discriminate between these two outcomes. The significant variables were language minority composite; whether the student intended to attend a private non-religious high school; and the student's grades for math from 6th to 8th grade. The significant variables and their associated effect on the model's probability of a STEM-Related outcome are shown in Table 7.11.

Table 7.11 Effect of Significant Predictors of STEM-Related vs. Non-STEM

Variable	Description	Effect on Probability of STEM-Related
Intercept	N/A	Positive
BYLMrb	Language Minority Composite	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS81Bro	Math Grades from Grade 6 Until Now	Positive

Figure 7.10 shows that the model is about as likely to predict a STEM-Related outcome as a Non-STEM outcome. The model possesses very little ability to correctly discriminate between the two educational outcomes. This is a very interesting result.

Consider that the STEM vs. STEM-Related, STEM vs. Non-STEM, and STEM vs. Other 4 Year Degree models possessed acceptable predictive accuracy. The finding that the STEM-Related vs. Non-STEM model had almost no predictive ability suggests that the STEM and STEM-Related students are more dissimilar than the STEM-Related and Non-STEM students. Therefore, if the STEM-Related category were to be discontinued it would make more sense to reclassify those students as Non-STEM rather than STEM.

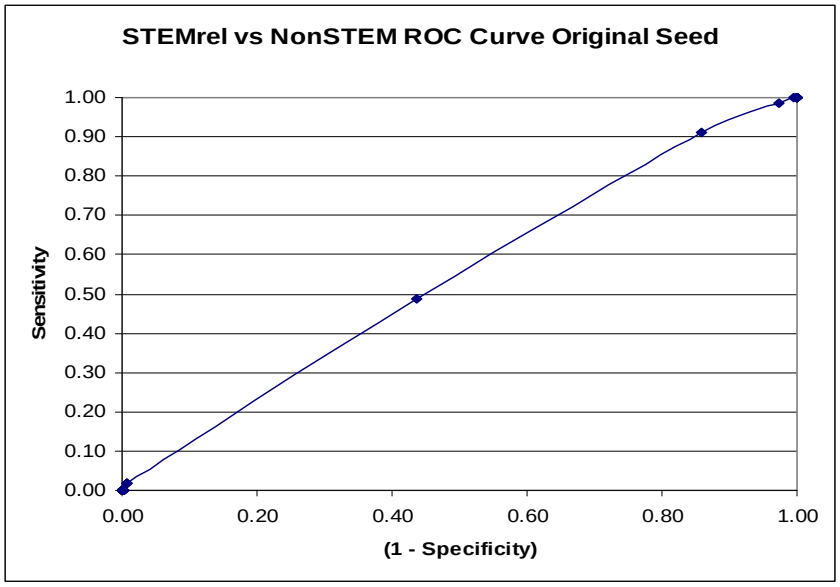

Figure 7.10 Sensitivity vs. (1-Specificity) for STEM-Related vs. Non-STEM model

7.1.9 STEM-Related vs. Sub-4 Yr Degree

The model fitted from the 76 recoded BY and F2 standardized score variables was significant with an AUC value of 0.885 indicating excellent predictive discrimination between the STEM-

Related and Sub-4 Year Degree outcomes. The significant variables were history and mathematics quartile; family composition; number of hours spent on homework per week; language minority composite; how far the parents expected their child to advance; the number of risk factors for dropping out; the type of high school the student expected to attend; the student's number of siblings; highest educational level of the parents; the number of hours per week the student watches television on weekdays; how far the student intends to advance in school; how far the student's father expects the student to advance; number of hours the student works per week for pay; the student's English and math grades from 6^{th} to 8^{th} grade; ACT reading score; SAT math scores; gender; and Hispanic race. The significant variables and their associated effect on the model's probability of a STEM-Related outcome are shown in Table 7.12.

Table 7.12 Effect of Significant Predictors of STEM-Related vs. Sub-4Yr Deg

Variable	Description	Effect on Probability of STEM-Related
Intercept	N/A	Negative
BY2XHQro	History/Cit/Geog Quartile	Positive
BY2XMQro	Mathematics Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYHOMEWKro	# of Hours Spent on Homework per Week	Positive
BYLMrb	Language Minority Composite	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYRISKro	# of BY Risk Factors for Dropping Out of School	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS14rPvRel	H.S. Student Plans to Attend: Private Religious	Positive
BYS17rb	Student Speaks Any Lang. Other Than English Before School	Positive
BYS32ro	Number of Siblings Student Has	Negative
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS34Brb	Mother's Highest Level of Education: College or Not	Positive
BYS42Aro	# of Hrs Student Watches TV on Weekdays	Negative
BYS47ro	How Sure Student Is To Go Further Than H.S.	Negative
BYS48Aro	How Far in School the Student's Father Wants Him/Her To Go	Positive
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS81Aro	English Grades from Grade 6 Until Now	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
F2RACTRro	ACT (Reading)	Negative
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F4RACE2rHi	Race of Student: Hispanic	Negative

Figure 7.11 shows that high levels of correct STEM-Related predictions can be achieved with relatively low levels of incorrect STEM-Related predictions for the STEM-Related vs. Sub-4 Year Degree model.

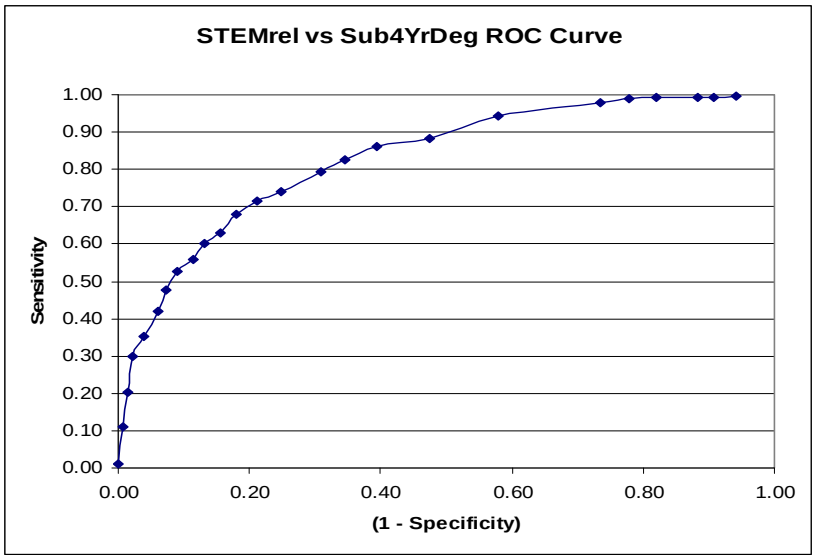

Figure 7.11 Sensitivity vs. (1-Specificity) for STEM-Related vs. Sub-4Yr Deg model

7.1.10 STEM-Related vs. No Degree

The model fitted from the 76 recoded BY and F2 standardized score variables was significant with an AUC value of 0.887 indicating excellent predictive discrimination between the STEM-Related and Sub-4 Year Degree outcomes. The significant variables were history and mathematics quartile; family composition composite variables; language minority composite; how often the parents assist the student with homework; the number of risk factors for dropping out; the type of high school the student expected to attend; the student's number of siblings; the father's highest educational attainment; the number of cigarettes the student smokes per; how far the student intends to advance in school; how far the student expects to advance in school; how sure the student is to go further than high school; number of hours the student works per week for pay; the student's English and math grades from 6th to 8th grade; % of white non-Hispanic 8th

graders in the student's school; % of 8th graders at the student's school in single parent families; ACT reading score; SAT math scores; Hispanic race; and gender. The significant variables and their associated effect on the model's probability of a STEM-Related outcome are shown in Table 7.13.

Table 7.13 Effect of Significant Predictors of STEM-Related vs. No Degree

Variable	Description	Effect on Probability of STEM-Related
Intercept	N/A	Negative
BY2XHQro	History/Cit/Geog Quartile	Positive
BY2XMQro	Mathematics Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYFCOMPr2	Family Composition: Father & Female Guardian	Negative
BYLMrb	Language Minority Composite	Negative
BYP69ro	How Often Parent Helps Child with Homework	Positive
BYRISKro	# of BY Risk Factors for Dropping Out of School	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Positive
BYS14rPvRel	H.S. Student Plans to Attend: Private Religious	Positive
BYS32ro	Number of Siblings Student Has	Negative
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS43ro	# of Cigarettes Student Smokes per Day	Negative
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS47ro	How Sure Student Is To Go Further Than H.S.	Negative
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS81Aro	English Grades from Grade 6 Until Now	Positive
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYSC13Erc	% of White Non-Hispanic 8th Graders	Positive
BYSC14ro	% of 8th Graders In Single Parent Family	Negative
F2RACTRro	ACT (Reading)	Negative
F2RSATMro	Scholastic Aptitude Test (Mathematics)	Positive
F4RACE2rAs	Race of Student: Asian	Positive
F4SEXrb	Sex of Student - binary (1 = Female)	Positive

Figure 7.12 shows that high levels of correct STEM-Related predictions can be achieved with relatively low levels of incorrect STEM-Related predictions for the STEM-Related vs. No Degree model.

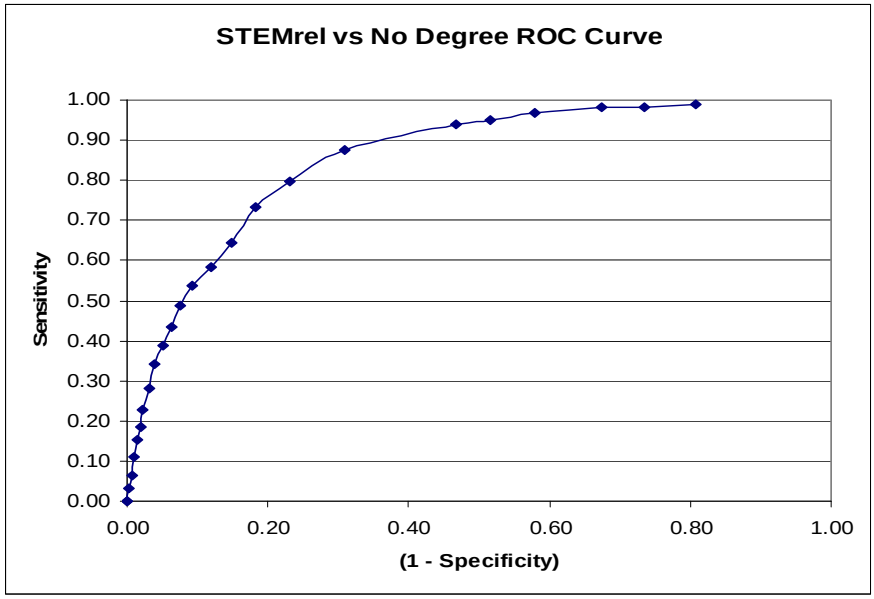

Figure 7.12 Sensitivity vs. (1-Specificity) for STEM-Related vs. No Degree model

Other models including Non-STEM vs. Sub 4 Yr Degree, Non-STEM vs. No Degree, and Sub 4 Yr Deg vs. No Degree were explored. However, the focus of the research was on STEM and STEM-Related students, and the discussion of the other models was not of sufficient interest to warrant its inclusion.

7.2 LOGISTIC REGRESSION PREDICTIONS FOR REVISED DATASET CLASSIFICATIONS

After the sample data was classified by STEM track departure type for survival analysis, several logistic regression models were re-fitted. The same random number seeds were used, but the resulting samples were slightly different since 200 records had been excluded from the analysis and the stratification of STEM vs. another outcome was based on the variable "STEM" initially and "STEM_Outcome" in the revised dataset. This was explained in greater detail in Section

6.3. The re-fitted models were for STEM vs. STEM-Related and STEM vs. All Else. Since these represented the least and most disparate comparisons, the results of re-fitting these models were judged to be sufficient to determine the effects of employing a new data classification approach. The results indicated that the models remained quite stable and did not change appreciably from the models created by with the initial dataset classification.

The STEM vs. All Else model had an AUC value of 0.852 for the Original Seed indicating excellent predictive ability between the STEM and All Else outcomes. The AUC for the other seeds ranged from 0.845 to 0.864 so the results were quite consistent across the multiple random samples. The ROC curve for the test sample using the original seed is shown in Figure 7.13. This model was able to correctly predict 50% of the STEM outcomes in conjunction with a 10% incorrect STEM prediction. The model was also able to correctly predict 85% of the STEM outcomes in conjunction with a 32% incorrect STEM prediction. These results are comparable to those obtained with the initial dataset classification. Equivalent values of Sensitivity and Specificity for this model were achieved at 76.6% when the probability cutpoint for prediction was set to 0.06. This can be seen in Figure 7.14. Table 7.14 summarizes the significant variables for this model across the eleven different random number seeds used to create the fit and test sample datasets.

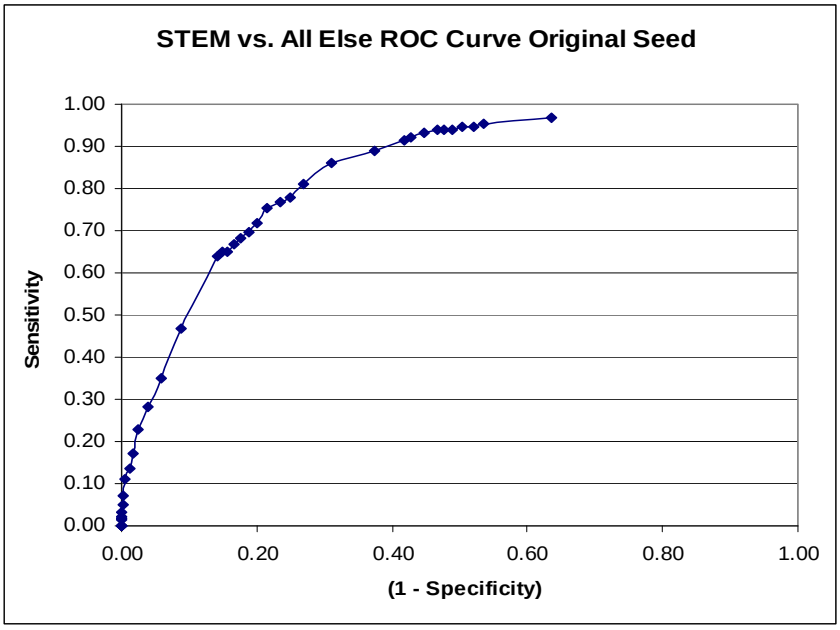

Figure 7.13 Sensitivity vs. (1-Specificity) for STEM vs. All Else model utilizing the Survival Analysis Classification Dataset

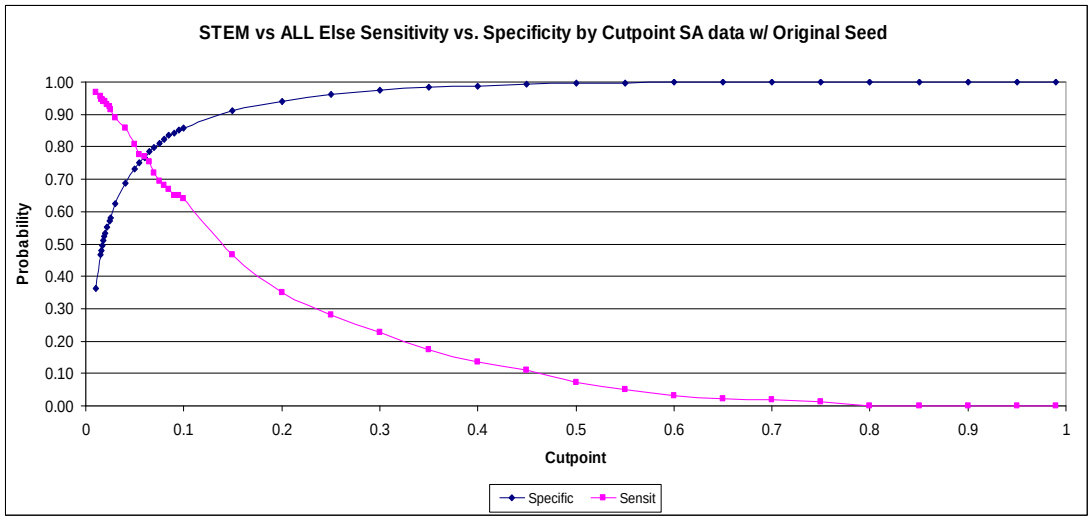

Figure 7.14 Sensitivity vs. Specificity by Cutpoint for STEM vs. All Else model utilizing the Survival Analysis Classification Dataset

Table 7.14 Coefficients of Logistic Regression Models for STEM vs. All Else for the Revised Dataset Classification

Variable	Description	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Intercept	n/a	-4.8602	-5.2736	-5.0606	-8.8106	-4.9600	-5.3720	-4.7395	-4.8744	-4.7164	-4.8951	-4.7939
BY2XMPROro	Overall Math Proficiency	0.1132	0.1419	0.1401	0.1226	0.1541	0.1637		0.1253			
BY2XMQro	Mathematics Quartile										0.2194	0.1808
BY2XRPROro	Overall Reading Proficiency		-0.1679									
BY2XRQro	Reading Quartile			-0.1561								
BY2XSQro	Science Quartile	0.2339	0.2613	0.3407	0.2353	0.3189	0.2158	0.2904	0.2750	0.2980	0.2258	0.2081
BYFAMINCro	Yearly Family Income						-0.0419					
BYFCOMPr1	Family Composition: Mother & male guardian	-0.4446	-0.5057	-0.5327		-0.5905	-0.6750	-0.5318	-0.5678	-0.7844	-0.5449	
BYFCOMPr3	Family Composition: Mother						-0.5303			-0.4060		
BYLMrb	Language Minority Composite	-2.7061	-2.7296	-3.3144		-2.7625	-2.8460	-2.1188	-2.7424	-4.3646	-2.4932	-3.8435
BYP64Bro	Family rule re how early/late child watches TV			0.3296								
BYP65Aro	Family rule about maintaining grade avg.	0.2246								0.1943		
BYP68ro	How often parent talks to child re post H.S. plans		-0.2012	-0.1955				-0.2096		-0.2439		-0.1661
BYP76ro	How far in school parent expects child to go		0.0808	0.0579			0.0734	0.0869		0.0503		0.0744
BYPARMARr1	Parents' Marital Status: Divorced		-0.7168						-0.4318		-0.4443	-0.6093
BYRISKro	# of BY Risk Factors for Dropping Out of School	-0.2066		-0.2888		-0.1874		-0.2679				
BYS14rPvNRel	HS Student Plans to Attend: Private Nonreligious		-0.7298	-0.6060		-0.6227		-0.4638	-0.5574	-0.5819	-0.4595	-0.5968
BYS14rPvRel	HS Student Plans to Attend: Private Religious	0.3428										
BYS34Arb	Father's Highest Level of Education: College or not	0.2186	0.2916		0.2443	0.3064	0.2213			0.3038	0.2304	0.2592
BYS42Aro	# of hrs Student watches TV on weekdays		-0.0783					-0.0651				
BYS43ro	# of Cigarettes Student Smokes per Day							0.3260		0.3359		
BYS45ro	How far in school do you think you will get	0.2305	0.2666	0.3558	0.3026	0.3070	0.2887	0.2370	0.3006	0.3795	0.2156	0.2354
BYS46ro	How sure that you will graduate from H.S.	-0.5709		-0.4770		-0.5043		-0.4769	-0.5156		-0.4777	
BYS53ro	# of Hrs student works for pay per week	-0.1551	-0.1383		-0.1088	-0.1078			-0.1143		-0.1221	-0.1232
BYS60Aro	Student's ability group for Mathematics	0.2087	0.1958	0.1636	0.2087		0.1868	0.2172	0.2341	0.2280	0.1896	0.2330
BYS60Bro	Student's ability group for Science	-0.0989	-0.1320	-0.1019	-0.1193		-0.1322	-0.1674	-0.1389	-0.1337	-0.1194	-0.1168
BYS81Bro	Math grades from Grade 6 until now	0.3401	0.3720	0.3361	0.3570	0.3790	0.2545	0.2493	0.2921	0.4043	0.2558	0.4109
BYS81Cro	Science grades from Grade 6 until now	0.2921	0.3237	0.3331	0.3648	0.2394	0.2569	0.2748	0.2922	0.3100	0.3339	0.2897

Table 7.14 (continued).

Variable	Description	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
F2RACTEro	ACT (English Score)	-0.0778	-0.0842	-0.0647	-0.0703	-0.0663	-0.0693	-0.0640	-0.0690	-0.0647	-0.0772	-0.0541
F2RACTMro	ACT (Math)	0.0967	0.1091	0.0868	0.0947	0.0874	0.0927	0.0899	0.0944	0.0860	0.0960	0.0741
SATQUANro	Scholastic Aptitude Test (Mathematics) expanded	0.0035	0.0014	0.0031	0.0038	0.0034	0.0047	0.0043	0.0031	0.0037	0.0033	0.0032
SATVERro	Scholastic Aptitude Test (Verbal) expanded	-0.0026		-0.0018	-0.0029	-0.0027	-0.0037	-0.0032	-0.0021	-0.0027	-0.0021	-0.0021
F4RACE2rAs	Race of student: Asian	0.6968	0.5047	0.6867	0.4720	0.4228	0.5754	0.6377	0.6238	0.3974	0.5867	0.4868
F4RACE2rBl	Race of student: African-American	0.6558	0.6621	0.6032	0.4644		0.4388	0.5628	0.4592	0.5752	0.5471	0.6136
F4SEXrb	Sex of student - binary (1 = Female)	-0.6007	-0.7214	-0.6506	-0.6521	-0.5958	-0.5780	-0.5655	-0.6268	-0.6082	-0.8168	-0.7142
AUC or "C"	Area under the ROC Curve	0.852	0.859	0.864	0.855	0.847	0.855	0.854	0.845	0.859	0.852	0.854

A visual comparison of Table 7.1 and Table 7.14 reveals that the sets of variables which were significant predictors for at least one fit data sample are virtually identical. The variables that had a positive effect on the probability of a STEM outcome in the initial classification scheme also have a positive effect under the revised classification scheme. Variables that previously exhibited a negative effect do the same under the revised classification scheme. The effect by variable is shown in Table 7.15. The estimated coefficients for the covariates are also similar between the two classification methods. The data classification method based on the survival analysis sample of 11,128 students grouped by departure type produced logistic regression models that are more consistent across the random samples in terms of the sets of significant predictors. These findings combined with the ROC curves produced by the models for the revised dataset suggest that the classification of student outcomes for the survival analysis module is acceptable. The excellent results obtained initially were not degraded by using the revised classification method.

Table 7.15 Effect of Consistently Significant Predictors of STEM vs. All Else for Revised Dataset Classification

Variable	Description	Effect on Probability of STEM
Intercept	N/A, constant in logistic regression model equation	Negative
BY2XMPROro	Overall Math Proficiency	Positive
BY2XSQro	Science Quartile	Positive
BYFCOMPr1	Family Composition: Mother & Male Guardian	Negative
BYLMrb	Language Minority Composite	Negative
BYP68ro	How Often Parent Talks To Child re Post H.S. Plans	Negative
BYP76ro	How Far in School Parent Expects Child To Go	Positive
BYPARMARr1	Parents' Marital Status: Divorced	Negative
BYS14rPvNRel	H.S. Student Plans to Attend: Private Nonreligious	Negative
BYS34Arb	Father's Highest Level of Education: College or Not	Positive
BYS45ro	How Far In School Do You Think You Will Get	Positive
BYS53ro	# of Hrs Student Works for Pay per Week	Negative
BYS60Aro	Student's Ability Group for Mathematics	Positive
BYS60Bro	Student's Ability Group for Science	Negative
BYS81Bro	Math Grades from Grade 6 Until Now	Positive
BYS81Cro	Science Grades from Grade 6 Until Now	Positive
F2RACTEro	ACT (English Score)	Negative
F2RACTMro	ACT (Math)	Positive
SATQUANro	Scholastic Aptitude Test (Mathematics)	Positive
SATVERro	Scholastic Aptitude Test (Verbal)	Negative
F4RACE2rAs	Race of Student: Asian	Positive
F4RACE2rBl	Race of Student: African-American	Positive
F4SEXrb	Sex of student – binary (1 = Female)	Negative

Another question that arose was how this model would perform if the standardized test scores were not included in the model fitting. The STEM vs. All Else model fitted with solely BY variables and the revised dataset classification was examined for the original seed. The resulting AUC value for the fitted model was 0.833 as opposed to 0.852 for model including the standardized test scores from F2. Table 7.16 compares the models fitted with and without the standardized test scores for the original seed. The set of significant variables and estimated coefficients of the fitted model are very similar other than the absence of the SAT and ACT test score variables.

Table 7.16 Coefficients of Logistic Regression Models for STEM vs. All Else for the Revised Dataset Classification Original Seed with and without F2 Standardized Test Scores

Recoded Var	Variable Description	Original Seed w/ std. tests	Original Seed w/o std. tests
Intercept	n/a	-4.8602	-4.6849
BY2XMPROro	Overall Math Proficiency	0.1132	0.1838
BY2XSQro	Science Quartile	0.2339	0.2333
BYFCOMPr1	Family Composition: Mother & male guardian	-0.4446	-0.4662
BYLMrb	Language Minority Composite	-2.7061	-3.2829
BYP65Aro	Family rule about maintaining grade avg.	0.2246	0.2398
BYPARMARr3	Parents' Marital Status: Separated		0.7215
BYRISKro	# of BY Risk Factors for Dropping Out of School	-0.2066	-0.2674
BYS14rPvRel	HS Student Plans to Attend: Private Religious	0.3428	0.4342
BYS34Arb	Father's Highest Level of Education: College or not	0.2186	0.3006
BYS45ro	How far in school do you think you will get	0.2305	0.2743
BYS46ro	How sure that you will graduate from H.S.	-0.5709	-0.6566
BYS53ro	# of Hrs student works for pay per week	-0.1551	-0.1544
BYS60Aro	Student's ability group for Mathematics	0.2087	0.2373
BYS60Bro	Student's ability group for Science	-0.0989	-0.1003
BYS81Bro	Math grades from Grade 6 until now	0.3401	0.4263
BYS81Cro	Science grades from Grade 6 until now	0.2921	0.3127
F2RACTEro	ACT (English Score)	-0.0778	N/A
F2RACTMro	ACT (Math)	0.0967	N/A
F4RACE2rAs	Race of student: Asian	0.6968	0.8879
F4RACE2rBl	Race of student: African-American	0.6558	0.5592
F4SEXrb	Sex of student - binary (1 = Female)	-0.6007	-0.6937
SATQUANro	Scholastic Aptitude Test (Mathematics) expanded	0.0035	N/A
SATVERro	Scholastic Aptitude Test (Verbal) expanded	-0.0026	N/A
AUC or "C"	Area under the ROC Curve	0.852	0.833

Figure 7.15 shows the comparison of the ROC Curves obtained when the fitted models were applied to the test data for the original seed. The ROC Curve for the model utilizing the standardized test scores has a slightly steeper gradient and encompasses more area under the curve than that of the model without the standardized test scores. However, the curves are very similar. This suggests that using variables obtained in the 8th grade produces a model that has slightly less predictive accuracy than a model which also includes the SAT and ACT test scores as potential predictor variables.

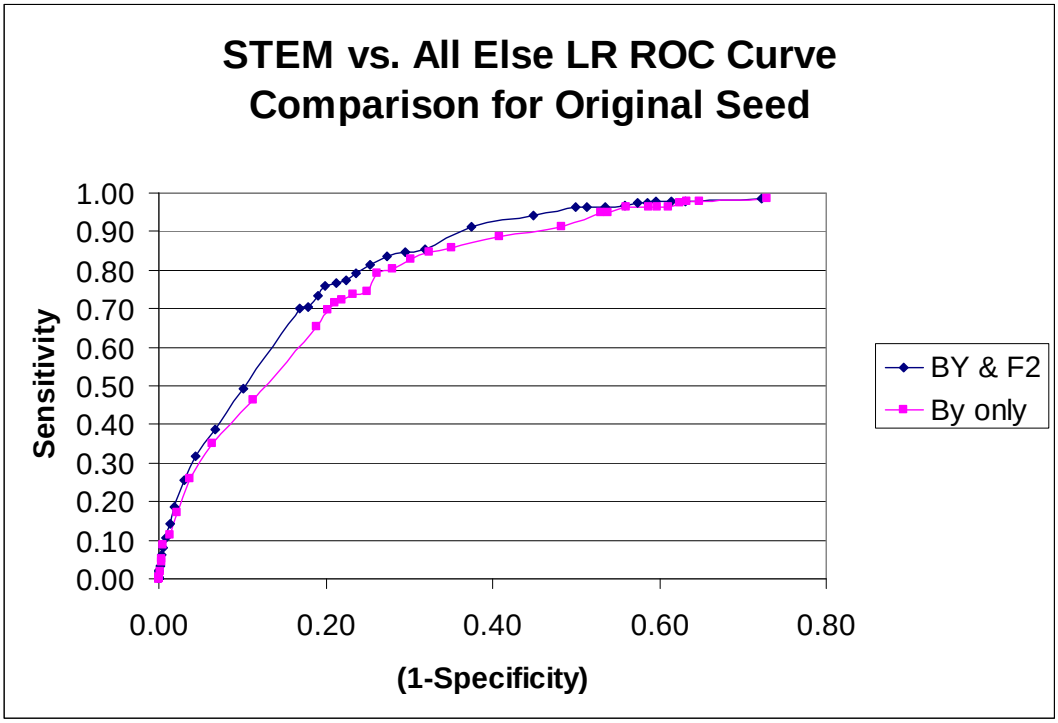

Figure 7.15 Sensitivity vs. (1-Specificity) for STEM vs. All Else model comparing the models utilizing the Survival Analysis Classification Dataset with and without Standardized Test Scores

The STEM vs. STEM-Related model had an AUC value of 0.722 for the original seed indicating good predictive ability between the STEM and STEM-Related outcomes. The AUC for the other seeds ranged from 0.699 to 0.732 so the results were consistent across the multiple random samples. The ROC curve for the test sample using the original seed is shown in Figure 7.16. Equivalent values of Sensitivity and Specificity for this model were achieved at 62.3% when the probability cutpoint for prediction was set to 0.40. This can be seen in Figure 7.17. The equivalent values for the other seeds ranged from approximately 60.4% to 65.4%.

Table 7.17 summarizes the significant variables for this model across the eleven different random number seeds used to create the fit and test sample datasets.

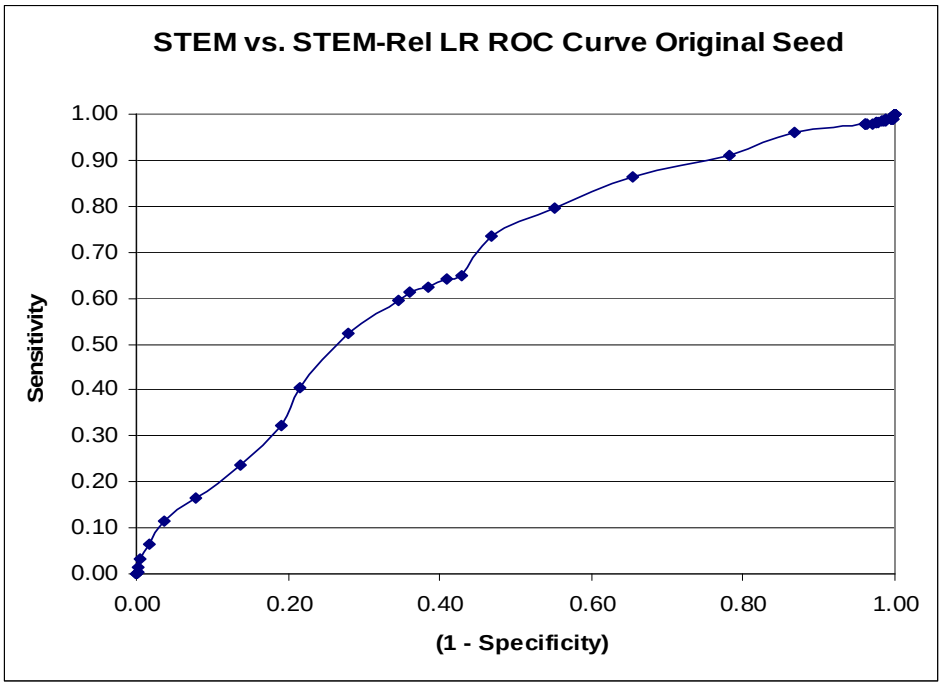

Figure 7.16 Sensitivity vs. (1-Specificity) for STEM vs. STEM-Related model utilizing the Survival Analysis Classification Dataset

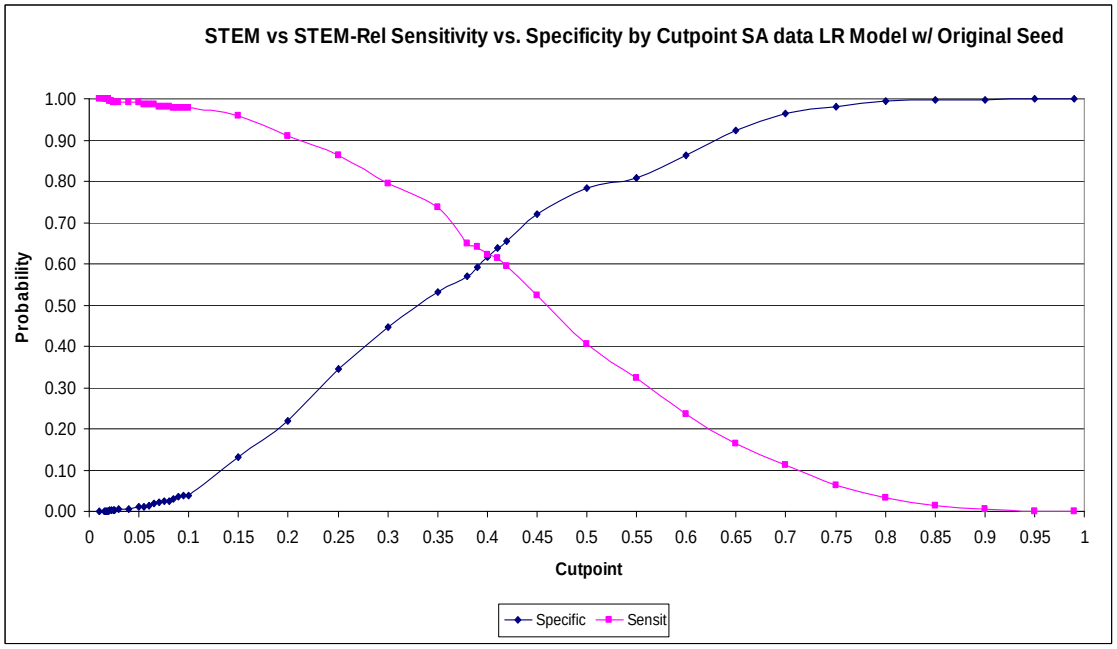

Figure 7.17 Sensitivity vs. Specificity by Cutpoint for STEM vs. STEM-Related model utilizing the Survival Analysis Classification Dataset

Table 7.17 Coefficients for the Logistic Regression Models for STEM vs. STEM-Related for the Revised Dataset Classification

Variable	Description	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Intercept	n/a	-2.1054	-2.2206	-1.9501	-2.5305	-2.6152	-2.5799	-3.4455	-2.5924	-2.4288	-2.8537	-2.0914
BY2XMPROro	Overall Math Proficiency			0.1716				0.1977	0.1480	0.1413	0.1647	0.1267
BY2XRQro	Reading Quartile	-0.2171				-0.2406			-0.2767			
BY2XSQro	Science Quartile	0.3311	0.2242		0.1546	0.3107	0.2168		0.3453	0.1762		
BYFCOMPr5	Family Composition: Other Relative/Nonrelative	1.3040	1.5796						1.8331	1.5149	1.5637	
BYHMLANGro	Home Language Background	-0.1895										
BYP64Bro	Family rule re how early/late child watches TV		0.6199									
BYP64Dro	Family rule how many hrs on school days child watches TV		-0.4053									
BYP69ro	How often parent helps child with homework			-0.1403	-0.1228							
BYPARMARr3	Parents' Marital Status: Separated								1.1318			1.2539
BYS14rPvNRel	HS Student Plans to Attend: Private Nonreligious		-0.6051		-0.4951			-0.5699				
BYS43ro	# of Cigarettes Student Smokes per Day	0.6046	0.6474	0.3679						0.6401	0.6713	0.4292
BYS45ro	How far in school do you think you will get							0.1644				
BYS53ro	# of Hrs student works for pay per week					-0.1344						
BYS60Aro	Student's ability group for Mathematics	0.1885	0.2161	0.1368		0.1312		0.2080		0.1363		0.2090
BYS60Bro	Student's ability group for Science		-0.1332					-0.1293				
BYS81Bro	Math grades from Grade 6 until now	0.2040			0.2515	0.2347	0.3652			0.2629	0.2309	
BYS81Cro	Science grades from Grade 6 until now	0.2483	0.2678	0.2565	0.1961	0.2431		0.1670	0.3200		0.1891	0.3550
BYSC13Erc	% of White Non-Hispanic 8th Graders	-0.0066	-0.0065	-0.0070	-0.0090	-0.0085	-0.0089		-0.0091	-0.0059	-0.0053	-0.0059
BYSC15ro	% of 8th Graders Limited English Proficient					-0.2769					-0.3228	
BYSC16Erc	# of students in English as 2nd Language				-0.0057				-0.0041			
BYSC19rc	Base Salary for Beginning Teacher w/ B.A.			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0001	
F2RACTMro	ACT (Math)		0.0178	0.0177	0.0157	0.0166	0.0148	0.0154	0.0214	0.0120	0.0119	0.0139
F2RACTSro	ACT (Science Reasoning)	0.0149										
F4RACE2rAs	Race of student: Asian				0.4295			0.5510	0.6624			0.6114
F4RACE2rBl	Race of student: African-American							0.5774				
F4SEXrb	Sex of student - binary (1 = Female)	-0.9327	-0.9302	-0.9800	-0.9936	-1.0265	-0.9639	-0.8932	-0.9571	-0.9528	-0.8861	-0.9818
SATQUANro	Scholastic Aptitude Test (Mathematics) expanded	0.0012	0.0009	0.0006	0.0009	0.0012	0.0008	0.0006	0.0009	0.0010	0.0025	
SATVERro	Scholastic Aptitude Test (Verbal) expanded										-0.0022	
AUC or "C"	Area under the ROC Curve	0.722	0.717	0.700	0.713	0.728	0.699	0.704	0.732	0.715	0.705	0.722

The positive or negative impact of the most frequently significant predictors across the samples was consistent with what was found in previous models. The probability of a STEM outcome is enhanced by strong mathematical capability; good academic performance in math and science; and higher standardized test scores in mathematics. The models created with the data reclassified for the survival analysis had comparable predictive ability to those created with the original classification scheme. Overall, the predictive strength of the models was clearly stronger for the models where the two-outcome pairs were more divergent.

7.3 INTEGRATED MODEL PREDICTIONS

Integrated models were fitted to predict STEM. vs. All Else for the same eleven random number seeds so that the accuracy of the predictions with the integrated model could be directly compared to that of the logistic regression model. ROC Curves were created to examine the sensitivity of the integrated model using the logistic regression and survival analysis linked in series. The integrated model using these techniques in parallel was applied to the same random samples with cutpoints of 0.07 for the logistic regression portion and 0.50 for the survival analysis portion. The cutpoint of 0.07 was based on a decision to balance the sensitivity and specificity of the logistic regression module when making a prediction that would be compared to the survival analysis module prediction. The cutpoint of 0.50 for the survival analysis module was chosen to offer an optimistic prediction that students still on the STEM track would remain on it. The combination of these cutpoints directed each module to predict a STEM outcome while balancing the probability of an incorrect prediction since a final STEM prediction would require both modules to agree. This lowered the chance of a true STEM student being

incorrectly predicted to have an All Else outcome because one of the two modules did not predict the student to have a STEM outcome.

7.3.1 Integrated Model in Series Results

Figure 7.18 shows the resulting ROC Curve for the integrated model fitted from the original seed applied to the test data sample in comparison to the ROC Curve for the logistic regression model. The ROC curves in this graph are plotted in a smoothed format just to show the comparison between them. The curve for the integrated model is less steep than that of the revised logistic regression model indicating that its accuracy is weaker. A sensitivity of 69.1% is achievable with a corresponding (1 - Specificity) value of approximately 26.9%. This is acceptable discrimination ability, but it is not as precise as the 70.0% sensitivity with a corresponding (1 - Specificity) value of approximately 16.8% provided by the logistic regression model. The logistic regression model created from the original seed fit data sample performed better when applied to the associated test data sample.

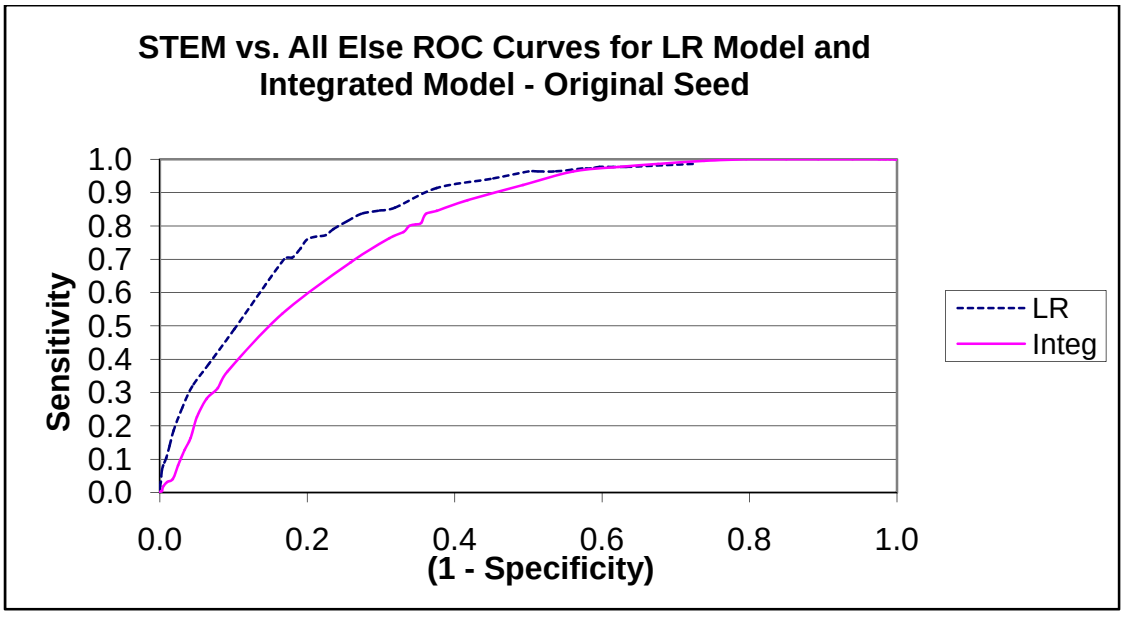

Figure 7.18 Sensitivity vs. (1-Specificity) for Integrated Model vs. the Logistic Regression Model for the Original Seed

Figure 7.19 shows the tradeoffs between sensitivity and specificity for the integrated model. The point at which the values are equivalent is approximately 72% vs. the 77% in Figure 7.14. The findings for the other randomly chosen test datasets are very similar.

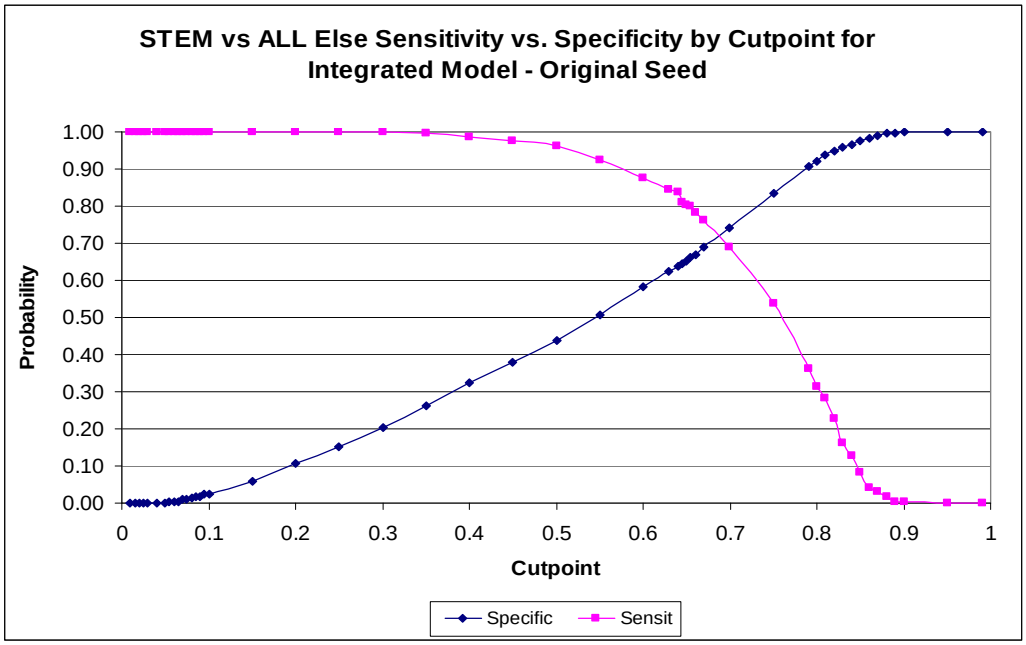

Figure 7.19 Sensitivity vs. Specificity by Cutpoint for the Integrated Model

To determine if the survival analysis module alone would perform differently than the integrated model, a log-logistic model was fitted from the original seed fit dataset without using the logistic regression module's estimated probability of a STEM outcome (LRprob_STEM) as a covariate. Essentially this model relied upon the 76 covariates from the NELS dataset, the survival times, and the censored vs. observed status of the survival times to create the model. The survival analysis model alone did not perform better. Figure 7.20 shows the resulting ROC Curve for the survival analysis model compared to that obtained for the logistic regression model when both were applied to the original seed test sample. The ROC curves in this graph are also plotted in a smoothed format just to show the comparison between them. The ROC curve for the

survival analysis model is flatter still than that depicted in Figure 7.18 for the integrated model indicating its accuracy was even weaker. A sensitivity of 70.9% is achievable with a corresponding (1 - Specificity) value of approximately 30.4%. This is less precise discrimination between the two outcomes than that offered by the integrated model or the logistic regression model.

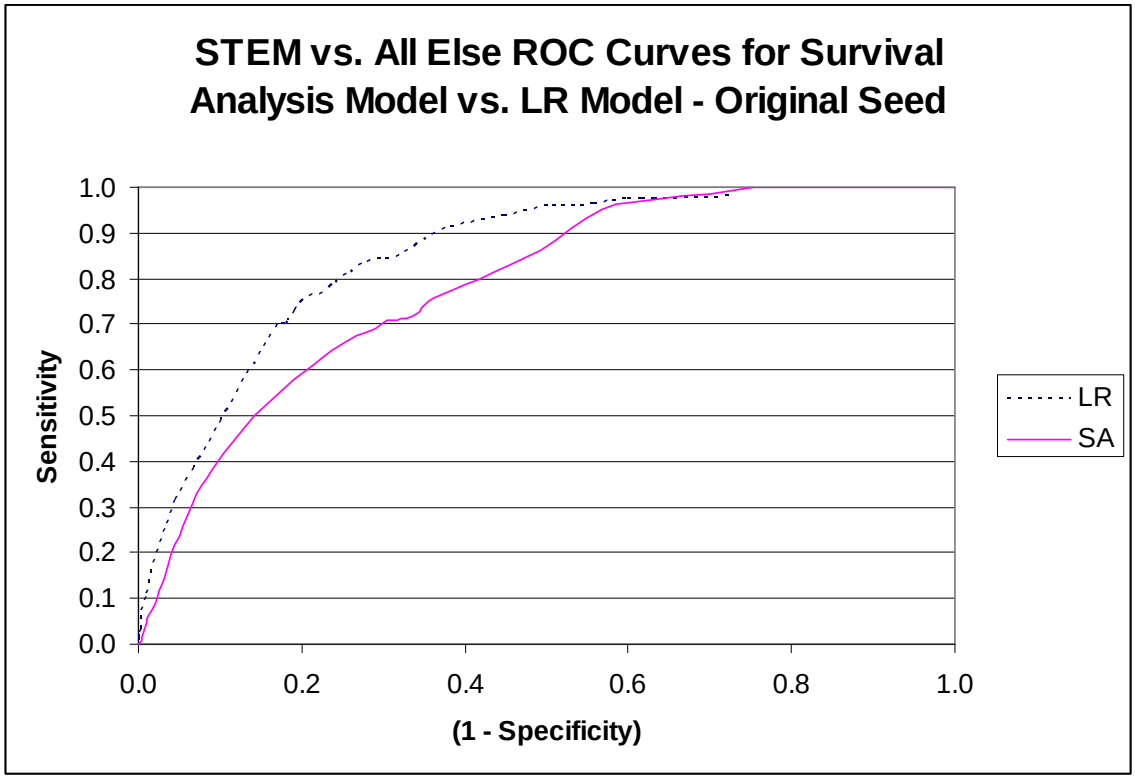

Figure 7.20 Sensitivity vs. (1-Specificity) for the Logistic Regression Model vs. the Survival Analysis Model without LR Module Input

Figure 7.21 shows the tradeoffs between sensitivity and specificity for the survival analysis model. The point at which the values are equivalent is approximately 69.5%. This finding suggests that the logistic regression module integrated with the survival analysis module produces better results than the survival analysis module alone for this test sample.

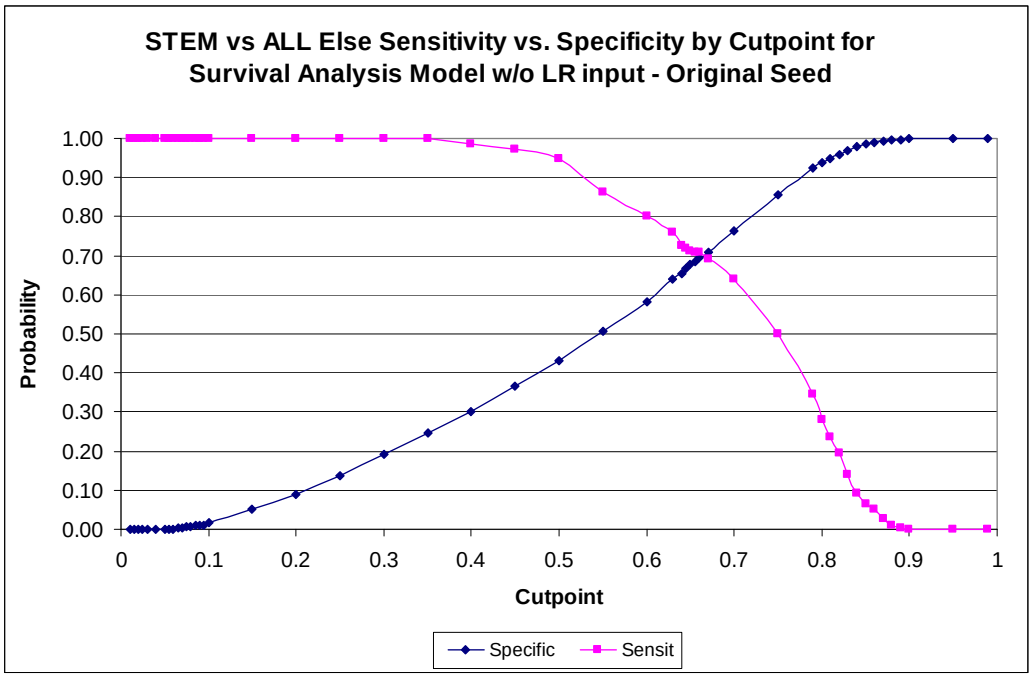

Figure 7.21 Sensitivity vs. Specificity by Cutpoint for the Survival Analysis Module

Table 7.18 lists the variables found to be significant predictors in the integrated model across the eleven different random number seeds used to create the fit and test sample datasets. Note that many of the same variables were also significant predictors in the logistic regression model. However, several variables which described aspects of the students' high schools were found to be significant predictors of survival on the STEM track beyond 7.25 years. The LRprob_STEM variable was found to be a significant predictor for each of the 11 samples in which it was a potential covariate. Since the integrated survival analysis module was fitting log-logistic probability models a scale parameter was also estimated and shown in the table. The table also shows the parameter estimates for the model fitted to the original seed sample without including the logistic regression model's estimated probability of a STEM degree (LRprob_STEM) as a covariate. This model was notably weaker because when applied to the same test data it provided lower values for sensitivity and specificity at the equivalence point

than either the logistic regression model or the integrated model with LRprob_STEM as a covariate.

Table 7.18 Integrated Model Parameters for STEM vs. All Else by Random Sample

Variable or Parameter	Description	Original Seed w/o Log. Reg.	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Scale	Scale parameter of log-logistic model	0.2216	0.2185	0.2207	0.2206	0.2202	0.2203	0.2220	0.2197	0.2206	0.2211	0.2199	0.2188
Intercept	Intercept parameter of log-logistic model	1.7878	1.8381	1.8072	1.7880	1.7702	1.6977	1.8005	1.8235	1.8689	1.7028	1.7985	1.7172
BY2XHQro	History/Cit/Geog Quartile	0.0201	0.0229	0.0232	0.0200		0.0143		0.0187	0.0203		0.0193	0.0224
BY2XMQro	Mathematics Quartile	0.0278	0.0256	0.0237	0.0266	0.0372	0.0276	0.0296	0.0182	0.0293	0.0367	0.0316	0.0281
BY2XSQro	Science Quartile						0.0119	0.0202	0.0138				
BYFCOMPr1	Family Composition: Mother & male guardian		-0.0381		-0.0347		-0.0418	-0.0405	-0.0404		-0.0510	-0.0405	
BYFCOMPr3	Family Composition: Mother	0.0804	0.0674	0.0531	0.0519	0.0618	0.0763		0.0514			0.0520	0.0495
BYHMLANGro	Home Language Background	-0.0269	-0.0244	-0.0246	-0.0331	-0.0302	-0.0286	-0.0209	-0.0310	-0.0273	-0.0281		-0.0395
BYHOMEWKro	# of Hours Spent on Homework per Week					0.0053							
BYLMrb	Language Minority Composite		-0.1875	-0.1671	-0.1932	-0.2297		-0.2299	-0.1617	-0.1757		-0.3408	
BYP64Bro	Family rule re how early/late child watches TV			-0.0219		-0.0352	-0.0357		-0.0264	-0.0354			
BYP64Dro	Family rule how many hrs on school days child watches TV							-0.0161					
BYP68ro	How often parent talks to child re post H.S. plans	-0.0221	-0.0173	-0.0150					-0.0171	-0.0142	-0.0171		-0.0208
BYP76ro	How far in school parent expects child to go	0.0158	0.0143	0.0149	0.0110	0.0124	0.0122	0.0131	0.0159	0.0148	0.0125	0.0109	0.0149
BYPARMARr5	Parents' Marital Status: Marriage-like relationship												-0.1033
BYRISKro	# of BY Risk Factors for Dropping Out of School	-0.0620	-0.0564	-0.0487	-0.0599	-0.0494	-0.0686	-0.0403	-0.0598	-0.0377	-0.0391	-0.0476	-0.0547
BYS14rPvNRel	HS Student Plans to Attend: Private Nonreligious						-0.0850			-0.0643	-0.0613	-0.0470	-0.0667
BYS14rPvRel	HS Student Plans to Attend: Private Religious	0.0469	0.0538	0.0507	0.0437			0.0481	0.0443		0.0400	0.0422	
BYS20rb	Language Student usually speaks now									-0.0611			
BYS32ro	Number of siblings student has	-0.0118	-0.0097		-0.0071			-0.0087	-0.0063				-0.0091
BYS34Arb	Father's Highest Level of Education: College or not	0.0433	0.0496	0.0445	0.0360	0.0579	0.0662	0.0419	0.0328	0.0341	0.0661	0.0481	0.0520
BYS41ro	Time spent after school with no adult present	0.0126	0.0134	0.0112	0.0166	0.0163	0.0120	0.0128	0.0193	0.0133	0.0107	0.0154	0.0118
BYS43ro	# of Cigarettes Student Smokes per Day							-0.0355					-0.0280

Table 7.18 (continued).

Variable or Parameter	Description	Original Seed w/o Log. Reg.	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
BYS45ro	How far in school do you think you will get	0.0510	0.0552	0.0493	0.0544	0.0549	0.0493	0.0554	0.0523	0.0549	0.0544	0.0561	0.0483
BYS46ro	How sure that you will graduate from H.S.	-0.0491	-0.0391	-0.0265	-0.0429	-0.0447	-0.0533	-0.0332	-0.0497	-0.0568	-0.0576	-0.0397	-0.0401
BYS47ro	How sure student is to go further than H.S.	-0.0478	-0.0360	-0.0358	-0.0298	-0.0345	-0.0371	-0.0263	-0.0260	-0.0341	-0.0357	-0.0318	-0.0380
BYS53ro	# of Hrs student works for pay per week		-0.0120	-0.0146	-0.0148	-0.0136	-0.0207	-0.0140	-0.0124	-0.0139		-0.0181	-0.0161
BYS60Aro	Student's ability group for Mathematics		0.0179	0.0212	0.0151	0.0169		0.0169	0.0231	0.0151	0.0142		0.0211
BYS60Bro	Student's ability group for Science		-0.0151	-0.0172	-0.0128	-0.0201	-0.0099	-0.0195	-0.0222	-0.0173	-0.0127	-0.0103	-0.0154
BYS81Aro	English grades from Grade 6 until now	0.0200	0.0186	0.0199	0.0283	0.0259	0.0192	0.0237	0.0171	0.0254	0.0189	0.0241	0.0191
BYS81Bro	Math grades from Grade 6 until now						0.0102						
BYS81Cro	Science grades from Grade 6 until now		0.0167	0.0167	0.0141	0.0297	0.0133	0.0208	0.0237	0.0140	0.0149	0.0214	0.0147
BYSC13Erc	% of White Non-Hispanic 8th Graders	-0.0006	-0.0005	-0.0005			-0.0006	-0.0007	-0.0005	-0.0006	-0.0007		
BYSC16Arc	# of students in Free Lunch Program	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001		-0.0001	-0.0001	-0.0001	-0.0001		-0.0001
BYSC16Brc	# of students in Remedial Reading						-0.0001						
BYSC16Drc	# of students in Bilingual Education					0.0002							
BYSC16Erc	# of students in English as 2nd Language	0.0004	0.0004	0.0005	0.0005	0.0006	0.0005	0.0004	0.0006	0.0005	0.0006	0.0005	0.0006
BYSC16Grc	# of students in Gifted, Talented Ed	0.0001	0.0001					0.0001					
BYSC29rb	Min. GPA Required to Participate in Activities											-0.0200	
F2RACTEro	ACT (English Score)		-0.0077	-0.0063	-0.0050	-0.0105	-0.0126	-0.0073	-0.0108	-0.0084	-0.0072	-0.0106	-0.0087
F2RACTMro	ACT (Math)		0.0134	0.0125	0.0108	0.0109	0.0116	0.0138	0.0106	0.0152	0.0064	0.0086	0.0088
F2RACTSro	ACT (Science Reasoning)					0.0063	0.0072		0.0070		0.0065	0.0071	0.0060
F4RACE2rAI	Race of student: Amer Ind						-0.0263						
F4RACE2rAs	Race of student: Asian		0.0425	0.0602	0.0596	0.0430	0.0420	0.0562	0.0699	0.0685		0.0837	0.0415
F4RACE2rBl	Race of student: African-American		0.0444	0.0441	0.0661	0.0534			0.0473	0.0430	0.0441	0.0397	0.0726
F4SEXrb	Sex of student - binary (1 = Female)		-0.0258	-0.0277	-0.0363	-0.0389		-0.0294	-0.0227	-0.0353		-0.0433	-0.0387

Table 7.18 (continued).

Variable or Parameter	Description	Original Seed w/o Log. Reg.	Original Seed	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
SATQUANro	Scholastic Aptitude Test (Mathematics) expanded	0.0001	0.0004	0.0002	0.0005	0.0005	0.0003	0.0006	0.0006	0.0003	0.0002	0.0005	0.0004
SATVERro	Scholastic Aptitude Test (Verbal) expanded		-0.0002		-0.0003	-0.0002		-0.0003	-0.0003			-0.0003	-0.0002
LRprob_STEM	Logistic Regression model estimated Prob(STEM)	N/A	-0.5747	-0.4192	-0.4877	-0.6549	-0.5904	-0.6168	-0.7163	-0.5805	-0.3961	-0.5516	-0.5798

7.3.2 Integrated Model in Parallel Results

Across the eleven random samples, the model achieved results ranging from 73.2 to 80.4% in sensitivity and 75.5 to 77.6% specificity. The sensitivity and specificity for each of the samples were optimized at a level that represented good predictive discrimination between the STEM and All Else students. The ROC curves for the integrated model in series indicate that the choice of cutpoint(s) strongly affects the models' predictive strength and ability to focus on the sub-population of interest.

7.4 VALIDITY OF THE STEM-RELATED CATEGORY

The logistic regression models developed for STEM vs. STEM-Related, STEM vs. Non-STEM, STEM vs. Other 4 Year Degree, and STEM-Related vs. Non-STEM indicated the hierarchical nature of the models. Models with more divergent two-outcome pairs were stronger than those with in which the two outcomes were more similar. Table 7.19 contains the ROC Curve AUC results of the fitted models for various two-outcome pairs. The models' strength tended to increase as the disparity between the outcomes increased.

Table 7.19 Hierarchy of Logistic Regression Model Accuracy by Outcome Pair

Outcome	STEM	STEM-Rel	Non-STEM	Sub 4Yr Deg	No Degree
STEM	N/A	0.720	0.743	0.924	0.919
STEM-Rel	-	N/A	0.550	0.885	0.887
Non-STEM	-	-	N/A	0.876	0.878
Sub 4Yr Deg	-	-	-	N/A	0.604
No Degree	-	-	-	-	N/A

These results are also found when the outcomes include combinations of categories. For example, the STEM vs. Other Degree model had an AUC of 0.742 which is comparable to a blending of the individual STEM vs. STEM-Related and STEM vs. Non-STEM AUC figures. The STEM vs. All Else model had an AUC of 0.848 which lies between the STEM vs. STEM-Related or Non-STEM figures and those of STEM vs. Sub-4 Year Degree or No Degree. The Degree vs. Non-Degree model predicted student outcomes to be either a 4 Year Degree or a Sub-4 Year Degree/No Degree, and its associated AUC value was 0.882. This reflects the clear divergence between students that did and did not earn a bachelor's degree. Thus having an AUC value larger than that of the STEM vs. All Else model is not surprising. The All Else category includes a diverse population of students including those who earned other 4 year college degrees and have more in common with the STEM students than the No Degree students.

One important issue was to determine the validity of creating the STEM-Related category. If this category represented a valid subdivision of the students with bachelor's degrees, significant differences were expected between this category and those of the STEM and Non-STEM categories. Evaluation of the STEM & STEM-Related vs. Non-STEM model indicates that it has slightly better predictive accuracy than the STEM-Related vs. Non-STEM model and less accuracy than the STEM vs. Other Degree model. This suggests that while there may be benefit to keeping STEM-Related as a separate category, the students within this category have more in common with the Non-STEM students than the STEM students. Another logistic regression model was fitted to predict between a combination of STEM and STEM-Related students and their Non-STEM counterparts. This STEM & STEM-Related vs. Non-STEM model had an AUC value of 0.613 which suggests that the STEM-Related students were more similar to the Non-STEM students than the STEM students.

Figure 7.22 illustrates the ROC curves for the STEM vs. STEM-Related, STEM vs. Other Degree, STEM-Related vs. Non-STEM, and STEM & STEM-Related vs. Non-STEM models when applied to the test data sets. The STEM vs. STEM-Related and STEM vs. Other Degree models were very similar with acceptable predictive accuracy. The STEM-Related vs. Non-STEM model was similar to a 45º line demonstrating the model had no real ability to discriminate between these outcomes. The ROC curve for the STEM & STEM-Related vs. Non-STEM model lies in the middle of the other curves showing it had poor discrimination performance. These results indicate that greater predictive accuracy is achieve by keeping STEM as a narrowly defined category rather than expanding it to include the majors comprising the STEM-Related category.

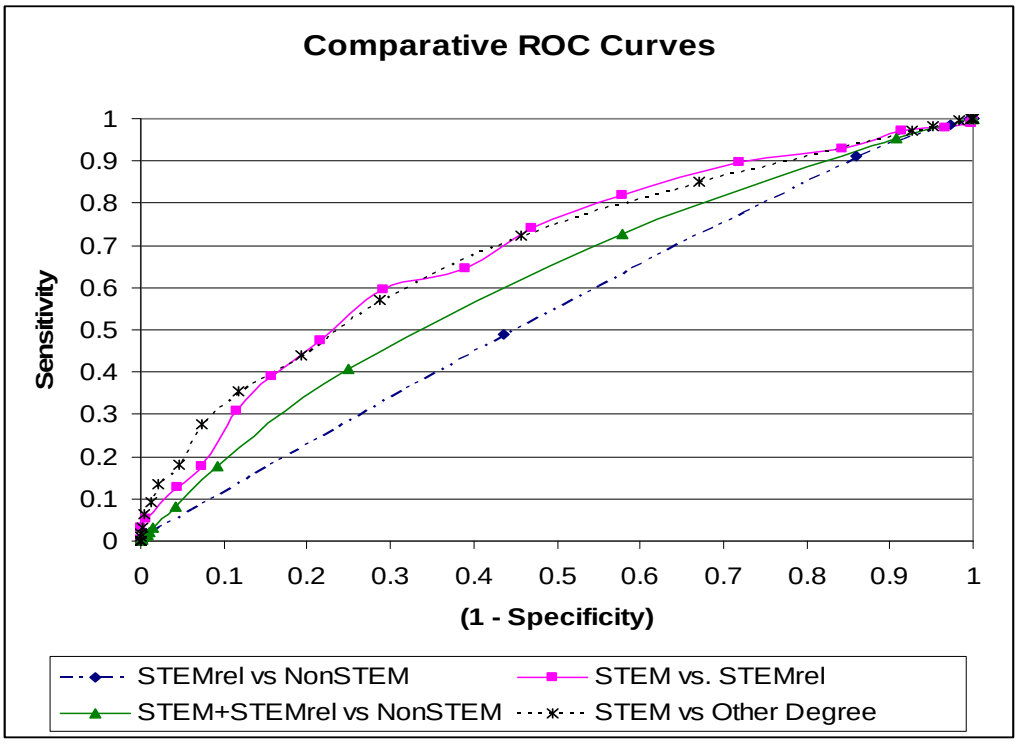

Figure 7.22 Comparison of ROC Curves for Modeling Different College Degree Outcomes

The dataset was then examined to determine how many of the 736 STEM students also earned a STEM-Related degree. A total of 20 STEM students also earned a college degree in a STEM-Related topic with 17 completing the STEM degree first or simultaneously. The remaining 3 STEM students earned a STEM-Related degree prior to the STEM degree. Of these 20 students, 12 of the STEM-Related degrees were of a professional nature with 1 in Health, 2 in Dentistry, 5 in Medicine, and 4 in Psychology. Since the number of STEM students that also had a STEM-Related degree was comparatively quite small, this was not judged to be a significant factor in explaining the models' predictive accuracy.

7.5 SUMMARY

The classification of the records by educational outcome permitted many two-outcome pairs including combinations of outcomes to be modeled. The stability of the models fitted by the multiple random samples was very good. Logistic regression models generally have greater predictive ability when the two outcomes are sharply different from one another. The predictive accuracy of the logistic regression models created in this research varied from negligible for the STEM-Related vs. Non-STEM model to outstanding for the STEM vs. Sub-4 Year Degree and STEM vs. No Degree models. The results for this application indicated a hierarchical relationship between the outcomes with STEM students exhibiting significant differences from the other students. The patterns of predictive accuracy for the models suggested that the five basic educational outcomes outlined in this research may be considered an ordered set as follows:

{STEM Degree, STEM-Related Degree, Non-STEM Degree, Sub-4 Year Degree, No Degree }

Considering this set to have an ordinal scale, the modeling accuracy improved as the two potential outcomes differed. For example, the models which predicted the probability of a STEM outcome improved as the alternative outcome modeled changed from a closely related category like STEM-Related to a more divergent category like Sub-4 Year Degree. The STEM-Related vs. No-Degree model had excellent predictive accuracy while the STEM-Related vs. Non-STEM model had little more predictive accuracy than tossing a fair coin. The STEM & STEM-Related vs. Non-STEM model was better than tossing a coin, but there was poor discrimination between these two outcomes. Table 7.20 lists the logistic regression models and the level of predictive accuracy associated with each fitted model.

Table 7.20 Comparison of Logistic Regression Model Accuracy

Model	Predictive Accuracy	Accuracy Scale
STEM vs. STEM-Related	Acceptable	$0.70 \leq AUC < 0.80$
STEM vs. Non-STEM	Acceptable	$0.70 \leq AUC < 0.80$
STEM vs. Sub-4 Year Degree	Outstanding	$AUC \geq 0.90$
STEM vs. No-Degree	Outstanding	$AUC \geq 0.90$
STEM vs. All Else	Excellent	$0.80 \leq AUC < 0.90$
STEM vs. Other 4 Year Degree	Acceptable	$0.70 \leq AUC < 0.80$
STEM & STEM-Related vs. Non-STEM	Poor	$0.50 < AUC < 0.60$
STEM-Related vs. Non-STEM	Negligible	$AUC \approx 0.50$
STEM-Related vs. Sub-4 Year Degree	Excellent	$0.80 \leq AUC < 0.90$
STEM-Related vs. No-Degree	Excellent	$0.80 \leq AUC < 0.90$
4 Year Degree vs. Non-4 Year Degree	Excellent	$0.80 \leq AUC < 0.90$

The ROC Curves permitted the models to be "tuned" by changing the prediction cutpoints to ensure a desired level of sensitivity in exchange for a corresponding level of specificity. The sensitivity analysis indicated that very high levels of correct STEM predictions could be achieved if the policy goal was to identify the maximum number of STEM students and students who could earn STEM degree if they chose to remain on the STEM track. If an intervention program were being considered that had a fixed cost per student, the prediction

cutpoint could be chosen to maximize the sensitivity while not exceeding the budgetary limits. One of the most powerful benefits of using the sensitivity analysis module is that decision makers are free to adjust the integrated model to optimize their policy goals.

8.0 EFFECTIVENESS OF THE VARIOUS MODELS

8.1 INTRODUCTION

The key aspect in validating the integrated model was using the same samples to fit and test the separate logistic regression model, the integrated model linked in series, and the integrated model linked in parallel. This allowed direct comparison of the actual outcomes with the predicted outcomes of the different models to determine the accuracy of each model. It also allowed direct comparison of accuracy between the different models.

If the results for a particular sample are put in the form of a 2 x 2 classification table they can be quickly summarized in terms of the true and false predictions of STEM vs. All Else as shown in Table 8.1. The number of True STEM, False STEM, True All Else, and False All Else predictions by sample will be compared for the different models tested.

Table 8.1 Example of a Results Classification Table

		Actual Outcome	
		STEM	All Else
Predicted Outcome	STEM	True STEM	False STEM
	All Else	False All Else	True All Else

8.2 COMPARISON OF LOGISTIC REGRESSION MODEL PREDICTIONS TO ACTUAL RESULTS

The results of the logistic regression analysis were very good. The STEM vs. All Else model applied to the test data for the various seeds correctly identified the majority of the 220 STEM students. Of the 3,117 All Else students, the number incorrectly predicted to be STEM varied widely. The levels of sensitivity and specificity at the cutpoint value producing equivalence ranged from 73.6 to 78.6% and 75.1 to 78.0%, respectively. These levels of accuracy were much better than initially anticipated. Table 8.2 shows the comparison across the samples of the accuracy achieved at the equivalence cutpoints. Between 162 and 173 of the STEM students were correctly predicted while 686 to 780 All Else students were incorrectly predicted to be STEM. One interpretation of this finding is that based on their attributes these All Else students could have been successful in a STEM major had they chosen to pursue one. There could certainly be additional factors and influences that were not included in the model. For example, a student may be influenced by a trusted mentor or role model that encourages study in a different domain. This is an interesting question for future research since not all of the students' attributes are known, and it would require collecting data about why the students chose a major other than STEM.

Table 8.2 STEM vs. All Else Logistic Regression Model Accuracy by Random Sample for the Cutpoint Producing Equivalent Sensitivity and Specificity Values

Item	Seed Orig.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Cutpoint	0.075	0.065	0.07	0.07	0.075	0.07	0.07	0.075	0.065	0.07	0.07
True STEM	170	168	162	164	167	169	173	170	164	169	169
True All Else	2415	2337	2343	2357	2403	2391	2386	2431	2341	2387	2376
False STEM	702	780	774	760	714	726	731	686	776	730	741
False All Else	50	52	58	56	53	51	47	50	56	51	51
Specificity	77.5%	75.0%	75.2%	75.6%	77.1%	76.7%	76.5%	78.0%	75.1%	76.6%	76.2%
(1 - Specificity)	22.5%	25.0%	24.8%	24.4%	22.9%	23.3%	23.5%	22.0%	24.9%	23.4%	23.8%
Sensitivity	77.3%	76.4%	73.6%	74.5%	75.9%	76.8%	78.6%	77.3%	74.5%	76.8%	76.8%

Table 8.3 shows the same comparison across the samples of the accuracy achieved at the cutpoint producing a sensitivity level of approximately 80.0%. This cutpoint produced similar results with between 174 and 180 correctly predicted STEM students and 733 to 1,028 All Else student incorrectly predicted to be STEM.

Table 8.3 STEM vs. All Else Logistic Regression Model Accuracy by Random Sample for the Cutpoint Producing Approximately 80% Sensitivity

Item	Seed Orig.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Cutpoint	0.07	0.055	0.05	0.055	0.065	0.06	0.06	0.07	0.055	0.06	0.065
True STEM	174	178	180	178	178	177	176	178	176	175	175
True All Else	2383	2215	2089	2179	2294	2304	2282	2384	2233	2281	2321
False STEM	734	902	1028	938	823	813	835	733	884	836	796
False All Else	46	42	40	42	42	43	44	42	44	45	45
Specificity	76.5%	71.1%	67.0%	69.9%	73.6%	73.9%	73.2%	76.5%	71.6%	73.2%	74.5%
(1 - Specificity)	23.5%	28.9%	33.0%	30.1%	26.4%	26.1%	26.8%	23.5%	28.4%	26.8%	25.5%
Sensitivity	79.1%	80.9%	81.8%	80.9%	80.9%	80.5%	80.0%	80.9%	80.0%	79.5%	79.5%

One aspect of evaluating the model's accuracy was examining the instances where the model incorrectly predicted an All Else student to have a STEM outcome (a False Positive). Table 8.4 illustrates the breakdown of the 734 false positives for the STEM vs. All Else model using the original seed at a 0.07 cutpoint threshold as shown in Table 8.3. Table 8.4 categorizes the false positives by the actual outcome in terms of STEM track departure type. The results for the other seeds were very consistent across the other ten random samples.

Table 8.4 False Positive Breakdown by STEM Track Departure Type for the Original Seed and Cutpoint = 0.07

STEM Track Departure Type	# of False Positives	# of Dep. Type in Sample	% of Other Than STEM	% Type of False Pos
H.S. Dropout	1	167	5.36%	0.14%
H.S. Graduate	37	720	23.10%	5.04%
College Dropout	108	629	20.18%	14.71%
Incomplete Deg.	36	141	4.52%	4.90%
Sub-4 Yr Degree	69	525	16.84%	9.40%
Other 4 Yr Degree	483	935	30.00%	65.80%
Total	734	3,117	100.00%	100.00%

Most of the incorrect STEM predictions were for students that ultimately went on to get a different four year degree. The logistic regression model did an excellent job of predicting an All Else outcome for the students that did not finish a college degree or achieved a Sub 4 Yr degree. The model was a little less adept at identifying the college dropouts as All Else. As previously discussed, this suggests that the model may have detected aspects of these students that could have allowed them to complete a STEM degree had they chosen to attempt and persist in that field of study.

8.3 COMPARISON OF INTEGRATED IN SERIES MODEL PREDICTIONS TO ACTUAL RESULTS

The integrated model with logistic regression and survival analysis linked in series also produced good results. When applied to the test data this model correctly identified between 151 and 179 of the 220 STEM students depending on which cutpoint was used for the logistic regression module. The levels of sensitivity and specificity at the cutpoint value producing equivalence ranged from 68.6 to 77.3% and 70.0 to 74.2%, respectively. Table 8.5 shows the comparison across the samples of the accuracy achieved at the equivalence cutpoints. Between 162 and 173 of the STEM students were correctly predicted while 686 to 780 All Else students were incorrectly predicted to be STEM.

Table 8.5 STEM vs. All Else Integrated Model Accuracy by Random Sample for the Cutpoint Producing Equivalent Sensitivity and Specificity Values

Item	Seed Orig.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Cutpoint	0.7	0.66	0.7	0.7	0.7	0.7	0.7	0.7	0.67	0.67	0.7
True STEM	152	156	158	157	157	156	160	167	156	170	151
True All Else	2310	2181	2313	2288	2272	2275	2246	2271	2221	2251	2364
False STEM	807	936	804	829	845	842	871	846	896	866	753
False All Else	68	64	62	63	63	64	60	53	64	50	69
Specificity	74.1%	70.0%	74.2%	73.4%	72.9%	73.0%	72.1%	72.9%	71.3%	72.2%	75.8%
(1 - Specificity)	25.9%	30.0%	25.8%	26.6%	27.1%	27.0%	27.9%	27.1%	28.7%	27.8%	24.2%
Sensitivity	69.1%	70.9%	71.8%	71.4%	71.4%	70.9%	72.7%	75.9%	70.9%	77.3%	68.6%

Table 8.6 shows the same comparison across the samples of the accuracy achieved at the cutpoint producing a sensitivity level of approximately 80.0% for the logistic regression module. This cutpoint produced similar results with between 174 and 179 correctly predicted STEM students and 921 to 1,113 All Else students incorrectly predicted to be STEM.

Table 8.6 STEM vs. All Else Integrated Model Accuracy by Random Sample for the Cutpoint Producing Approximately 80% Sensitivity

Item	Seed Orig.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
Cutpoint	0.655	0.63	0.655	0.65	0.645	0.645	0.655	0.67	0.64	0.66	0.66
True STEM	176	176	175	177	175	174	176	179	175	176	177
True All Else	2063	2039	2055	2029	2004	2030	2014	2116	2075	2196	2147
False STEM	1054	1078	1062	1088	1113	1087	1103	1001	1042	921	970
False All Else	44	44	45	43	45	46	44	41	45	44	43
Specificity	66.2%	65.4%	65.9%	65.1%	64.3%	65.1%	64.6%	67.9%	66.6%	70.5%	68.9%
(1 - Specificity)	33.8%	34.6%	34.1%	34.9%	35.7%	34.9%	35.4%	32.1%	33.4%	29.5%	31.1%
Sensitivity	80.0%	80.0%	79.5%	80.5%	79.5%	79.1%	80.0%	81.4%	79.5%	80.0%	80.5%

If each prediction category (True STEM, True All Else, False STEM, False All Else) for each sample for the logistic regression model is compared to its counterpart integrated model they can be tested for statistically significant differences. Treating the logistic regression predictions as the "Expected" figures and the integrated model predictions as the "Observed" figures, a chi-square statistic can be calculated as $\chi^2 = \sum_{i=1}^{4} \frac{[Obs_i - Exp_i]^2}{Exp_i}$ and compared to a critical value of χ^2 at the alpha (α) = 0.05 level with 1 degree of freedom = 3.843. Performing this analysis for each of the sample/model combinations, there are significant differences between the logistic regression and integrated model for each sample except the two samples for random seed 2.

Table 8.7 compares the true and false STEM predictions across the 11 random samples for the models producing 80% sensitivity. Testing the variances between the two models finds that the *p*-value for an *F* test of equal variances is 0.3875 for the True STEM predictions and 0.2553 for the False STEM predictions. This suggests that we cannot reject the hypothesis of equal variances for the True STEM predictions or the False STEM predictions. Assuming a

normal distribution for these figures, a Student's *t*-test of the True STEM predictions has an associated *p*-value of 0.2374 suggesting there is no significant difference between the means of these predictions. Using Welch's[99] *t*-test for small samples with unequal variances for the False STEM predictions results in an associated *p*-value of < 0.0001 suggesting there is evidence to conclude the models' mean predictions are significantly different. Overall, a visual inspection of the sample/model combinations reveals that the logistic regression model performs better at providing fewer False STEM predictions. So the model integrating logistic regression and survival analysis in series does not provide improved accuracy.

Table 8.7 True vs. False STEM Predictions for the Logistic Regression and Integrated Models by Random Sample at 80% Sensitivity

Seed	True STEM Predictions		False STEM Predictions	
	Logistic Regression	Integrated (Series)	Logistic Regression	Integrated (Series)
Original	174	176	734	1054
1	178	176	902	1078
2	180	175	1028	1062
3	178	177	938	1088
4	178	175	823	1113
5	177	174	813	1087
6	176	176	835	1103
7	178	179	733	1001
8	176	175	884	1042
9	175	176	836	921
10	175	177	796	970
average	176.82	176.00	847.45	1047.18
variance	3.16	1.80	7579.67	3595.36

8.4 COMPARISON OF INTEGRATED IN PARALLEL MODEL PREDICTIONS TO ACTUAL RESULTS

The integrated model with logistic regression and survival analysis linked in series and then combined in parallel also produced good results. This layered model predicted a STEM outcome if the logistic regression probability of STEM was ≥ 0.07 and the integrated model probability of survival beyond 7.25 years on the STEM track was ≥ 0.50. When applied to the test data this model correctly identified between 158 and 177 of the 220 STEM students. The number of All Else students incorrectly predicted to be STEM ranged from 698 to 765 out of a total of 3,117 All Else students. The levels of sensitivity and specificity ranged from 71.8 to 80.4% and 75.5 to 77.6%, respectively. Table 8.8 shows the accuracy results by sample.

Table 8.8 STEM vs. All Else Integrated in Parallel Model Accuracy by Random Sample

Item	Seed Orig.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Seed 7	Seed 8	Seed 9	Seed 10
True STEM	173	163	161	164	173	168	172	177	158	167	168
True All Else	2400	2411	2352	2373	2355	2392	2391	2398	2404	2398	2390
False STEM	717	706	765	744	762	725	726	719	713	719	727
False All Else	47	57	59	56	47	52	48	43	62	53	52
Specificity	77.0%	77.4%	75.5%	76.1%	75.6%	76.7%	76.7%	76.9%	77.1%	76.9%	76.7%
(1 - Specificity)	23.0%	22.6%	24.5%	23.9%	24.4%	23.3%	23.3%	23.1%	22.9%	23.1%	23.3%
Sensitivity	78.6%	74.1%	73.2%	74.5%	78.6%	76.4%	78.2%	80.5%	71.8%	75.9%	76.4%

Setting the critical logistic regression probability cutpoint to 0.07 meant that module of the integrated model was set to the point that optimized both sensitivity and specificity. As a

result this model is most directly comparable to the logistic regression model at the cutpoint that produces equivalent sensitivity and specificity values.

Table 8.9 illustrates the breakdown of the 717 false positives for the STEM vs. All Else model using the original seed as shown in Table 8.8 and categorizes the false positives by the actual outcome in terms of STEM track departure type. The results for the other seeds were very consistent across the other ten random samples.

Table 8.9 Integrated Model False Positive Breakdown by STEM Track Departure Type for the Original Seed

STEM Track Departure Type	# of False Positives	# of Dep. Type in Sample	% of Other Than STEM	% Type of False Pos
H.S. Dropout	0	167	5.36%	0.00%
H.S. Graduate	35	720	23.10%	4.88%
College Dropout	103	629	20.18%	14.37%
Incomplete Deg.	35	141	4.52%	4.88%
Sub-4 Yr Degree	64	525	16.84%	8.93%
Other 4 Yr Degree	480	935	30.00%	66.95%
Total	717	3117	100.00%	100.00%

As with the logistic regression model, most of the incorrect STEM predictions were for students that ultimately went on to get a different four year degree. The integrated model did an excellent job of predicting an All Else outcome for the students that did not finish a college degree or achieved a Sub 4 Yr degree. The model was a little less able to identify the college dropouts as All Else as with the model integrated in series. This suggests that the model may have detected aspects of these students that could have allowed them to complete a STEM degree had they chosen to attempt and persist in that field of study.

The chi-square test comparing the observed numbers of True STEM, False STEM, True All Else, and False All Else predictions to those produced by the logistic regression model found

that only 3 of the seeds exhibited a significant difference. The results were then examined by comparing the numbers of true and false STEM numbers across all the seeds as shown in Table 8.10.

Table 8.10 True vs. False STEM Predictions for the Logistic Regression and Integrated in Parallel Models by Random Sample

Seed	True STEM Predictions		False STEM Predictions	
	Logistic Regression	Integrated (Parallel)	Logistic Regression	Integrated (Parallel)
Original	170	173	702	717
1	168	163	780	706
2	162	161	774	765
3	164	164	760	744
4	167	173	714	762
5	169	168	726	725
6	173	172	731	726
7	170	177	686	719
8	164	158	776	713
9	169	167	730	719
10	169	168	741	727
average	167.73	167.64	738.18	729.36
variance	10.42	33.65	982.96	376.65

Testing the variances between the two models finds that the *p*-value for an *F* test of equal variances is 0.0782 for the True STEM predictions and 0.065 for the False STEM predictions. This suggests that we cannot reject the hypothesis of equal variances for the True STEM predictions at an alpha (α) level of 0.05 but it could be rejected at an alpha level of 0.10. Assuming a normal distribution for these figures, a Student's *t*-test of the True STEM predictions has an associated *p*-value of 0.9643 suggesting there is no significant difference between the means of these predictions. The Student's *t*-test of the False STEM predictions has an associated *p*-value of 0.4370 suggesting there is also no significant difference between the means of these

predictions. Therefore the two models appear to be performing at the same level of accuracy, and the integrated in parallel model does not offer significantly more accurate predictions.

The results for the variance tests suggest that the two models may not be performing with identical variances. For this set of 11 samples the integrated in parallel model predicts True STEM numbers that vary more than those of the logistic regression model by a factor of 3.23. The integrated in parallel model predicts False STEM numbers with lower variation than that of the logistic regression model by a factor of 2.61. So the model integrating logistic regression and survival analysis in parallel provides very comparable accuracy in terms of average predictions of True and False STEM numbers, but the variation of the integrated model for the False STEM predictions is better. However, since an alpha level of 0.05 has been consistently applied throughout this analysis, the sample results are not sufficient to conclude that the two models perform with significantly different variances.

Another way of examining the results is to compare the models' performance for each seed on the basis of percent correct/incorrectly. Table 8.11 lists the performance of the logistic regression models in terms of the percent of the STEM [All Else] students correctly predicted to have a STEM [All Else] outcome or incorrectly predicted to have an All Else [STEM] outcome. The results show that on average the logistic regression models had correct predictions for 76.2% of all the students.

Table 8.11 Percentage of True vs. False STEM/All Else Predictions for the Logistic Regression Models by Random Sample

Seed	% True STEM	%False All Else	%True All Else	%False STEM
Original	0.773	0.227	0.775	0.225
1	0.764	0.236	0.750	0.250
2	0.736	0.264	0.752	0.248
3	0.745	0.255	0.756	0.244
4	0.759	0.241	0.771	0.229
5	0.768	0.232	0.767	0.233
6	0.786	0.214	0.765	0.235
7	0.773	0.227	0.780	0.220
8	0.745	0.255	0.751	0.249
9	0.768	0.232	0.766	0.234
10	0.768	0.232	0.762	0.238
Average	0.762	0.238	0.763	0.237

Table 8.12 lists the performance of the Integrated in Parallel models in terms of the percent of the STEM [All Else] students correctly predicted to have a STEM [All Else] outcome or incorrectly predicted to have an All Else [STEM] outcome. The results show that on average these integrated models also had correct predictions for 76.2% the STEM students. However, this set of models had a slightly better average percentage of correct predictions for the All Else students at 76.6% vs. 76.3% for the logistic regression models.

Table 8.12 Percentage of True vs. False STEM/All Else Predictions for the Integrated in Parallel Models by Random Sample

Seed	% True STEM	%False All Else	%True All Else	%False STEM
Original	0.786	0.214	0.770	0.230
1	0.741	0.259	0.774	0.226
2	0.732	0.268	0.755	0.245
3	0.745	0.255	0.761	0.239
4	0.786	0.214	0.756	0.244
5	0.764	0.236	0.767	0.233
6	0.782	0.218	0.767	0.233
7	0.805	0.195	0.769	0.231
8	0.718	0.282	0.771	0.229
9	0.759	0.241	0.769	0.231
10	0.764	0.236	0.767	0.233
Average	0.762	0.238	0.766	0.234

These tables further illustrate that for 7 of the 11 random seeds, the Integrated in Parallel models performed slightly better in correctly predicting All Else outcomes.

8.5 FINDINGS

The logistic regression models were developed to provide a basis for comparing the accuracy of the integrated methodology models. As the standard statistical technique for modeling a binary result, logistic regression was the logical method to choose for the initial analysis. It was anticipated that the logistic regression models would be of acceptable predictive accuracy in most cases, but not necessarily very strong. The logistic regression models turned out to be much more accurate than expected and extremely beneficial in analyzing the dataset. This in part was due to the effort in preparing the dataset in advance to ensure the variables were recoded to have binary, ordinal, or continuous scales.

Since the results for some of the logistic regression models for the STEM vs. All Else samples were so strong, there was less opportunity for the integrated methodology to provide a significant improvement. Across each of the random samples tested, the model integrated in series failed to show an improvement. The integrated models linked in series or parallel to the logistic regression model did an excellent job of distinguishing between the STEM students and the students that did not earn a bachelor's degree. Most of the errors in prediction concerned falsely predicting other four year degree students to have a STEM outcome. However, the model that was fitted to test the predictive strength of the survival analysis module applied to this data found that it performed poorly as a single technique in this instance. Thus the integration was able to improve upon the results of survival analysis alone, but it was not able to improve upon the results of logistic regression alone.

The integrated in parallel model did not consistently produce an improvement in the identification of True STEM students. While this model provided an improvement in that the numbers of False STEM students was reduced for 8 of the 11 random samples and the variability of these predictions was lower these results were not statistically significant. Overall, the integrated model did not perform in a significantly different manner from the logistic regression model in this application.

9.0 CONCLUSIONS AND RECOMMENDATIONS

9.1 CONCLUSIONS

9.1.1 Use of Integrated Models

The findings illustrated that models integrating multiple statistical techniques can be developed and applied to complex problems. This conclusion relates to the first research question outlined in Section 5.4: "Can multiple statistical techniques such as logistic regression analysis, ROC curve analysis, and survival analysis be successfully integrated and applied to a complex problem such as the achievement of a STEM degree?"

The integrated model approach is feasible and can achieve accuracy comparable to and, perhaps better than a standard single technique such as logistic regression. In this particular application, the improved results that were hypothesized did not occur because the logistic regression models were so strong. The model that integrated the survival analysis and logistic regression techniques in a combination of series and parallel exhibited some signs of refining the predictions to offer lower variability in the False STEM predictions but not enough to offer a statistically significant improvement. Therefore the original research hypothesis that the integrated model would produce more accurate answers than a single standard technique such as logistic regression was not proven in this case. However, the results suggest that the integrated

model methodology is feasible and could offer improvement in other applications for which a single standard technique might not be as accurate.

In retrospect, survival analysis, in itself, is a very powerful tool and has the potential to build upon the logistic regression's findings to improve accuracy. One aspect that may explain the particular result in this application is that the survival times of the 736 STEM students and the 3,156 students who earned bachelors degrees in other subjects were very similar. The similarity of the survival times hampered the survival analysis module portion of the integrated model by making it harder to discriminate between them. The cumulative hazard functions depicted for the four year degree students in Figure 6.1 were closely related with an intersection at approximately year 7. Defining the departure times for college graduates as the graduation date meant that most of the students that completed bachelor's degrees in any subject earned them in the year 1997. The mean departure time for the STEM students was 1997.42 compared to 1997.00 for the Other Degree students and 1997.04 for all four year degree students. The standard deviation of the departure times was 4.36 years for the STEM students, 1.04 years for the Other Degree students, and 4.89 for all four year degree students. The greater standard deviation for the STEM students may be explained by the cases where students earned a different degree before continuing on to complete a STEM degree as well as students taking longer than 4 years to complete a more demanding technical degree. The similar mean values for the departure times of the bachelor degree holders made it more difficult for the integrated model to distinguish between the groups despite treating the departure times for the STEM students as censored rather than observed events.

One approach that might improve the power of the survival analysis module would be changing the STEM track departure time for students that earned other college degrees. If the

time at which the students declared a major other than STEM were defined as the departure time, there would be a greater disparity in the departure times between STEM and Other Degree students. In addition, this approach would better reflect the phenomena of students starting a STEM major and later switching to a different major. In future research, it would be worthwhile to explore the feasibility of collecting this data. In the absence of information showing when a student declared a Not-STEM major or switched from a STEM major, the date the Other Degree students began attending college could be tested as a revised interpretation of the departure time.

The integrated model might offer an improved level of accuracy if the outcomes being predicted had more dissimilar hazard functions. For example, the cumulative hazard functions shown in Figure 6.1 reveal a period between years 6 and 7 when most of the students who ultimately earned a bachelor's degree had not yet begun to graduate, but those who never completed a bachelor's degree had departed the STEM track in large numbers. This suggests that an integrated model predicting a Degree vs. No Degree outcome might benefit from the inclusion of the survival analysis module.

9.1.1.1 The Need for Data Refinement with Large Datasets

The findings demonstrated that logistic regression is an excellent technique to apply in situations where predicting between two very diverse outcomes is desired. This technique is especially applicable to educational research instances when data that are categorical in nature are recoded to create new covariates that are binary or ordinal in scale. Developing models with a large set of potential covariates requires an extensive effort in advance to prepare the data for model fitting with logistic regression. There is no substitute for carefully examining each variable being considered for the model to understand what information it conveys and what the possible values mean. Variables with nominal values need to be recoded to create sets of binary

"dummy" variables if an ordinal scale cannot be designed. Records with variable values that are missing must be examined to determine if values may be imputed or if the records must be excluded. Some variables may legitimately not have responses for every record. For example, a variable indicating the date that a student first dropped out of high school may have many records coded as "legitimate skip" because many students never dropped out. In cases such as this the value of the variable should be coded for those records in a way that does not inadvertently affect the model. One option is to code the variable value as 0. If the data is not originally binary, ordinal, or continuous in scale the data must be adjusted in advance. Failure to perform this data refinement in advance will lead to models that do not properly utilize the data and may be less accurate.

9.1.2 Identification of Significant Predictors

This research found a set of significant predictors for STEM outcomes in response to the issue: "Can a set of variables that were measured for a group of students as they progress through high school and beyond be shown to affect the probability that a given student fails to "survive" to achieve a bachelor's degree in Science, Technology, Engineering, or Mathematics (STEM)?"

This analysis has determined that there is a set of predictor variables for modeling the different educational outcomes that is reasonably consistent across 11 random samples of the students. These variables include measures of academic skills in math, science, and reading; measures of personal confidence in academic capabilities; measures of the students' academic focus on schoolwork and career; and measures of demography including sex, race/ethnicity, and native fluency in English. The variables found to be significant predictors in the logistic regression and integrated models confirmed the findings of prior research. In addition, through

testing a much larger set of potential variables this research found additional significant predictors of earning a STEM degree including family composition; parental marital status; type of high school [private religious or private nonreligious] the student expected to attend; father's highest level of education; student television consumption habits; hours student worked for pay per week; and base salary of beginning teachers at student's school. Thus, additional predictors have been identified as affecting the probability that a given student failed to survive on the STEM track. The majority of the significant variables were those available by 8th grade. The STEM vs. All Else model tested with solely 8th grade variables performed almost as well as the one that also included the SAT and ACT scores as potential predictors.

9.1.2.1 Controllable and Not Controllable Predictors

The significant predictors fall into two categories: Not Controllable and Controllable. The Not Controllable predictors are beyond the influence of education policy makers. These include variables that reflect the student's fluency in English, the student's family composition, the parents' highest levels of education, and the parents' marital status. While these variables cannot be influenced directly, they can be useful in determining the students that may be "at risk" of departing the STEM track. In addition, the student's racial/ethnic group and gender may also be helpful in identifying these "at risk" students for the models in which they are significant predictors. Some of the school characteristics which were found to be significant for a few models are not directly controllable. These school characteristics include the percent of white non-Hispanic 8th grade students and the percent of 8th grade students that are in single parent families.

The Controllable predictors potentially can be influenced by education policy. These include the student's mathematics proficiency, Science proficiency, English proficiency,

academic performance, standardized test scores, and school characteristics. Improvement of academic instruction and encouragement towards positive attitudes for STEM careers are within the direct control of educators. Some school characteristics such as the starting salary for a new teacher with a B.A. degree are also within the control of education policy makers.

Other significant predictors are capable of being influenced, but they lie outside the direct control of educators. These include parental expectations of student's educational achievement; family rules regarding the student's time spent working for pay, doing homework, maintaining grade point average, or watching television; parental involvement in encouraging students to pursue a college degree; and parental choices in the type of high school for the student. These predictors could be changed by the student's parents so it is possible that they could be indirectly influenced by educators encouraging parents to consider beneficial changes. For example, a school could invite parents to attend an educational seminar that discusses the factors which influence students' interest in STEM and capability of pursuing a STEM degree. Other significant predictors are under the control of students and these factors can be potentially influenced by educators. These include attitudes towards school, attitudes towards educational attainment, and the investment of personal time towards school/work/social activities.

Table 9.1 lists the variables that were found to be significant for at least one of the 11 samples used to fit STEM vs. All Else models. The table classifies the variables by the extent to which they can be influenced by educators.

Table 9.1 Summary of Educator's Ability to Affect Significant Predictors of STEM

Variable	Directly Controllable	Indirectly Controllable	Non-Controllable
Overall Math Proficiency or Math Quartile	X		
Overall Reading Proficiency or Reading Quartile	X		
Science Quartile	X		
Student's ability group for Mathematics		X	
Student's ability group for Science		X	
Mathematics grades from Grade 6 to 8	X		
Science grades from Grade 6 to 8	X		
ACT (English Score)	X		
ACT (Mathematics)	X		
SAT (Verbal)	X		
SAT (Mathematics)	X		
Min. GPA Required to Participate in Activities	X		
# of students in Remedial Reading		X	
# of students in Bilingual Education		X	
# of students in English as 2nd Language		X	
# of students in Gifted, Talented Ed		X	
% of White Non-Hispanic 8th Graders			X
# of students in Free Lunch Program			X
Family rule re: how early/late child watches TV		X	
# of hrs Student watches TV on weekdays		X	
Family rule re: maintaining grade average		X	
How often parent talks to child re post H.S. plans		X	
How far in school parent expects child to go		X	
How far in school student thinks he/she will get		X	
How sure that you will graduate from H.S.		X	
# of Hrs student works for pay per week		X	
# of BY Risk Factors for Dropping Out of School		X	
# of Cigarettes Student Smokes per Day		X	
Language Minority Composite			X
H.S. Student Plans to Attend: Private Nonreligious			X
H.S. Student Plans to Attend: Private Religious			X
Family Composition: Mother & male guardian			X
Family Composition: Mother			X
Parents' Marital Status: Divorced			X
Yearly Family Income			X
Father's Highest Level of Education			X
Sex			X
Race/Ethnicity			X

It should be noted that these significant predictors measure aspects of students that were in 8th grade in 1988 and 12th grade in 1992. American culture has not remained static during the

past 20 years. The activities, experiences, and attitudes of current day students cannot be assumed to have remained constant. For example, computer literacy has become a vital skill in today's educational and employment spheres. Time spent socializing with friends via the Internet or in self-entertainment with computer games have become much more prevalent in 2008. The exact set of predictors found to be significant in this analysis may not be the ideal set to use in attempts to replicate this research in the future. This set of predictors is a logical place to start in replicating the research, but it would also be wise to consider what new measures may be useful.

Given that various measures of mathematical ability at different points (BY and F2 standardized test scores) were significant in most models, assessment of mathematical competence is likely to remain important in future modeling. Science, English, and reading proficiency are also likely to remain valid in future models. Variables that assess the behaviors and attitudes of the students and their families towards educational attainment and career development are worth evaluation.

9.1.3 Application of Survival Analysis to STEM Research

This research found that survival analysis could be applied to the STEM degree acquisition process and it provided valuable insights into variations between different groups of students over time in the probability of earning a STEM degree. These insights were obtained in the process of addressing the following questions: "Will Survival Analysis of the NELS:88 data reveal that the probability of a student achieving a STEM degree differ over time for students in different outcome groups?" and "Are there key time points in the educational process where distinct decreases or slight increases in the probability of achieving a STEM degree occur as

students developed academically? If so, are these key time points at which students were most likely to depart the STEM track sufficiently common for different student profiles that they could suggest the timing for delivery of pro-STEM intervention?"

Differences in the STEM probabilities by outcome group were found. This can be seen by examining the cumulative hazard functions in Figure 6.1 as developed from the logic described in Section 6.3 for establishing the track departure times. The high school dropouts had a markedly different hazard function with all the students in this group departing by year 5 of the study. The high school graduates departed the STEM track at an increasing pace with a sharp jump during year 4 when most earned their high school diplomas. The college dropouts and students who earned a sub-4 year degree had very similar curves that were dissimilar to the other hazard functions until after year 8. As mentioned earlier, the hazard functions for the STEM and Other Degree students were similar at some time points and divergent at others. These two groups had hazard functions that were clearly different from the other departure types.

It was envisioned that key time points in the educational process could be found to exhibit distinct changes in the probability of earning a STEM degree over time across different groups of students that could indicate the best timing for a pro-STEM intervention. Again, examination of the cumulative hazard functions for each group as shown in Figure 6.1 provides evidence to answer this research question affirmatively. The probability of earning a STEM degree drops as the hazard function increases. The students that graduate high school and go on to attend college did not experience increasing hazard function values until 3 years past the study start in approximately 1991. This suggests that a pro-STEM intervention conducted in high school would be able to target these students prior to their leaving the STEM track. Year 3-5

represents the period at which the probability of departing the STEM track rises the most sharply for the students that drop out of college or complete less than 4 year degrees.

The hazard functions for the high school dropout and students completing their education by graduating high school exhibit different track departure patterns. The high school dropouts experienced the sharpest increase in the probability of STEM departure at the study's start with a more gradual increase until midway through their junior year. The students that graduated high school experienced a less steep but steady increase in the probability of departure until year 4.

The conclusion reached from examining the hazard functions is that to target potential STEM degree students successfully, the pro-STEM intervention must occur before 8th grade. To reach all of these students, the intervention may have to occur in the 7th grade or earlier. The curves for the students whose educations did not go beyond high school were sufficiently dissimilar to those of the other students that it may be worthwhile to consider developing more than one intervention program. The first would occur prior to 8th grade and would focus on assisting students that would not otherwise be predicted to go on to college. The second intervention program would occur prior to 11th grade and focus on encouraging students that are predicted to attend college to consider pursuing a STEM degree. A potential third intervention program would take place after the first year of college for students as STEM students consider switching to a major outside STEM.

9.1.4 Potential Intervention Programs

The findings of this research suggest that there are three types of intervention programs that could be developed to increase the number of STEM students. The first would focus on improving students' capabilities to pursue STEM and be delivered prior to 8th grade. The second

would be oriented towards increasing students' interest in a STEM career and would be delivered in the last two years of high school when students are making college plans. The third would concentrate on encouraging STEM students considering switching majors to remain within STEM.

Students with greater capability and prior academic performance in mathematics are more likely to succeed in a STEM subject. Some measures of capability and performance in Science are also predictive of STEM success. Proficiency in speaking English is undoubtedly also a positive factor in completing a STEM degree in the United States. Each of these subject areas can be addressed through concentrated educational efforts to improve students' skills in these topics. The first type of intervention program could be directed towards assisting students in acquiring greater skills in the critical subject matter topics.

The second type of intervention program would address the issue of students that could perform well enough to earn a STEM degree but might not choose to pursue this line of study. These are students who either have little awareness of the benefits of a STEM career or have more interest in another subject. This intervention program would be designed to educate students about the interesting careers available to STEM graduates and wide-ranging applications of a STEM degree. This would serve to encourage the many capable students that might otherwise pursue a STEM-Related or Non-STEM degree to at least consider STEM. It would also encourage the students that might already be drawn to STEM to develop a stronger interest.

The third type of intervention program would attempt to reduce the number of capable STEM students that choose to change to a non-STEM major. These are students that are performing well in STEM but for some reason are less interested in STEM or have a greater

interest in a Non-STEM subject. The intervention would attempt to determine if the students have acquired a negative misimpression of STEM that could be countered with positive encouragement or suggestions to consider a different major within STEM that might be a better fit. Students that have genuinely lost interest in STEM should be encouraged to pursue a more personally appealing major.

Using ROC Curves to adjust the models' sensitivity would help education policymakers select the optimum target audience for an intervention program. The optimum target audience depends on the policy goals. Such goals may include balancing the models' sensitivity and specificity if the costs of correct and incorrect predictions are the same; reaching the largest number of potential STEM students that the budget will permit; or selecting the students that fall within a middle strata of STEM interest to focus on students that could pursue a STEM degree with extra encouragement but wouldn't necessarily pursue it on their own.

9.1.5 Defining Educational Outcomes

Dividing the potential educational outcomes into the five basic categories of No Degree, Sub-4 Year Degree, Non-STEM, STEM-Related, and STEM clarified the definition of STEM. The finer divisions in this categorization made it easier to compare these results to those of previous analyses since it was clear which educational outcomes were included in each model. This division allowed great flexibility in modeling potential outcomes since two-outcome pairs of the basic categories can be modeled as well as combinations of the categories. Prior educational research focused on modeling a few of these potential outcomes such as STEM vs. All Else or STEM vs. Other 4 year Degree. Comparing new results to the prior research depends on the ability to determine which potential outcomes are included in the new and prior models.

Further subdividing the No Degree category into high school dropouts, high school graduates, college dropouts, and students with degrees in progress at the study's end was also useful. It allowed more detailed examination of these students' similarities and dissimilarities with students in the other categories. The results of the integrated model and analysis of the instances of false STEM predictions suggested that many of the college dropouts shared qualities of the STEM students and had the potential to complete a college degree.

The formal definition of STEM as a narrowly defined vs. expansive collection of majors was supported by this analysis. It appears that the students who go on to obtain a four year degree have more in common as a group than they do with those students that do not achieve a four year degree. This suggests that if a student is interested and capable of obtaining a four year college degree then he or she is at least a fair candidate for considering a STEM degree. These students represent the most obvious pool of students to target as potential STEM candidates. If such students are identified at an early point in the secondary school process and encouraged to consider STEM, there will be several years to improve their academic capabilities as needed to earn a STEM degree. Overall, this means that future educational research should focus narrowly on defining STEM when attempting to model earning a STEM degree as an outcome.

9.1.5.1 The Advantages of Defining a STEM-Related Category

The STEM-Related category has utility as a way to examine the limits of the STEM and Non-STEM categories. Selectively reclassifying majors between STEM, STEM-Related, and Non-STEM enables exploration of the sensitivity of the models to a reclassification of a single major provided that enough students in the sample earned a degree in that major. The STEM category should be narrowly defined in order to maintain a higher degree of predictive accuracy in future modeling exercises. While the STEM and STEM-Related students take many of the same

quantitative coursework in college their focus is different. Intuitively, it would seem natural that the STEM and STEM-Related students would be more alike than the STEM-Related and Non-STEM students. Yet this was not found to be the case. The STEM-Related students are acquiring skills that they can apply directly in their professional careers. The STEM students are learning the reasons why the analytical tools work and how they can be adapted to work in new ways and for new applications.

9.2 RECOMMENDATIONS

9.2.1 When to Use Integrated Models

Developing an integrated model is more involved than applying a single analytical technique. An analyst might very well wonder why the extra effort should be made. For a less complex problem, a single technique might be sufficient. When the problem is so complex that a simplified subproblem is not realistic enough to capture enough of the factors to offer informative results, multiple techniques should be considered. Adding another analytical technique should make it possible to examine more of the factors that affect the problem provided that the techniques are additive. For example, if the standard analytical technique produces a 60% accuracy level with predictions and integrating a second technique increases the accuracy level to 75% then the techniques are additive. The type of data that can be collected will guide the selection of potential analytical tools. The nature of the problem and the focus of the improvement sought will also affect the choice of analytical tools. When techniques can be combined to offer the desired additive improvement, they are candidates for integration. The

next question to consider is whether the improvement would be significant or explain the results in a better way. If the first technique explains so much of the variation that it provides a superb level of predictive accuracy then the potential benefit from integrating another technique may not be worth the additional effort required.

In the application considered in this research, the standard technique did a very good job of predicting the educational outcomes and left little room for improvement by the integration of survival analysis. The reason that the logistic regression models were able to achieve such good levels of accuracy was the extensive effort made in preparing the data for modeling outcomes via logistic regression. This effort began with the selection of potentially promising variables after reviewing prior research to identify factors previously found to exhibit significant differences between STEM and Not-STEM students. The findings of prior research led to the selection of variables in NELS that provided the same or comparable information. Then the net was cast more widely to select variables that would provide additional insights into the family structure, academic capabilities, experiences, and attitudes of the students towards school.

The variables had to be carefully considered since the dataset had an enormous number of potential variables with nearly 7,000 variables from the students' high school years. Of these variables, the vast majority were redundant, not practical, or irrelevant to this research. For example, the students' race and gender were recorded in each of the five waves of data collection. This was done to ensure the most accurate accounting of the students, but only one set of these variables was ultimately used for the analysis. Other variables measured details about the students' teachers and schools. While these variables offered interesting insights into the students' educational experiences, their use in modeling was not supported by prior research. Including these extraneous variables would have required extra analytical time, and they were

not good candidates to recommend for future data collection to analyze the prediction of quantitative degree outcomes. Still other variables would have been useful for analyzing different research hypotheses but offered little insight into educational outcomes. A separate variable selection effort was required to identify variables that could be used to determine the students' final educational outcomes.

The use of the high school graduation status and the first two majors and degrees reported by the students proved to be insufficient to determine the final outcomes when attempting to derive the students' time to departing the STEM track. Additional variables from the post-secondary education transcript file were required to resolve gaps and inconsistencies in the data. Future attempts to analyze students' educational outcomes and persistence should involve gathering similar data that not only reflects significant predictors, but also identifies the outcomes in question.

Once the NELS variables were selected the set of potential values had to be reviewed so that they could be converted into recoded variables with binary (i.e. 0 or 1 values), ordinal, or continuous values such that the logistic regression model could employ them. Without the careful recoding of the original variables selected for the modeling the regression models would have been less accurate and using such a large set of potential variables in the modeling would have been infeasible. Examining the data with survival analysis provided additional insights into the research by clearly highlighting the differences and similarities in the cumulative hazard functions for the different groups. This provided insights into the timing for different potential intervention programs.

9.2.2 Data for Educational Outcome Research

The NELS dataset created by NCES contained a wealth of variables that made it extremely useful for examining the educational progression of students from high school through college. These included demographic, experiential, and attitudinal variables for the students as well as objective measures of their academic capacity and performance. Other variables indicating the students' course of study at college and work experiences were provided. Information about the students' schools, communities, and parents were made available. The dataset was gathered for the purpose of enabling a vast array of potential educational research rather than a single focused line of analysis. As such, it is very applicable for different purposes, but such application may require intensive examination of the data to develop a way of using it for a single purpose. There are several things that could be done in future data collection designs by NCES to better support research that continues the analysis performed in this dissertation as well as other qualitative analyses of educational and societal outcomes.

Since the findings of this research indicate that mathematical skills are so important to the prediction of educational outcomes, it would make sense to focus heavily on this subject in future data collection designs. Surveys of students should contain separate questions to gather the students' experiences in taking different mathematic subjects such as trigonometry, geometry, algebra, calculus, probability, and statistics to learn if and when the students took these classes and the grades they earned. The NELS:88 design included variables to measure the number of years of coursework and total number of "Carnegie units" that were taken in various math subjects. Explicit measures of how well each student performed in each math class taken would be particularly valuable. Any standardized tests taken by the students to measure their mathematical competency including the PSAT, SAT, and ACT should continue to be obtained.

These variables have consistently been found to be valuable predictors of educational outcomes but they may not be enough to completely explain results. The examination of these variables may offer additional insights into the differences between STEM and STEM-Related or Non-STEM students.

The strong logistic regression models developed in this research were the product of extensive manipulation of variables to create recoded versions with a binary or ordinal scale. Future longitudinal datasets from NCES would benefit from the creation of binary/ordinal versions of key variables such as the parents college degree status, standardized test scores, and language minority status. The variable "sex" had a dichotomous scale, but the responses had been set to 1 and 2 rather than making it binary originally. There were several variables within the NELS:88 dataset reporting race/ethnicity, but a complementary set of binary dummy variables would have better enabled the modeling process.

As previously mentioned, American culture has changed dramatically since the late 1980's, and the pace of these changes has been very rapid. The range of activities available to students has expanded accordingly. Several variables that measured television watching habits of students were significant in the modeling process. These variables also may be valuable in the future, but it would be logical to increase the set of variables collected by examining currently popular activities such as playing video games via the internet or handheld devices, participating in social networking websites, communicating via instant messaging software, communicating via text messages, communicating via cellular telephones, shopping for products via the internet, etc. Many of these activities may stimulate mental skills in mathematics, science, or strategic planning such as those which encourage students to take on the role of planning a city's development. Other games that have a physical interaction such as the Nintendo Wii™ units

have implications for physical development. Still other videogames serve solely entertainment purposes.

The manner in which educational material is disseminated and assimilated has also expanded. More schools have integrated multi-media in the educational process most notably by including personal computers as a teaching and learning tool. Thus a student's level of computer literacy can affect how well the student is able to use computer based learning tools, online reference materials, and analytical software. A very basic skill that directly affects the speed at which students can use a computer is their ability and proficiency to type. This presents other potential variables that should be collected in future attempts to replicate this research.

9.2.3 Implications for Intervention Programs

The findings from the cumulative hazard functions and the sets of significant variables for different models suggest that multiple intervention programs may be required to increase the numbers of students graduating with STEM degrees. Some of the students in the NELS:88 dataset clearly were at great risk of departing the STEM track early in the study by dropping out of high school. This suggests they were inadequately prepared academically and motivationally to continue in school. Any intervention program to advance these students towards a STEM outcome would have to have been delivered earlier then the 8^{th} grade to improve their skills, encourage them towards a technical subject, and increase their desire to at least complete high school. Additional consideration of the dropout phenomena is required to assess when this intervention would be optimally delivered.

A second potential time point for intervention concerns the later high school years when students are considering their post-high school plans and possible college major choices. This

intervention would concentrate on educating students about the interesting careers available in STEM, encouraging them to consider applying to a STEM program that suits their academic preparation, and motivating them to follow through. Most high school students have some idea what doctors and lawyers do, but they may not appreciate how a chemist, physicist, or engineer is employed. Without more awareness of the exciting array of careers available to STEM graduates, the choice to focus on a challenging STEM degree may be less appealing. A STEM career may be more appealing once students understand how these careers fit into the modern American culture. For example, a student that enjoys playing interactive videogames may be more drawn to STEM when understanding that videogame designers and special effects creators for the entertainment industry study a lot of computer programming, mathematics, and physics. The prospect of working on space exploration and development should also engage the interest of many students.

The findings that approximately 15% of the students that entered college but dropped out before graduating were incorrectly predicted to have a STEM outcome suggests many of these students could have succeeded had they stayed in school. Even more dramatically, many of the students that earned other four year degrees appeared to have the capacity to earn a STEM degree. Some of these students may well have started in a STEM major and then switched out. A third potential intervention program could be timed to the first year of college to encourage capable students to stay within or switch into STEM majors. The goal would be to retain as many capable students as possible and recruit students from other majors with an interest and aptitude for STEM coursework. One way to accomplish the retention goal might be to encourage students intending to switch out of STEM to consider a different major within STEM

that more closely matches their interests. A possible source of potential recruits is students that are taking quantitative coursework as an elective.

9.2.4 Educational Policy Implications

Preliminary findings of this research have been shared through presentations to professional associations. One common theme of the responses received has been concern that this research could be used as a tool to "cherry-pick" the best students to increase the supply of STEM students without regard to the students not selected. This is a valid concern. With the proper data, similar models could be constructed to predict the probability of a given student achieving a STEM degree so that only strong candidates would be accepted to a particular STEM program. The capable students in a particular high school could be encouraged to apply to STEM programs while the less capable students are ignored. A specific concern that has been expressed is that the variables for race/ethnicity and gender could be used to select the students that have a greater estimated probability of achieving a STEM degree with the result that desired goals for a diverse student population are not met. Each of these scenarios is possible.

An observation that may allay some of these concerns is that the need for increasing numbers of STEM degree-holders means that each student should be considered a potential candidate in junior high school. Students that are not necessarily the strongest in math may possess sufficient drive and initiative to succeed in STEM if motivated by personal interest despite their academic disadvantage. It should be the goal of educational policy makers to use this sort of research to improve their delivery of education for the benefit of all the students. If more students are better prepared academically to pursue a STEM degree the concerns about the implications of this research may be unwarranted.

9.2.5 Recommended Approach in Using Integrated Models

The results of this research led to the development of a recommended process for using an integrated model for a specific problem application. The process is depicted in Figure 9.1. As with any model, the first step is to formulate the problem to determine its parameters and decide the scope that will be explored. From there the problem's complexity must be assessed and a set of goals for the proposed solution developed. If the problem isn't particularly complex, then an integrated model may not be required. The goals for the solution will help to decide which parts of the problem are critical in data collection and how to determine if a proposed solution can produce an acceptable level of improvement. Then it is important to assess what sort of data can be obtained to describe the problem. Where in the process can data be found? What is the format of the data in terms of qualitative or quantitative? If some or all of the data is quantitative does it possess a continuous, integer, ordinal, binary, or nominal scale? The type of data that is required to measure the problem and the desired improvement will lead to a set of potential analytical methods to study the data. From this a standard single technique may be selected to serve as the benchmark or starting point for an integrated approach.

As an example, if the problem concerns missing production deadlines then it is reasonable to explore the sorts of data that go into planning and manufacturing the product. This would include how production orders are generated; how the information is conveyed to manufacturing; how raw materials and subcomponents are obtained; how the production is scheduled; how well the manufacturing process is operating; and how the work in process is transported for subsequent production processes. If preliminary analysis suggests that the process is failing to adequately plan production schedules, then production level forecasting may be the standard single technique that is employed. Other potential analytical methods that may

be useful include statistical process control for evaluating the process quality; simulation for testing alternative manufacturing process scenarios; and operations research optimization to explore the effects of product mix decisions and make vs. buy choices for subcomponents upon the production process.

Once the types of data needed and the primary analytical method are selected, it is important to consider what data is currently available and how it can be obtained. Ideally, the data is already collected and in a means that can be readily accessed. If this is the case, then the data is obtained and prepared as required for the methodology chosen. If the data is not already available, then it is necessary to design a data collection plan and implement it before beginning the analysis. The process of data acquisition may have to be repeated in several cycles if preliminary analysis indicates that additional data is required. Then the problem is modeled with the standard single technique.

The solution obtained from the initial model is evaluated to determine if the solution quality is acceptable and whether sufficient improvement has been gained. If the solution meets the criteria, then it is provided to the decision maker, implemented if approved, and the actual results are evaluated. Additional improvements are made to the process as needed.

If the solution provided by the standard single technique isn't sufficient, it may be worthwhile to consider developing an integrated model. This would start by considering other analytical methods that are appropriate to the problem data which could provide an additive benefit to the standard technique. This part of the process relies on the creativity and skill of the analyst as well as the accessibility of any additional data that might be needed for a particular technique. If additional data is required, a new data collection plan is developed and implemented. The various analytical techniques are combined in an integrated fashion and the

model solution is evaluated. Successive iterations may be required to experiment with different approaches in integrating the different techniques until a suitable method is found. The integration process is repeated as needed until an acceptable solution is developed.

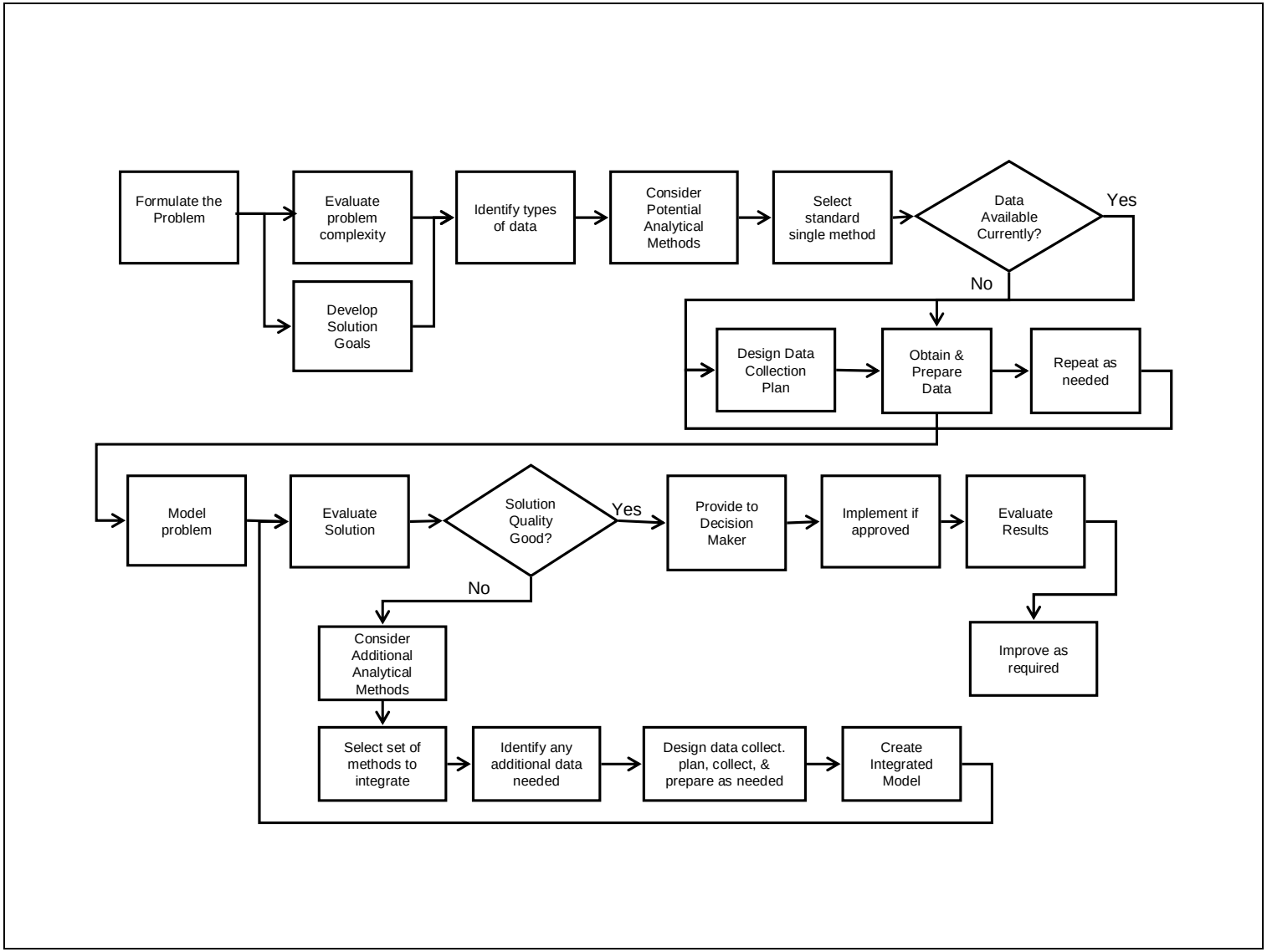

Figure 9.1 Recommended Process for Considering and Developing Integrated Models for Analysis

10.0 CONTRIBUTIONS AND FUTURE WORK

10.1 SUMMARY

This research has demonstrated that integrated models are feasible and hold promise for examining complex problems. It has produced a set of models which predict educational outcomes with accuracy that ranged from acceptable to outstanding. The logistic regression and integrated models were fitted using a large set of variables and a very large number of records. The models were created and tested with separate fit and test datasets for greater rigor in examining the predictive accuracy. ROC Curves were used to evaluate the accuracy of the models' predictions based on different threshold probability values. A formal definition was proposed for the college majors that should be considered part of STEM, and this definition was tested. The grouping of college majors into three categories (STEM, STEM-Related, and Non-STEM), showed interesting similarities and dissimilarities between the students in the different groups. A large set of variables that have been found to be predictive of educational outcomes in prior research was tested and several were found to be consistently significant predictors of post-secondary educational outcomes in this analysis. The findings led to conclusions about the logical next steps in continuing this research.

10.2 CONTRIBUTIONS OF THE DISSERTATION

10.2.1 Creation and Testing of an Integrated Model

The main contribution of this dissertation is the development and evaluation of an integrated model that could be applied to more general problems. The feasibility of employing multiple statistical techniques to data collected at different points in time across a complicated problem with interconnected factors and linking the techniques together has been demonstrated. The result in this case is a tool that can be used to predict post-secondary educational outcomes with good accuracy and identify during high school those students with a higher potential for completing a bachelor's degree in STEM.

While the integrated models developed in this research did not provide a statistically significant improvement in predictive accuracy, they did provide good accuracy and a starting point for further exploration of the integrated modeling approach. As discussed in section 9.2, there were two aspects of this analysis that may have hampered the integrated models' functioning. The first aspect that made it harder for the integrated models to show a significant improvement was simply that the logistic regression models were more accurate than anticipated. The second aspect was defining the STEM track departure times in a manner that did not provide much contrast between the STEM and Other Degree students. The results in this application suggest if the integrated approach were employed with a different design interpretation or to a different problem application, significant improvements might be gained over a single technique solution.

10.2.2 Developing a Process to Create and Evaluate Large Logistic Regression Models

Multivariate logistic regression models have been developed in prior education research, but the scope of the models created in this research was much larger. The models created in this analysis examine student records over a longer time period and are able to achieve very good accuracy in predicting post-secondary educational outcomes from an early point in high school. These logistic regression models were fitted with the use of a much larger number of records and potential covariates than has been the case in prior research. The process outlined for selecting potential covariates, recoding the covariates for maximum utility, selecting random samples to fit and test the models, generating prediction results, and evaluating the results is a sound approach for future research. Logistic regression models are often evaluated solely on the basis of how well a particular model explains the results observed in the data used to fit it. Extending the evaluation to show how well the fitted model works when applied to a new set of records and exploring how responsive the predictive accuracy is to a change in the prediction probability cutpoint for the test data has not been done in prior education research. The unusual combination of a very large dataset, a large set of potential variables, a longitudinal breadth of 12 years, a rigorous evaluation method, and a means of easily adjusting the sensitivity/specificity made these logistic regression models different from those of prior research.

The models achieve a great deal of accuracy in discriminating between diverse potential outcomes using a large assortment of demographic, attitudinal, and experiential data gathered during the 8^{th} grade from the students and their parents. The analysis confirms the utility of many variables previously found to be significant predictors of STEM interest as well as identifying new predictors. The models cover a wider scope of potential educational outcomes

including high school dropout, high school graduate, college dropout, and an array of different college degrees than has been examined simultaneously in prior research.

The value that was gained by using the NELS:88 dataset was a direct result of extensive preparation of the data. The NELS:88 data that was gathered for general research purposes and not with this particular type of analysis in mind. Adapting it for this research required extensive manipulation of the data to select the most appropriate set of variables; interpreting a large set of variables to develop record classifications; recoding potential covariates to a purely binary, ordinal, or continuous scale; reconciling conflicting information from different variables; and developing logical rules for handling cases of missing data. During the course of this analysis a number of methods were designed for the efficient manipulation of huge quantities of data and particularly when much of the data is categorical in nature. These methods are applicable to more general problems than just educational research.

10.2.3 Extending the Application of ROC Curve Analysis to Education Modeling

ROC Curves are commonly used in other fields such as medical research to evaluate the predictive ability of diagnostic tests and statistical models. The manner in which they have been applied in this research is new to educational research modeling. Previous analyses have relied on the ROC Curve by using the area under the ROC Curve (AUC) value to assess how well a particular model explained the observed results in the fit dataset. The AUC values have also been used in the past to compare the strengths of different models. The sensitivity and specificity at different cutpoints has been mentioned as an option for the analyst to select a preferred value, but this has not been previously utilized.

The extension offered by this research is incorporating the ROC Curves to visually depict the different combinations of sensitivity and specificity achievable with a particular model that not only permits adjustments to the model but provides a simple way to directly compare the strengths of different models when applied to test data rather than merely fit data. It also provides a powerful means for an analyst or policy maker to adjust a particular model to suit individual policy goals. This greatly increases the potential utility of the models by minimizing the need to do additional sophisticated analysis or further modeling in order to apply a given model for differing goals. One of the greatest strengths of the ROC Curves is their ability to communicate the models' strength and tradeoffs in predictive accuracy to a wide audience in a manner that can be easily understood. Policy makers can use the curves to explain the impact of particular policy goals and how best to use the models once decisions have been made. Without seeing the illustration of balance between sensitivity and specificity it can be hard to grasp why a particular target audience of students should be selected for a proposed intervention program.

10.2.4 Creating a Formal Definition of "STEM"

This dissertation led to a proposed definition of STEM beyond the conventional practice of grouping majors that are labeled as part of Science, Technology, Engineering, and Math. Other research efforts have employed varying combinations of majors based on assorted criteria, but the concept of formally testing groups of majors to determine if a narrow or expansive definition of STEM was warranted is new. Formally proposing and testing whether a distinct third category of majors between STEM and Non-STEM should exist is also new. The concept of the STEM-Related category is to reflect the degree programs that involve extensive quantitative and technical coursework, but whose graduates will apply differently in their careers than the STEM

graduates. The STEM and STEM-Related students may have a great deal of common coursework, but the STEM-Related graduates will not be applying their skills in the research and development capacity that STEM graduates are expected to. STEM graduates are trained to use the acquired knowledge and skills to adapt them in new ways that extend the body of knowledge. In contrast, the STEM-Related graduates are envisioned to apply the knowledge and tools they have acquired to specific problems rather than creating new ways to use the tools or new methods. It should be noted that some STEM-Related students may also be oriented towards STEM and the converse is true. In this research, students that earned degrees in both STEM and STEM-Related topics were classified as having a STEM outcome.

Use of the STEM-Related category allows a more precise categorization of STEM and Non-STEM. The classification of majors meant that the "Not-STEM" outcome could be broken down much more finely than before to suit the analyst's goals whether looking at all possible educational outcomes, just those that involved completing a bachelor's college degree at minimum, or college degree outcomes that involved more vs. less quantitative coursework.

The effort to create and test a definition of STEM was instrumental in deciding to create a series of finely divided post-secondary educational outcomes. Breaking the potential outcomes up into a series of clear and precisely determined results made a much wider array of models possible. This series of outcomes combined with the extensive number of records in the dataset allowed many more two-outcome models to be fitted and evaluated. It provided a means for future research to directly compare results for a wide set of potential outcomes with those of prior research. As long as it can be determined which group of outcomes a previous research effort considered, results can be compared to those obtained by using these models.

10.2.5 Applying Survival Analysis in a Unique Manner

As discussed earlier, survival analysis has been used in previous educational research to explore trends in dropping out of high school, remaining employed as a secondary educational teacher, and completing a college degree. In those cases analysts were attempting to spot critical points in time when the risk of an event occurring changed dramatically and to determine if the probability of "surviving" past a given time was significantly different for various groups of people. This research is unique in applying survival analysis to predict which of two educational outcomes individual students will have as well as exploring the sensitivity of the predictions via ROC Curves. Numerous issues in applying survival analysis to an educational application that covers such a large, disparate group of students were encountered and resolved. This provides a very useful support for additional research to build upon in applying the powerful abilities of survival analysis.

10.3 FUTURE RESEARCH

The potential lines of inquiry suggested by this dissertation include examining the integrated model approach in applications other than educational research; evaluating other statistical analysis tools for incorporation into an integrated model; and exploring additional integrated models for educational research that could achieve greater predictive accuracy. Additional educational analyses that are logical extensions of this research are testing the limits of the STEM definition by selectively including or excluding individual majors; examining the sensitivity of the models to alterations in key significant variables to obtain a road map for pro-

STEM intervention programs, developing intervention programs and forecasting their potential impact on the probability of a student earning a STEM degree; developing models to predict academic strength at high school graduation from an earlier point in the educational process; and examining the high school dropout phenomena in greater detail.

One of the most important tests of the integrated model approach is applying it to applications beyond educational research. The transportation network described in section 1.2 would be a good test of the integrated methodology. Another scenario would be examining the effectiveness of a manufacturing process by predicting the lifespan of a product based on reliability data and potential covariates of the manufacturing process. Medical research applications would be another area in which integrated models could be beneficial. The human body is an extremely complex system in its own right. The effects of genetics, environment, treatment decisions, and age provide a rich opportunity for data collection and analysis using multiple techniques focused on the different sorts of data available.

Models that predict commodity pricing would be a potentially valuable use of the integrated model in the current economic environment. Commodity prices are affected by a complex system of supply, demand, transportation requirements, pre-consumption processing, supplier channels, competing products/uses, tax policy, import/export policy, political influences, and unexpected events that can shock the systems. For example, the price of gasoline sold in the U.S. is a function of foreign/domestic petroleum supplies; international and domestic demands for petroleum products; seasonal variations in demand; costs to access and deliver petroleum products; refining costs; refinery capacity and mix of product requirements; distribution network efficiency; prices of biofuels, nuclear, and other energy sources; demand for petroleum used as a raw material in other production processes; state, federal, and national taxation levels; the

currency exchange rate; the import/export policies of producing and consuming nations; the effects of political decisions in numerous countries; and factors that are difficult to predict such as adverse weather events, failures in drilling/refining equipment, transportation accidents, terrorist acts, and political instability. The price consumers in the U.S. ultimately pay depends on how all of these factors play out over a long period of time. The petroleum industry and petroleum consumption are of great concern due to the critical effect of this resource on international economies and the lives of ordinary people, and there are certainly other commodities for which integrated pricing models could be beneficial.

It is logical to consider alternative analytical techniques in assembling the modules of a potential integrated model. For example, a particular application may feature data that would be suited to time series analysis, factor analysis, principle components analysis, or other types of nonlinear regression analysis. It would be critical to consider which techniques are best suited to the particular type of data available to measure the problem across its scope. In the pricing commodity example, time-series forecasting could be tested to project seasonal demands at a future time; a nonlinear regression module could be used to examine the effects of disparate influences on the demand and supply; and a game theory approach could be explored as a way of capturing the effects of nations and groups in opposition to one another.

Among the potential extensions of this research in the educational field, there are options that could be explored to determine if greater predictive accuracy can be achieved with the integrated models when applied to the NELS:88 data. These range from modeling more starkly disparate pairs of outcomes to adjusting the survival times to illustrate the differences between the STEM and All Else students more clearly.

The STEM vs. Non-STEM classification of majors can be tested further by selectively changing the assignment of majors classified as STEM-Related to the other four year degree categories. This would provide a rigorous test of the limits of the classification proposed in this dissertation by considering each candidate for the STEM category to see if it warrants inclusion. In addition, it would be beneficial to test the classification against student outcomes observed in different studies. Additional educational datasets can be examined with this technique to ensure that the classification scheme is valid regardless of the dataset employed. It is also worth noting that since the students in the NELS:88 dataset completed their college education the array of subjects that a student may earn a bachelors degree in has expanded. Expansion of previous subjects such as biotechnology and entire new fields such as nanotechnology has dramatically expanded the array of academic degree subjects. Considering the expansion of academic inquiry, it would be valuable to see how the definition of STEM might change.

The classical vision of sensitivity analysis in optimization involves determining how the results change if the value of a key input variable is altered. In this setting, that would involve measuring the impact on the probability of a STEM outcome or survival on the STEM track past a particular point from changing a significant predictive variable by one unit. For example, if a student with a given vector of covariates had a 100 point increase in the SAT mathematics score would this change the predicted educational outcome for this student and if so, by how much? This version of sensitivity analysis would be complicated by the likely correlation between variables. Since multiple measures of mathematical skills were found to be significant, a change in one could affect the values for other measures. Another consideration is the classification of variables as controllable or not controllable. If the goal is to effect change in a positive direction, the sensitivity of non controllable variables may provide insights, but the sensitivity of

controllable variables may prove more valuable. Such analysis could provide the basis for designing an intervention program by indicating which of the factors education policy makers control have the greatest impact on the probability of a STEM outcome. The design of potential intervention programs is a logical extension from there. Given that a specific set of vital covariates that can be influenced exists, what form should a pro-STEM intervention take? How would a given program affect the likely numbers of STEM students resulting? What is the impact of including different students in the target audience?

It would be easier to strengthen students' academic skills if deficiencies that could lessen their STEM potential were identified at an earlier time in the educational process and dealt with then. By 8^{th} or 10^{th} grade the students' prior academic preparation has become critical in their potential for going on to college. While this may be an opportune moment to target the college-bound students for a pro-STEM intervention, it may not be in time to appreciably alter the trajectory of other students. It would be interesting to explore data from earlier in students' academic career to determine if it could accurately predict what their academic performance, attitudes towards STEM, and intention to pursue a college degree would be as they near the high school graduation. If significant predictors could be found at an earlier point in the educational process it would open up the possibility of designing effective intervention methods to assist students in being better prepared for high school as well as college.

The importance of strong skills in mathematics has been apparent through out the different models. Although multiple variables assessing mathematical skills were used in this analysis, none dealt with the question of which mathematical skills were the most critical to the probability of achieving a STEM degree. Among the mathematics courses that are potentially available to students in high school are algebra, geometry, trigonometry, statistics, probability,

and calculus. Future research into the STEM degree acquisition process should examine the effects of competence in these different mathematical topics to determine which have the most effect on the probability of earning a STEM degree.

As discussed in section 2.5, there has been prior educational research done into the subject of high school dropouts. This is an area of concern to society since failing to complete high school can limit employment and future educational prospects. Unless students are motivated to earn a high school diploma with strong academic skills they will also be less likely to complete a STEM degree. Modeling the dropout phenomena to identify factors that predict whether and when students will leave high school is a potentially useful application of the integrated methodology.

All of these options for future research offer the opportunity to expand knowledge of how an integrated methodology can be employed to solve complex problems. With all of the analytical tools available to industrial engineers the possible combinations of techniques that can be integrated for application to complex problems are numerous.

APPENDIX A

BACKGROUND FOR NELS:88 VARIABLES

A.1 NELS:88 "UNIVERSE" VARIABLES

The F4UNIV1 variable indicates the status of the students in the NELS:88 dataset during each of the five waves of data collection. It consists of a four digit code that corresponds to a combination of five alphanumeric codes that are the status indicators. The alphanumeric codes for the waves begin with two digits that indicate the status related to either the base year (BY) in 1988 or one of the four follow up waves of data collection (F1, F2, F3, or F4). The only exceptions to this are for the codes BNA for Base Year Not Applicable and 1NA for First Follow-up Not Applicable. After the digits indicating the wave of data collection is a one to two digit code for the status of the student in that wave. Table A.1 lists the set of potential status codes for the data collection waves.

Table A.1 Summary of Student Status Codes for the Data Collection Waves

Status Code	Status Description
BNA	Base Year Not Applicable
1NA	First Follow-up Not Applicable
A	In-school, in grade
B	In-School out-of-grade
D	Dropout
E	Eligible
F	Freshened
G	Received GED/HS Equivalent
H	Received HS diploma
I	Ineligible
N	Not Pursuing GED/HS Diploma
O	Subsampled Out or Equivalent
P	Pursuing GED/HS Diploma
X	Out of Scope (e.g., deceased)
1ER	F1 Sampling Error
2ER	F2 Sampling Error
?	Status Unknown
Q	Respondents

For example, cases where F4UNIV1 = 1038 indicate that the students had status codes of "BYE F1B F2D F3G F4Q." These codes translated into an overall status summary of Base Year Eligible (participated in the 1988 wave), in school during F1 but not in the grade 10th grade that would have been expected, dropped out status in F2, received a GED or high school diploma equivalent by F3, and responded during the F4 wave.

Table A.2 lists the set of F4UNIV1 numeric codes with their associated alphanumeric status meanings and the number of students in each category. The table also indicates if the students participated in all five waves of data collection, were ever in drop out status, or ever failed to respond during one of the waves of data collection.

Table A.2 Summary of "Universe" Variables indicating Student status during the waves of NELS data collection

F4UNIV1	BY	F1	F2	F3	F4	Freq.	Dropout at some point	Non-response at some point	Partic. in all 5 waves	Freq. of All 5 partic.
1001	BYI	F1A	F2A	F3H	F4Q	84			No	
1002	BYI	F1A	F2A	F3G	F4Q	1			No	
1003	BYI	F1A	F2A	F3P	F4Q	1			No	
1004	BYI	F1A	F2A	F3N	F4Q	4			No	
1007	BYI	F1A	F2B	F3H	F4Q	2			No	
1008	BYI	F1A	F2B	F3G	F4Q	1			No	
1009	BYI	F1A	F2B	F3P	F4Q	1			No	
1010	BYI	F1A	F2B	F3N	F4Q	1			No	
1011	BYI	F1A	F2D	F3H	F4Q	7	Yes		No	
1012	BYI	F1A	F2D	F3G	F4Q	6	Yes		No	
1013	BYI	F1A	F2D	F3P	F4Q	1	Yes		No	
1014	BYI	F1A	F2D	F3N	F4Q	4	Yes		No	
1015	BYI	F1A	F2D	F3?	F4Q	1	Yes	Yes	No	
1020	BYI	F1A	F2?	F3G	F4Q	1		Yes	No	
1021	BYI	F1A	F2?	F3N	F4Q	2		Yes	No	
1025	BYE	F1B	F2A	F3H	F4Q	115			Yes	115
1026	BYE	F1B	F2A	F3G	F4Q	7			Yes	7
1027	BYE	F1B	F2A	F3P	F4Q	6			Yes	6
1028	BYE	F1B	F2A	F3N	F4Q	5			Yes	5
1031	BYE	F1B	F2B	F3H	F4Q	48			Yes	48
1032	BYE	F1B	F2B	F3G	F4Q	8			Yes	8
1033	BYE	F1B	F2B	F3P	F4Q	17			Yes	17
1034	BYE	F1B	F2B	F3N	F4Q	12			Yes	12
1037	BYE	F1B	F2D	F3H	F4Q	11	Yes		Yes	11
1038	BYE	F1B	F2D	F3G	F4Q	49	Yes		Yes	49
1039	BYE	F1B	F2D	F3P	F4Q	37	Yes		Yes	37
1040	BYE	F1B	F2D	F3N	F4Q	54	Yes		Yes	54
1041	BYE	F1B	F2D	F3?	F4Q	1	Yes	Yes	No	
1045	BYE	F1B	F2?	F3H	F4Q	2		Yes	No	
1046	BYE	F1B	F2?	F3N	F4Q	3		Yes	No	
1049	BYE	F1D	F2A	F3H	F4Q	22	Yes		Yes	22
1050	BYE	F1D	F2A	F3G	F4Q	5	Yes		Yes	5
1051	BYE	F1D	F2A	F3P	F4Q	4	Yes		Yes	4
1052	BYE	F1D	F2A	F3N	F4Q	1	Yes		Yes	1
1055	BYE	F1D	F2B	F3H	F4Q	5	Yes		Yes	5
1056	BYE	F1D	F2B	F3G	F4Q	3	Yes		Yes	3
1057	BYE	F1D	F2B	F3P	F4Q	2	Yes		Yes	2
1058	BYE	F1D	F2B	F3N	F4Q	1	Yes		Yes	1
1061	BYE	F1D	F2D	F3H	F4Q	20	Yes		Yes	20
1062	BYE	F1D	F2D	F3G	F4Q	107	Yes		Yes	107
1063	BYE	F1D	F2D	F3P	F4Q	74	Yes		Yes	74

Table A.2 (continued).

F4UNIV1	BY	F1	F2	F3	F4	Freq.	Dropout at some point	Non-response at some point	Partic. in all 5 waves	Freq. of All 5 partic.
1064	BYE	F1D	F2D	F3N	F4Q	185	Yes		Yes	185
1070	BYE	F1D	F2?	F3G	F4Q	1	Yes	Yes	No	
1071	BYE	F1D	F2?	F3P	F4Q	2	Yes	Yes	No	
1072	BYE	F1D	F2?	F3N	F4Q	5	Yes	Yes	No	
1076	BYE	F1I	F2A	F3H	F4Q	4			Yes	4
1077	BYE	F1I	F2A	F3G	F4Q	2			Yes	2
1080	BYE	F1I	F2B	F3H	F4Q	1			Yes	1
1082	BYE	F1I	F2D	F3H	F4Q	1	Yes		Yes	1
1083	BYE	F1I	F2D	F3G	F4Q	3	Yes		Yes	3
1084	BYE	F1I	F2D	F3P	F4Q	1	Yes		Yes	1
1087	BYE	F1X	F2A	F3H	F4Q	7		Yes	No	
1089	BYE	F1X	F2B	F3N	F4Q	1		Yes	No	
1090	BYE	F1X	F2D	F3H	F4Q	1	Yes	Yes	No	
1091	BYE	F1X	F2D	F3G	F4Q	4	Yes	Yes	No	
1092	BYE	F1X	F2D	F3N	F4Q	4	Yes	Yes	No	
1100	BYE	F1?	F2A	F3H	F4Q	139		Yes	No	
1102	BYE	F1?	F2A	F3P	F4Q	3		Yes	No	
1103	BYE	F1?	F2A	F3N	F4Q	1		Yes	No	
1106	BYE	F1?	F2B	F3H	F4Q	4		Yes	No	
1107	BYE	F1?	F2B	F3G	F4Q	1		Yes	No	
1108	BYE	F1?	F2B	F3P	F4Q	1		Yes	No	
1109	BYE	F1?	F2B	F3N	F4Q	1		Yes	No	
1112	BYE	F1?	F2D	F3H	F4Q	5	Yes	Yes	No	
1113	BYE	F1?	F2D	F3G	F4Q	6	Yes	Yes	No	
1114	BYE	F1?	F2D	F3P	F4Q	6	Yes	Yes	No	
1115	BYE	F1?	F2D	F3N	F4Q	9	Yes	Yes	No	
1120	BYE	F1?	F2?	F3H	F4Q	3		Yes	No	
1121	BYE	F1?	F2?	F3G	F4Q	1		Yes	No	
1122	BYE	F1?	F2?	F3N	F4Q	1		Yes	No	
1126	BYE	F1A	F2A	F3H	F4Q	9486			Yes	9486
1127	BYE	F1A	F2A	F3G	F4Q	71			Yes	71
1128	BYE	F1A	F2A	F3P	F4Q	77			Yes	77
1129	BYE	F1A	F2A	F3N	F4Q	47			Yes	47
1131	BYE	F1A	F2A	F3?	F4Q	2		Yes	No	
1133	BYE	F1A	F2B	F3H	F4Q	74			Yes	74
1134	BYE	F1A	F2B	F3G	F4Q	13			Yes	13
1135	BYE	F1A	F2B	F3P	F4Q	22			Yes	22
1136	BYE	F1A	F2B	F3N	F4Q	13			Yes	13
1140	BYE	F1A	F2D	F3H	F4Q	130	Yes		Yes	130
1141	BYE	F1A	F2D	F3G	F4Q	221	Yes		Yes	221
1142	BYE	F1A	F2D	F3P	F4Q	158	Yes		Yes	158
1143	BYE	F1A	F2D	F3N	F4Q	206	Yes		Yes	206

Table A.2 (continued).

F4UNIV1	BY	F1	F2	F3	F4	Freq.	Dropout at some point	Non-response at some point	Partic. in all 5 waves	Freq. of All 5 partic.
1149	BYE	F1A	F2?	F3H	F4Q	10		Yes	No	
1150	BYE	F1A	F2?	F3G	F4Q	2		Yes	No	
1151	BYE	F1A	F2?	F3P	F4Q	2		Yes	No	
1152	BYE	F1A	F2?	F3N	F4Q	3		Yes	No	
1156	BNA	F1FA	F2A	F3H	F4Q	205			No	
1157	BNA	F1FA	F2A	F3G	F4Q	2			No	
1158	BNA	F1FA	F2A	F3P	F4Q	4			No	
1159	BNA	F1FA	F2A	F3N	F4Q	3			No	
1162	BNA	F1FA	F2B	F3H	F4Q	7			No	
1163	BNA	F1FA	F2B	F3G	F4Q	3			No	
1165	BNA	F1FA	F2B	F3N	F4Q	3			No	
1168	BNA	F1FA	F2D	F3H	F4Q	10	Yes		No	
1169	BNA	F1FA	F2D	F3G	F4Q	37	Yes		No	
1170	BNA	F1FA	F2D	F3P	F4Q	21	Yes		No	
1171	BNA	F1FA	F2D	F3N	F4Q	29	Yes		No	
1176	BNA	F1FA	F2?	F3H	F4Q	3		Yes	No	
1179	BNA	F1FA	F2?	F3N	F4Q	1		Yes	No	
1183	BNA	F1FI	F2A	F3H	F4Q	1			No	
1185	BNA	F1FI	F2D	F3G	F4Q	1	Yes		No	
1186	BNA	F1FI	F2D	F3N	F4Q	1	Yes		No	
1199	BNA	F1F?	F2A	F3H	F4Q	2		Yes	No	
1204	BNA	F1F?	F2D	F3G	F4Q	1	Yes	Yes	No	
1205	BNA	F1F?	F2D	F3P	F4Q	1	Yes	Yes	No	
1206	BNA	F1F?	F2D	F3N	F4Q	3	Yes	Yes	No	
1213	BNA	1NA	F2FA	F3H	F4Q	56			No	
1214	BNA	1NA	F2FA	F3G	F4Q	1			No	
1215	BNA	1NA	F2FA	F3P	F4Q	2			No	
1216	BNA	1NA	F2FA	F3N	F4Q	4			No	
1225	BYI	F1B	F2A	F3H	F4Q	9			No	

Table A.2 (continued).

F4UNIV1	BY	F1	F2	F3	F4	Freq.	Dropout at some point	Non-response at some point	Partic. in all 5 waves	Freq. of All 5 partic.
1226	BYI	F1B	F2A	F3G	F4Q	1			No	
1229	BYI	F1B	F2B	F3H	F4Q	1			No	
1231	BYI	F1B	F2B	F3P	F4Q	2			No	
1232	BYI	F1B	F2B	F3N	F4Q	2			No	
1235	BYI	F1B	F2D	F3P	F4Q	2	Yes		No	
1236	BYI	F1B	F2D	F3N	F4Q	3	Yes		No	
1240	BYI	F1B	F2?	F3N	F4Q	1		Yes	No	
1243	BYI	F1D	F2B	F3N	F4Q	1	Yes		No	
1245	BYI	F1D	F2D	F3G	F4Q	2	Yes		No	
1246	BYI	F1D	F2D	F3P	F4Q	6	Yes		No	
1247	BYI	F1D	F2D	F3N	F4Q	6	Yes		No	
1250	BYI	F1D	F2?	F3H	F4Q	1	Yes	Yes	No	
1253	BYI	F1I	F2A	F3H	F4Q	14			No	
1254	BYI	F1I	F2A	F3G	F4Q	1			No	
1255	BYI	F1I	F2A	F3P	F4Q	1			No	
1259	BYI	F1I	F2B	F3H	F4Q	1			No	
1261	BYI	F1I	F2B	F3N	F4Q	2			No	
1266	BYI	F1I	F2D	F3P	F4Q	3			No	
1267	BYI	F1I	F2D	F3N	F4Q	4			No	
1275	BYI	F1X	F2A	F3H	F4Q	1			No	
1276	BYI	F1X	F2D	F3G	F4Q	1			No	
1277	BYI	F1X	F2D	F3N	F4Q	1			No	
1280	BYI	F1?	F2D	F3H	F4Q	1		Yes	No	
Total F4 Respondents						12144	Total Respondents in All 5 waves			11328

Table A.3 Glossary of NELS Variables Used in Model Building

Variable Name	Variable Label	Potenial Values	Recoded Values
BY2XHQ	HISTORY/CIT/ GEOG QUARTILE	Code Freq Percent Label 1 2212 18.2 QUARTILE 1 LOW 2 2682 22.1 QUARTILE 2 3 2848 23.5 QUARTILE 3 4 3178 26.2 QUARTILE 4 HIGH 6 760 6.3 {Legitimate skip/not in wave} 8 63 0.5 {MISSING} 9 401 3.3 {TEST NOT COMP}	Use ordinal scale and set >4 = 0
BY2XMPRO	OVERALL MATH PROFICIENCY	Code Freq Percent Label 0 1563 12.9 BELOW LEVEL 1 1 3844 31.7 LEVEL 1 2 2414 19.9 LEVEL 1 AND 2 3 2372 19.5 ALL 3 LEVELS 6 760 6.3 {Legitimate skip/not in wave} 8 790 6.5 {MISSING} 9 401 3.3 {TEST NOT COMP}	> 2 = 9, 3 = 4, 2 = 3, 1 = 2, 0 = 1, and 9 = 0
BY2XMQ	MATHEMATICS QUARTILE (1=LOW)	Code Freq Percent Label 1 2146 17.7 QUARTILE 1 LOW 2 2589 21.3 QUARTILE 2 3 2891 23.8 QUARTILE 3 4 3339 27.5 QUARTILE 4 HIGH 6 760 6.3 {Legitimate skip/not in wave} 8 18 0.1 {MISSING} 9 401 3.3 {TEST NOT COMP}	Use ordinal scale and set >4 = 0
BY2XRPRO	OVERALL READING PROFICIENCY	Code Freq Percent Label 0 1229 10.1 BELOW LEVEL 1 1 5315 43.8 LEVEL 1 2 4019 33.1 LEVEL 2 6 760 6.3 {Legitimate skip/not in wave} 8 420 3.5 {MISSING} 9 401 3.3 {TEST NOT COMP}	> 2 = 9, 2 = 3, 1 = 2, 0 = 1, and 9 = 0
BY2XRQ	READING QUARTILE (1=LOW)	Code Freq Percent Label 1 2259 18.6 QUARTILE 1 LOW 2 2572 21.2 QUARTILE 2 3 2869 23.6 QUARTILE 3 4 3264 26.9 QUARTILE 4 HIGH 6 760 6.3 {Legitimate skip/not in wave} 8 19 0.2 {MISSING} 9 401 3.3 {TEST NOT COMP}	Use ordinal scale and set >4 = 0
BY2XSPRO	OVERALL SCIENCE PROFICIENCY	Code Freq Percent Label 0 2608 21.5 BELOW LEVEL 1 1 4783 39.4 LEVEL 1 2 2768 22.8 LEVEL 2 6 760 6.3 {Legitimate skip/not in wave} 8 824 6.8 {MISSING} 9 401 3.3 {TEST NOT COMP}	> 2 = 9, 2 = 3, 1 = 2, 0 = 1, and 9 = 0
BY2XSQ	SCIENCE QUARTILE (1=LOW)	Code Freq Percent Label 1 2196 18.1 QUARTILE 1 LOW 2 2699 22.2 QUARTILE 2 3 2896 23.8 QUARTILE 3 4 3162 26.0 QUARTILE 4 HIGH 6 760 6.3 {Legitimate skip/not in wave} 8 30 0.2 {MISSING} 9 401 3.3 {TEST NOT COMP}	Use ordinal scale and set >4 = 0

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYFAMINC	YEARLY FAMILY INCOME	Code Freq Percent Label 1 40 0.3 NONE 2 86 0.7 LESS THAN $1,000 3 147 1.2 $1,000 - $2,999 4 183 1.5 $3,000 - $4,999 5 305 2.5 $5,000 - $7,499 6 352 2.9 $7,500 - $9,999 7 823 6.8 $10,000-$14,999 8 788 6.5 $15,000-$19,999 9 1078 8.9 $20,000-$24,999 10 1967 16.2 $25,000-$34,999 11 2182 18.0 $35,000-$49,999 12 1450 11.9 $50,000-$74,999 13 397 3.3 $75,000-$99,999 14 395 3.3 $100,000-199,999 15 155 1.3 $200,000 OR MORE 98 1036 8.5 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	>15 = 0
BYFCOMP	FAMILY COMPOSITION COMPOSITE	Code Freq Percent Label 1 7882 64.9 MOTHER & FATHER 2 1051 8.7 MOTHER & MALE GUARDN 3 228 1.9 FATHER & FEM GUARD. 4 1584 13.0 MOTHER ONLY 5 248 2.0 FATHER ONLY 6 259 2.1 OTH REL/NON-RELATIVE 98 132 1.1 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Use set of dummy var. BYFCOMPr1, BYFCOMPr2, BYFCOMPr3, BYFCOMPr4, BYFCOMPr5 with 1 or > 6 having all = 0. Ref. case is Mother & Father
BYHMLANG	HOME LANGUAGE BACKGROUND	Code Freq Percent Label 1 417 3.4 NON-ENGLISH ONLY 2 1020 8.4 NON-ENGLISH DOMINANT 3 1078 8.9 ENGLISH DOMINANT 4 8846 72.8 ENGLISH ONLY 8 23 0.2 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and set >4 = 0
BYHOMEWK	NUMBER OF HRS SPENT ON HOMEWORK PER WEEK	Code Freq Percent Label 1 274 2.3 NONE 2 767 6.3 .50 TO 1.99 HOURS 3 2422 19.9 2.00 TO 2.99 HOURS 4 3613 29.8 3.00 TO 5.49 HOURS 5 2061 17.0 5.50 TO 10.49 HOURS 6 501 4.1 10.50 TO 12.99 HOURS 7 777 6.4 13.00 TO 20.99 HOURS 8 332 2.7 21.00 AND UP HOURS 98 637 5.2 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	>8 = 0
BYLEP	LIMITED ENGLISH PROFICIENCY COMPOSITE	Code Freq Percent Label 0 11024 90.8 STUDENT NOT LEP 1 271 2.2 STUDENT IS LEP 8 89 0.7 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use binary recode with > 1 = 0
BYLM	LANGUAGE MINORITY COMPOSITE	Code Freq Percent Label 0 9684 79.7 NOT LANG MINORITY 1 1698 14.0 LANGUAGE MINORITY 8 2 0.0 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use binary recode with > 1 = 0

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYP64B	FAMILY RULE HOW EARLY/LATE CHLD WATCH TV	like: Code Freq Percent Label 1 7159 59.0 YES 2 3290 27.1 NO 6 2 0.0 {MULTIPLE RESPNSE} 8 321 2.6 {MISSING} 9 1372 11.3 {Legitimate skip/not in wave}	> 6 = 0
BYP64C	FAMILY RULE HOW MANY HRS CHILD WATCH TV	Same format as BYP64B	> 6 = 0
BYP64D	FMLY RULE HOW MNY HRS WTCH TV ON SCH DYS	Same format as BYP64B	> 6 = 0
BYP65A	FAMILY RULE ABOUT MAINTAINING GRADE AVG	Code Freq Percent Label 1 7486 61.6 YES 2 2976 24.5 NO 8 310 2.6 {MISSING} 9 1372 11.3 {Legitimate skip/not in wave}	> 2 = 0
BYP68	HOW OFT TALKS TO CHLD RE POST H.S. PLANS	Code Freq Percent Label 1 382 3.1 NOT AT ALL 2 1107 9.1 RARELY 3 5262 43.3 OCCASIONALLY 4 3999 32.9 REGULARLY 7 2 0.0 {REFUSAL} 8 20 0.2 {MISSING} 9 1372 11.3 {Legitimate skip/not in wave}	> 4 = 0
BYP69	HOW OFTEN HELP CHILD WITH HOMEWORK	Code Freq Percent Label 1 3213 26.5 SELDOM OR NEVER 2 2931 24.1 ONCE/TWICE A MONTH 3 3340 27.5 ONCE/TWICE A WEEK 4 1025 8.4 ALMOST EVERY DAY 8 263 2.2 {MISSING} 9 1372 11.3 {Legitimate skip/not in wave}	> 4 = 0
BYP76	HOW FAR IN SCHOOL R EXPECT CHILD TO GO	Code Freq Percent Label 1 39 0.3 LESS THN H.S DIPLOMA 2 21 0.2 GED 3 1205 9.9 HIGH SCHL GRADUATION 4 123 1.0 VOC,TRD,BUS < 1YR 5 416 3.4 VOC,TRD,BUS 1-2 YRS 6 330 2.7 VOC,TRD,2YRS OR MORE 7 536 4.4 < 2YRS OF COLLEGE 8 1020 8.4 2 / MORE YRS COLLEGE 9 500 4.1 FINISH A 2YR PROGRAM 10 4109 33.8 FINISH 4/5 YR PROG 11 1266 10.4 MASTER^S DEGREE 12 1151 9.5 PH.D., M.D., 96 13 0.1 {MULTIPLE RESPNSE} 97 13 0.1 {REFUSAL} 98 30 0.2 {MISSING} 99 1372 11.3 {Legitimate skip/not in wave}	> 12 = 0
BYPARMAR	PARENTS^ MARITAL STATUS	Code Freq Percent Label 1 1085 8.9 DIVORCED 2 255 2.1 WIDOWED 3 332 2.7 SEPARATED 4 212 1.7 NEVER MARRIED 5 141 1.2 MARRIAGE-LIKE RELAT 6 8493 69.9 MARRIED 98 866 7.1 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Use a set of dummy variables with all = 0 if Married (ref. case)

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYRISK	BY RISK OF DROPPING OUT OF SCHOOL	Code Freq Percent Label 0 6583 54.2 NO RISK FACTORS 1 2833 23.3 ONE RISK FACTOR 2 1348 11.1 TWO RISK FACTORS 3 482 4.0 THREE RISK FACTORS 4 121 1.0 FOUR RISK FACTORS 5 17 0.1 FIVE RISK FACTORS 99 760 6.3 {Legitimate skip/not in wave}	> 6 = 0
BYS14	SECTOR OF HIGH SCHOOL R PLANS TO ATTEND (pub/priv rel/priv non-rel)	Code Freq Percent Label 1 9550 78.6 PUBLIC 2 1031 8.5 PRIVATE RELIGIOUS 3 466 3.8 PRVT NON-RELIGIOUS 4 185 1.5 DON^T KNOW 8 152 1.3 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Create 2 dummy variables for Private Religious (yes/no), Private Nonreligous (yes/no)
BYS17	R SPEAK ANY LANG OTH THN ENGLISH BFR SCH	Code Freq Percent Label 1 1764 14.5 YES 2 9562 78.7 NO 8 58 0.5 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Make binary with > 1 = 0
BYS20	LANGUAGE R USUALLY SPEAKS NOW	Code Freq Percent Label 1 10867 89.5 ENGLISH 2 146 1.2 SPANISH 3 16 0.1 CHINESE 4 1 0.0 JAPANESE 5 5 0.0 KOREAN 6 10 0.1 FILIPINO LANGUAGE 7 3 0.0 ITALIAN 8 14 0.1 FRENCH 9 4 0.0 GERMAN 10 2 0.0 GREEK 11 1 0.0 POLISH 12 1 0.0 PORTUGUESE 13 32 0.3 OTHER (SPECIFY) 96 103 0.8 {MULTIPLE RESPNSE} 98 179 1.5 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Make binary with 1 or > 96 = 0 (English) and all else = 1
BYS32	NUMBER OF SIBLINGS R HAS	Code Freq Percent Label 0 705 5.8 NONE 1 3673 30.2 ONE 2 3057 25.2 TWO 3 1703 14.0 THREE 4 905 7.5 FOUR 5 491 4.0 FIVE 6 788 6.5 SIX OR MORE 96 10 0.1 {MULTIPLE RESPNSE} 98 52 0.4 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and make > 6 = 0
BYS34A	FATHER^S HIGHEST LEVEL OF EDUCATION	Code Freq Percent Label 1 1680 13.8 NOT FINISH H.S. 2 3078 25.3 GRADUATED H.S. 3 1106 9.1 JUNIOR COLLEGE 4 799 6.6 COLLEGE LT 4 YRS 5 1533 12.6 GRADUATED COLLEGE 6 872 7.2 MASTER^S DEGREE 7 619 5.1 PH.D., M.D., ETC. 8 1528 12.6 DON^T KNOW 97 41 0.3 {REFUSAL} 98 128 1.1 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Make binary with <4 or >7 = 0 and 4-7 = 1.

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYS34B	MOTHER^S HIGHEST LEVEL OF EDUCATION	Code Freq Percent Label 1 1686 13.9 NOT FINISH H.S. 2 3698 30.5 GRADUATED H.S. 3 1228 10.1 JUNIOR COLLEGE 4 908 7.5 COLLEGE LT 4 YRS 5 1520 12.5 GRADUATED COLLEGE 6 788 6.5 MASTER^S DEGREE 7 239 2.0 PH.D., M.D., ETC. 8 1236 10.2 DON^T KNOW 97 16 0.1 {REFUSAL} 98 65 0.5 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Make binary with <4 or >7 = 0 and 4-7 = 1.
BYS41	TIME SPENT AFTER SCHL WTH NO ADULT PRSNT	Code Freq Percent Label 0 1553 12.8 NONE 1 3744 30.8 LESS THAN 1 HOUR 2 3104 25.6 1-2 HOURS 3 1406 11.6 2-3 HOURS 4 1392 11.5 MORE THAN 3 HRS 6 5 0.0 {MULTIPLE RESPNSE} 8 180 1.5 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and make > 6 = 0
BYS42A	NO. OF HOURS R WATCHES TV ON WEEKDAYS	Code Freq Percent Label 0 322 2.7 DON^T WATCH TV 1 885 7.3 LT 1 HOUR A DAY 2 2370 19.5 1-2 HOURS 3 2456 20.2 2-3 HOURS 4 1815 14.9 3-4 HOURS 5 1215 10.0 4-5 HOURS 6 1259 10.4 OVER 5 HRS A DAY 96 774 6.4 {MULTIPLE RESPNSE} 98 288 2.4 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and make > 6 = 0
BYS42B	NO. OF HOURS R WATCHES TV ON WEEKENDS	Same format as BYS42A	Use ordinal scale and make > 6 = 0
BYS43	NO. OF CIGARETTES R SMOKES PER DAY	Code Freq Percent Label 0 10587 87.2 I DON^T SMOKE 1 385 3.2 1-5 CIGARETTES 2 124 1.0 ABOUT 1/2 PACK 3 65 0.5 MT 1/2,LT 2 PACKS 4 24 0.2 2 PACKS OR MORE 8 199 1.6 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and make > 4 = 0
BYS45	HOW FAR IN SCH DO YOU THINK YOU WILL GET	Code Freq Percent Label 1 136 1.1 WON^T FINISH H.S 2 1024 8.4 WILL FINISH H.S 3 972 8.0 VOC,TRD,BUS AFTR H.S 4 1467 12.1 WILL ATTEND COLLEGE 5 4848 39.9 WILL FINISH COLLEGE 6 2850 23.5 HIGHER SCH AFTR COLL 98 87 0.7 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	> 6 = 0
BYS46	HOW SURE THAT YOU WILL GRADUATE FROM H.S	Code Freq Percent Label 1 9504 78.3 VERY SURE WILL 2 1619 13.3 PROBABLY WILL 3 100 0.8 PROBABLY WON^T 4 56 0.5 VERY SURE WON^T 6 3 0.0 {MULTIPLE RESPNSE} 8 102 0.8 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	> 4 = 0

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYS47	HOW SURE R IS TO GO FURTHER THAN H.S.	Code Freq Percent Label 1 7004 57.7 VERY SURE WILL 2 3247 26.7 PROBABLY WILL 3 664 5.5 PROBABLY WON^T 4 253 2.1 VERY SURE WON^T 8 160 1.3 {MISSING} 9 816 6.7 {Legitimate skip/not in wave}	> 4 = 0
BYS48A	HOW FAR IN SCHL R^S FATHER WANTS R TO GO	Code Freq Percent Label 1 78 0.6 LESS THAN HIGH SCHL 2 519 4.3 GRADUATE HIGH SCHOOL 3 583 4.8 VOC,TRD,BUS AFTR H.S 4 986 8.1 ATTEND COLLEGE 5 4711 38.8 GRADUATE FRM COLLEGE 6 2750 22.6 HIGHER SCH AFTR COLL 7 908 7.5 DON^T KNOW 98 849 7.0 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	> 6 = 0
BYS48B	HOW FAR IN SCHL R^S MOTHER WANTS R TO GO	Same format as BYS48A	> 6 = 0
BYS53	NO. OF HOURS R WORKS FOR PAY PER WEEK	Code Freq Percent Label 0 3538 29.1 NONE 1 3987 32.8 UP TO 4 HOURS 2 2262 18.6 5-10 HOURS 3 841 6.9 11-20 HOURS 4 595 4.9 21 OR MORE HOURS 6 3 0.0 {MULTIPLE RESPNSE} 8 158 1.3 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	Use ordinal scale and make > 4 = 0
BYS60A	R^S ABILITY GROUP FOR MATHEMATICS	Code Freq Percent Label 1 3629 29.9 HIGH 2 4523 37.2 MIDDLE 3 750 6.2 LOW 4 1676 13.8 AREN^T GROUPED 5 545 4.5 I DON^T KNOW 6 9 0.1 {MULTIPLE RESPNSE} 8 252 2.1 {MISSING} 9 760 6.3 {Legitimate skip/not in wave}	> 4 = 0, 1 goes to 3, 3 goes to 1.
BYS60B	R^S ABILITY GROUP FOR SCIENCE	Same format as BYS60A	> 4 = 0, 1 goes to 3, 3 goes to 1.
BYS60C	R^S ABILITY GROUP FOR ENGLISH	Same format as BYS60A	> 4 = 0, 1 goes to 3, 3 goes to 1.
BYS81A	ENGLISH GRADES FROM GRADE 6 UNTIL NOW	Code Freq Percent Label 1 3879 31.9 MOSTLY AS 2 4352 35.8 MOSTLY BS 3 2244 18.5 MOSTLY CS 4 473 3.9 MOSTLY DS 5 166 1.4 MOSTLY BELOW D 6 45 0.4 NOT GRADED 96 139 1.1 {MULTIPLE RESPNSE} 97 7 0.1 {REFUSAL} 98 79 0.7 {MISSING} 99 760 6.3 {Legitimate skip/not in wave}	>6 = 0, 1 goes to 5, 2 goes to 4, 4 goes to 2, 5 goes to 1.

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYS81B	MATH GRADES FROM GRADE 6 UNTIL NOW	Same format as BYS81A	>6 = 0, 1 goes to 5, 2 goes to 4, 4 goes to 2, 5 goes to 1.
BYS81C	SCI GRADES FROM GRADE 6 UNTIL NOW	Same format as BYS81A	>6 = 0, 1 goes to 5, 2 goes to 4, 4 goes to 2, 5 goes to 1.
BYSC13E	% OF WHITE NON-HISPANIC 8TH GRADERS	Code Freq Percent Label 0 3525 29.0 {zero} {cont} 7656 63.0 {1-100;15.59/22.30} 996 12 0.1 {DON^T KNOW} 997 7 0.1 {REFUSAL} 998 37 0.3 {MISSING} 999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >100 = 0
BYSC14	% OF 8TH GRADERS IN SINGLE PARENT FAMILY	Code Freq Percent Label 1 94 0.8 NONE 2 5977 49.2 1% - 25 3 3718 30.6 26% - 50 4 838 6.9 51% - 75 5 172 1.4 76% - 99 7 401 3.3 CANNOT ESTIMATE 98 37 0.3 {MISSING} 99 907 7.5 {Legitimate skip/not in wave}	Use ordinal scale and set > 5 = 0
BYSC15	% OF 8TH GRADERS LIMITED ENGL PROFICIENT	Code Freq Percent Label 1 10197 84.0 10% OR LESS 2 563 4.6 11 - 20 3 227 1.9 21 - 30 4 109 0.9 31 - 40 5 25 0.2 41 - 50 6 14 0.1 51 - 60 7 10 0.1 61 - 70 8 14 0.1 71 - 80 9 49 0.4 81% OR MORE 98 29 0.2 {MISSING} 99 907 7.5 {Legitimate skip/not in wave}	Use ordinal scale and set >9 = 0
BYSC16A	NUMBER OF STUDENTS IN FREE LUNCH PROGRAM	Code Freq Percent Label 0 1585 13.1 {zero} {cont} 9599 79.0 {1-3230;182.33/253.24} 9996 18 0.1 {DON^T KNOW} 9998 35 0.3 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0
BYSC16B	NUMBER OF STUDENTS IN REMEDIAL READING	Code Freq Percent Label 0 2236 18.4 {zero} {cont} 8985 74.0 {1-2700;87.30/158.89} 9998 16 0.1 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0
BYSC16C	NUMBER OF STUDENTS IN REMEDIAL MATH	Code Freq Percent Label 0 3460 28.5 {zero} {cont} 7761 63.9 {1-2700;76.32/141.16} 9998 16 0.1 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0
BYSC16D	NUMBER OF STUDENTS IN BILINGUAL EDUCATN	Code Freq Percent Label 0 9716 80.0 {zero} {cont} 1482 12.2 {1-1100;85.08/157.66} 9998 39 0.3 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0
BYSC16E	NUMBER OF STUDNTS IN ENGLISH AS 2ND LANG	Code Freq Percent Label 0 7804 64.3 {zero} {cont} 3393 27.9 {1-0730;40.11/82.18} 9998 40 0.3 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
BYSC16F	NUMBER OF STUDENTS IN SPECIAL ED	Code Freq Percent Label 0 1910 15.7 {zero} {cont} 9298 76.6 {1-0265;50.78/39.65} 9998 29 0.2 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >4000 = 0
BYSC16G	NUMBER OF STUDENTS IN GIFTED,TALENTED ED	Code Freq Percent Label 0 3103 25.6 {zero} {cont} 7801 64.2 {1-0900;63.55/81.49} 9998 333 2.7 {MISSING} 9999 907 7.5 {Legitimate skip/not in wave}	Use as cont. var and set >1000 = 0
BYSC19	BASE SALARY FOR BEGINNING TEACHER W/ BA	Code Freq Percent Label {cont} 11083 91.3 {5500-25428;17526.37/2926.15} 99996 7 0.1 {DON^T KNOW} 99997 20 0.2 {REFUSAL} 99998 127 1.0 {MISSING} 99999 907 7.5 {Legitimate skip/not in wave	Use as cont. var and set >50000 = 0
BYSC29	MIN. GPA REQUIRD TO PARTIC IN ACTIVITIES	Code Freq Percent Label 1 8328 68.6 YES 2 2880 23.7 NO 8 29 0.2 {MISSING} 9 907 7.5 {Legitimate skip/not in wave}	Use binary recode with > 1 = 0
F2RACTC	ACT (COMPOSITE)	Code Freq Percent Label 4 1 0.0 {04} 10 3 0.0 {10} 11 4 0.0 {11} 12 18 0.1 {12} 13 47 0.4 {13} 14 82 0.7 {14} 15 122 1.0 {15} 16 167 1.4 {16} 17 211 1.7 {17} 18 246 2.0 {18} 19 263 2.2 {19} 20 288 2.4 {20} 21 233 1.9 {21} 22 218 1.8 {22} 23 226 1.9 {23} 24 173 1.4 {24} 25 166 1.4 {25} 26 141 1.2 {26} 27 102 0.8 {27} 28 120 1.0 {28} 29 74 0.6 {29} 30 58 0.5 {30} 31 40 0.3 {31} 32 22 0.2 {32} 33 20 0.2 {33} 34 4 0.0 {34} 35 2 0.0 {35} 98 7259 59.8 {MISSING DATA} 99 1834 15.1 {Legitimate skip/not in wave}	> 36 = 0
F2RACTE	ACT (ENGLISH SCORE)	Same format as F2RACTC	> 36 = 0
F2RACTM	ACT (MATH)	Same format as F2RACTC	> 36 = 0
F2RACTR	ACT (READING)	Same format as F2RACTC	> 36 = 0
F2RACTS	ACT (SCIENCE REASONING)	Same format as F2RACTC	> 36 = 0

Table A.3 (continued).

Variable Name	Variable Label	Potenial Values	Recoded Values
F2RSATM	SCHOLASTIC APTITUDE TEST (MATHEMATICS)	Code Freq Percent Label {cont} 3499 28.8 {200-800;502.09/121.80} 998 6811 56.1 {MISSING DATA} 999 1834 15.1 {Legitimate skip/not in wave}	>800 = 0
F2RSATV	SCHOLASTIC APTITUDE TEST (VERBAL)	Code Freq Percent Label {cont} 3500 28.8 {200-780;446.48/111.26} 998 6810 56.1 {MISSING DATA} 999 1834 15.1 {Legitimate skip/not in wave}	>800 = 0
SATM	SAT MATH SCORE W/CORRECTIONS	Code Freq Percent Label {cont} 3,547 29.21 { 200 - 800; 502.1 / 121.66 } -1 6,592 54.28 No claim, no score -9 2,005 16.51 Claim, no score	< 0 = 0
SATV	SAT VERBAL SCORE W/CORRECTIONS	Code Freq Percent Label {cont} 3,548 29.22 { 200 - 780; 447.3 / 111.14 } -1 6,591 54.27 No claim, no scores -9 2,005 16.51 Claim, no scores	< 0 = 0
F4RACE2	New definition of race-primary choice	Code Freq Percent Label 1 131 1.1 American Indian or Alaska Native 2 712 5.9 Asian or Pacific Islander 3 1120 9.2 Black, not Hispanic 4 8203 67.5 White, not Hispanic 5 1687 13.9 Hispanic or Latino -9 291 2.4 {Missing}	Set of dummy var. F4RACE2rAI = 1 for Amer. Ind., F4RACE2rAs = 1 for Asian/Pac. Isl., F4RACE2rBl = 1 for Black, F4RACE2rHi = 1 for Hispanic/Latino, all = 0 for White.
F4SEX	Gender	Code Freq Percent Label 1 5782 47.6 Male 2 6362 52.4 Female	Recode Male = 0 as reference, Female = 1

Table A.4 Summary of NELS Variables Used in Record Classification

Variable Name	Variable Description
AAMJR	ASSOCIATE^S DEGREE AGGREGATE MAJOR
ACTTEST	ACT COMPOSITE SCORE: CORRECTED/EXPANDED
AGREEDEG	STUDENT V. TRANSCRIPT ON HIGHEST DEGREE
ALLHDEG	FINAL EDUCATIONAL STATUS - DERIVED
ASSOTIME	TRUE ELAPSED TIME TO ASSOCIATE^S DEGREE
BACHTME	TRUE TOTAL ELAPSED TIME TO BACHELOR^S
BACHTTD	STANDARD ACCOUNT OF TIME TO BACHELOR^S
BALIKELY	BACH DEGREE LIKELY BY DECEMBER, 2001
BAMJR	BACHELOR^S DEGREE AGGREGATE MAJOR
CONSDEG	CONSOLIDATED HIGHEST DEGREE
CREDRET	CREDIT-RETENTION ACCOUNT OF ATTAINMENT
DEG1	NO DEGREE ON AT LEAST 1 TRANSCRIPT
DEG2	CERTIFICATE ON AT LEAST 1 TRANSCRIPT
DEG3	ASSOCIATE^S DEG ON AT LEAST 1 TRANSCRIPT
DEG4	BACHELOR^S DEG ON AT LEAST 1 TRANSCRIPT
DEG5	POST-BACC COURSE WORK ON TRANSCRIPT
DEG6	INCOMPLETE GRAD DEG ON AT LEAST 1 TRANS
DEG7	MASTERS DEGREE ON AT LEAST 1 TRANSCRIPT
DEG8	1ST PROFESS DEG ON AT LEAST 1 TRANSCRIPT
DEG9	PHD DEGREE ON TRANSCRIPT
DEGDAT2	DATE OF 1ST CERTIFICATE EARNED
DEGDAT3	DATE OF 1ST ASSOCIATE^S DEGREE EARNED
DEGDAT4	DATE OF 1ST BACHELOR^S DEGREE EARNED
DEGDAT5	DATE OF 1ST POST-BACC CERTIFICATE EARNED
DEGDAT7	DATE OF 1ST MASTER^S DEGREE EARNED
DEGDAT8	DATE OF 1ST PROFESSIONAL DEGREE
DELAY	TIME BETWEEN HS GRAD DATE AND REFDATE
DOUBBACH	DOUBLE OR 2 BACHELORS DEGREES EARNED
ENDDAT	LAST DATE ENROLLED AS AN UNDERGRADUATE
GPA	UNDERGRADUATE GRADE POINT AVERAGE
HDEG	HIGHEST DEGREE EARNED: TRANSCRPT ACCOUNT
HSGPAV	HIGH SCHOOL GRADE POINT AVERAGE
IN2000SC	STUDENT CLAIM TO ENROLLMENT IN 2000
INSCHOOL	STUDENT PSE STATUS IN 2000
MAJCOD1	DETAILED FIELD FOR NO DEGREE
MAJCOD2	DETAILED FIELD FOR CERTIFICATE
MAJCOD3	DETAILED FIELD FOR ASSOCIATE^S DEGREE
MAJCOD4	DETAILED FIELD FOR BACHELOR^S DEGREE
MAJCOD5	DETAILED FIELD FOR POST-BACC COURSE WORK
MAJCOD6	DETAILED FIELD FOR INCOMPLETE GRAD DEG
MAJCOD7	DETAILED FIELD FOR MASTER^S DEGREE
MAJCOD8	DETAILED FIELD FOR 1ST PROFESS DEGREE
MAJCOD9	DETAILED FIELD FOR PHD
MATHINT	DEGREE OF INTEREST IN MATH, 1992
PETSGTYP	TYPE OF HIGH SCHOOL CREDENTIAL: REVISED
PETSHSDT	DATE OF HIGH SCHOOL GRADUATION: REVISED
PETSID	ID FOR STUDENTS IN THE PETS FILES
PSBEG	OSTENSIBLE DATE OF PSE ENTRY

Table A.4 (continued).

Variable Name	Variable Description
PSEND	LAST MONTH OF ENROLLMENT IN PSE
QUALDAT4	LAST DATE ENROLLED FOR BACHELORS DEGREE
REFDATE	TRUE FIRST DATE OF PSE ATTENDANCE
STUHDEG	STUDENT ACCOUNT OF HIGHEST DEGREE EARNED
F1D7MNTH	MONTH R LAST ATTENDED SCHOOL
F1D7YEAR	YEAR R LAST ATTENDED SCHOOL
F2D6M	MONTH R LAST ATTENDED SCHOOL
F2D6Y	YEAR R LAST ATTENDED SCHOOL
F4EDGR1	Degree/certificate earned-1
F4EDGR2	Degree/certificate earned-2
F4EDMJ1	Double major indicator-degree 1
F4EDMJ2	Double major indicator-degree 2
F4EMJ1D	Major/field of study code-1
F4EMJ2D	Major/field of study code-2

APPENDIX B

CLASSIFICATION OF COLLEGE MAJORS

Table B.1 Classification of College Majors by Seymour & Hewitt[100]

Field	College Major	Discipline
Sci./Math/Engr.	Biological Sciences	Biology (general); Biochemistry/Biophysics; Botany; Marine (life) Science; Microbiology/Bacteriology; Zoology
Sci./Math/Engr.	Physical Sciences	Astronomy; Atmospheric Science; Chemistry; Earth Science; Marine Science; Physics; Other physical science
Sci./Math/Engr.	Engineering	Aeronautical or Astronautical; Civil; Chemical; Electrical or Electronic; Industrial; Mechanical, Other
Sci./Math/Engr.	Mathematics/Statistics	Mathematics; Statistics
Sci./Math/Engr.	Math (only)	Mathematics
Sci./Math/Engr.	Agriculture	Agriculture; Forestry
Humanities/Soc. Sci.	History/Political Sci.	History; Political Science
Humanities/Soc. Sci.	Social Sciences	Anthropology; Economics; Ethnic Studies; Geography; Psychology; Social Work; Sociology; Women's Studies; Other social sciences
Humanities/Soc. Sci.	Fine Arts	Art, Fine and Applied; Music; Speech; Architecture/Urban Planning
Humanities/Soc. Sci.	English	English (language or literature)
Humanities/Soc. Sci.	Other Humanities	Languages (except English); Philosophy; Theater or drama; Theology or Religion; Other
Other	Health Professions	Nursing; Pharmacy; Pre-medicine; Pre-dentistry; Pre-veterinary; Clinical Therapies (Physical, Occupational, Speech)

Table B.1 (continued).

Field	College Major	Discipline
Other	Computer Science/Technical	Computer Science; Data Processing or Computer Programming; Communications; Drafting or Design; Mechanics; Electronics, Other technical
Other	Business	Accounting; Business Administration; Finance; Marketing; Management; Secretarial Studies; Other business
Other	Education	Business; Elementary; Music or Art; Physical Education or Recreation; Secondary; Special
Other	Other Non-technical	Journalism; Home Economics; Library/Archival Science; Law Enforcement; Military Science; Other

Table B.2 Majors Classified as "SME" by Frederick Smythe by Dataset [101]

Dataset	Majors
"College & Beyond"	Biological Sciences
	Pre-Med
	Dentistry
	Computers
	Material Sciences
	Mechanical Engineering
	Engineering
	Computer and Information Sciences
	Mathematics
	Astronomy, Atmospheric Sciences
	Chemistry
	Geology
	Geological Sciences
	Physics
	Other Physical Sciences
CIRP	Biology (general)
	Biochemistry or Biophysics
	Botany
	Marine (Life) Sciences
	Microbiology or Bacteriology
	Zoology
	Other Biological Science
	Astronomy
	Atmospheric (incl. Meteorology)
	Chemistry
	Earth Science
	Marine (incl. Oceanography)
	Mathematics
	Physics
	Statistics
	Other Physical Science
	Aeronautical or Astronautical Engineering
	Civil Engineering
	Chemical Engineering
	Electrical or Electronic Engineering
	Industrial Engineering
	Mechanical Engineering
	Other Engineering
	Professional: Pre-Med, Pre-Vet., Pre-Dental
	Computer Science
	Data Processing
	Computer Programming

Table B.3 Classification of College Majors by NSF [102]

Field	College Major	Discipline
Science	Agricultural sciences	Animal breeding/genetics
Science	Agricultural sciences	Animal husbandry
Science	Agricultural sciences	Animal nutrition
Science	Agricultural sciences	Dairy science
Science	Agricultural sciences	Poultry science
Science	Agricultural sciences	Animal sciences, other
Science	Agricultural sciences	Agronomy/crop science
Science	Agricultural sciences	Agricultural/horticultural plant breeding/genetics
Science	Agricultural sciences	Plant pathology/phytopathology
Science	Agricultural sciences	Plant protection/pest management
Science	Agricultural sciences	Plant sciences, other
Science	Agricultural sciences	Food sciences
Science	Agricultural sciences	Food distribution
Science	Agricultural sciences	Food science
Science	Agricultural sciences	Food sciences/technology, other
Science	Agricultural sciences	Soil sciences
Science	Agricultural sciences	Soil chemistry/microbiology
Science	Agricultural sciences	Soil sciences, other
Science	Agricultural sciences	Horticulture science
Science	Agricultural sciences	Fish and wildlife
Science	Agricultural sciences	Fishing and fisheries sciences/management
Science	Agricultural sciences	Wildlife management
Science	Agricultural sciences	Forestry science
Science	Agricultural sciences	Forest sciences/biology
Science	Agricultural sciences	Forest engineering
Science	Agricultural sciences	Forest management/resources
Science	Agricultural sciences	Wood science and pulp/paper technology
Science	Agricultural sciences	Natural resources conservation, other
Science	Agricultural sciences	Forestry and related sciences, other
Science	Agricultural sciences	Wildlife/range management
Science	Agricultural sciences	Environmental science
Science	Agricultural sciences	Agricultural science, general
Science	Agricultural sciences	Agricultural science, other
Science	Agricultural sciences	Environmental science
Science	Biological Sciences	Biochemistry
Science	Biological Sciences	Biomedical sciences
Science	Biological Sciences	Biophysics
Science	Biological Sciences	Biotechnology research

Table B.3 (continued).

Field	College Major	Discipline
Science	Biological Sciences	Bacteriology
Science	Biological Sciences	Plant genetics
Science	Biological Sciences	Plant pathology/phytopathology
Science	Biological Sciences	Plant physiology
Science	Biological Sciences	Botany/plant biology
Science	Biological Sciences	Anatomy
Science	Biological Sciences	Biometrics/biostatistics
Science	Biological Sciences	Cell/cellular biology and histology
Science	Biological Sciences	Ecology
Science	Biological Sciences	Hydrobiology
Science	Biological Sciences	Developmental biology/embryology
Science	Biological Sciences	Endocrinology
Science	Biological Sciences	Entomology
Science	Biological Sciences	Biological immunology
Science	Biological Sciences	Molecular biology
Science	Biological Sciences	Microbiology/bacteriology
Science	Biological Sciences	Microbiology
Science	Biological Sciences	Neuroscience
Science	Biological Sciences	Nutritional sciences
Science	Biological Sciences	Parasitology
Science	Biological Sciences	Toxicology
Science	Biological Sciences	Genetics, human/animal
Science	Biological Sciences	Genetics
Science	Biological Sciences	Pathology, human/animal
Science	Biological Sciences	Pharmacology, human/animal
Science	Biological Sciences	Physiology, human/animal
Science	Biological Sciences	Animal/plant physiology
Science	Biological Sciences	Zoology, other
Science	Biological Sciences	Biology/biological sciences, general
Science	Biological Sciences	Biology/biomedical sciences, other
Science	Computer sciences	Computer/information sciences, general
Science	Computer sciences	Computer programming
Science	Computer sciences	Data processing technology/technician
Science	Computer sciences	Information sciences/systems
Science	Computer sciences	Computer systems analysis
Science	Computer sciences	Computer science
Science	Computer sciences	Web page design, computer graphics, database management

Table B.3 (continued).

Field	College Major	Discipline
Science	Computer sciences	Computer systems networking and telecommunications
Science	Computer sciences	System administration, networking, management
Science	Computer sciences	Computer/information sciences, other
Science	Atmospheric sciences	Atmospheric chemistry/climatology
Science	Atmospheric sciences	Atmospheric physics/dynamics
Science	Atmospheric sciences	Meteorology
Science	Atmospheric sciences	Atmospheric science/meteorology, general
Science	Atmospheric sciences	Atmospheric science/meteorology, other
Science	Earth sciences	Geology
Science	Earth sciences	Geochemistry
Science	Earth sciences	Geophysics/seismology
Science	Earth sciences	Geophysics (solid earth)
Science	Earth sciences	Paleontology
Science	Earth sciences	Mineralogy/petrology
Science	Earth sciences	Mineralogy, petrology, geochemistry
Science	Earth sciences	Stratigraphy/sedimentation
Science	Earth sciences	Geomorphology/glacial geology
Science	Earth sciences	Applied geology
Science	Earth sciences	Applied geology/geological engineering
Science	Earth sciences	Geological/earth sciences, general
Science	Earth sciences	Geological/earth sciences, other
Science	Earth sciences	Hydrology/water resources
Science	Ocean Sciences	Oceanography
Science	Ocean Sciences	Marine sciences
Science	Ocean Sciences	Ocean/marine sciences, other
Science	Mathematics	Applied mathematics
Science	Mathematics	Algebra
Science	Mathematics	Analysis/functional analysis
Science	Mathematics	Geometry/geometric analysis
Science	Mathematics	Logic
Science	Mathematics	Number theory
Science	Mathematics	Mathematical statistics
Science	Mathematics	Topology
Science	Mathematics	Computing theory/practice
Science	Mathematics	Operations research
Science	Mathematics	Mathematics/statistics, general
Science	Mathematics	Mathematics/statistics, other
Science	Physical sciences: Astronomy	Astronomy

Table B.3 (continued).

Field	College Major	Discipline
Science	Physical sciences: Astronomy	Astrophysics
Science	Physical sciences: Astronomy	Astronomy/astrophysics
Science	Physical sciences: Chemistry	Analytical chemistry
Science	Physical sciences: Chemistry	Agricultural/food chemistry
Science	Physical sciences: Chemistry	Inorganic chemistry
Science	Physical sciences: Chemistry	Nuclear chemistry
Science	Physical sciences: Chemistry	Organic chemistry
Science	Physical sciences: Chemistry	Medicinal/pharmaceutical chemistry
Science	Physical sciences: Chemistry	Physical chemistry
Science	Physical sciences: Chemistry	Polymer chemistry
Science	Physical sciences: Chemistry	Theoretical chemistry
Science	Physical sciences: Chemistry	Chemistry, general
Science	Physical sciences: Chemistry	Chemistry, other
Science	Physical sciences: Physics	Acoustics
Science	Physical sciences: Physics	Chemical and atomic/molecular physics
Science	Physical sciences: Physics	Electron physics
Science	Physical sciences: Physics	Electromagnetism
Science	Physical sciences: Physics	Elementary particle physics
Science	Physical sciences: Physics	Biophysics
Science	Physical sciences: Physics	Fluids
Science	Physical sciences: Physics	Mechanics
Science	Physical sciences: Physics	Nuclear physics
Science	Physical sciences: Physics	Optics/photonics
Science	Physical sciences: Physics	Plasma/high-temperature physics
Science	Physical sciences: Physics	Polymer physics
Science	Physical sciences: Physics	Thermal physics
Science	Physical sciences: Physics	Solid state/low-temperature physics
Science	Physical sciences: Physics	Theoretical physics
Science	Physical sciences: Physics	Applied physics
Science	Physical sciences: Physics	Physics, general
Science	Physical sciences: Physics	Physics, other
Science	Physical sciences: Other	Physical sciences, other
Science	Psychology	Clinical psychology
Science	Psychology	Cognitive psychology/psycholinguistics
Science	Psychology	Comparative psychology
Science	Psychology	Counseling psychology
Science	Psychology	Developmental/child psychology
Science	Psychology	Human development/family studies

Table B.3 (continued).

Field	College Major	Discipline
Science	Psychology	Experimental psychology
Science	Psychology	Experimental, comparative, physiological psychology
Science	Psychology	Educational psychology
Science	Psychology	Human engineering
Science	Psychology	Family psychology
Science	Psychology	Industrial/organizational psychology
Science	Psychology	Personality psychology
Science	Psychology	Physiological psychology/psychobiology
Science	Psychology	Psychometrics
Science	Psychology	Quantitative psychology
Science	Psychology	School psychology
Science	Psychology	Social psychology
Science	Psychology	Psychology, general
Science	Psychology	Psychology, other
Science	Soc. Sciences: Economics	Agricultural economics
Science	Soc. Sciences: Economics	Economics
Science	Soc. Sciences: Economics	Econometrics
Science	Soc. Sciences: Political Science	International relations/affairs
Science	Soc. Sciences: Political Science	Political science/government
Science	Soc. Sciences: Political Science	Political science/public administration
Science	Soc. Sciences: Political Science	Public policy analysis
Science	Soc. Sciences: Political Science	Public administration
Science	Soc. Sciences: Sociology	Demography/population studies
Science	Soc. Sciences: Sociology	Sociology
Science	Soc. Sciences: Other	Anthropology
Science	Soc. Sciences: Other	Area studies
Science	Soc. Sciences: Other	Criminology
Science	Soc. Sciences: Other	Geography
Science	Soc. Sciences: Other	Statistics
Science	Soc. Sciences: Other	Urban affairs/studies
Science	Soc. Sciences: Other	Social sciences, general
Science	Soc. Sciences: Other	Social sciences, other
Science	Soc. Sciences: Other	History/philosophy of science and technology
Science	Soc. Sciences: Other	Linguistics
Science	Soc. Sciences: Other	American studies
Science	Soc. Sciences: Other	Archeology
Engineering	Aeronautical/astronautical engineering	Aerospace, aeronautical, astronautical engineering

Table B.3 (continued).

Field	College Major	Discipline
Engineering	Chemical engineering	Chemical engineering
Engineering	Chemical engineering	Petroleum engineering
Engineering	Chemical engineering	Polymer/plastics engineering
Engineering	Chemical engineering	Fuel technology/petroleum engineering
Engineering	Civil engineering	Civil engineering
Engineering	Civil engineering	Environmental health engineering
Engineering	Electrical engineering	Communications engineering
Engineering	Electrical engineering	Computer engineering
Engineering	Electrical engineering	Electrical engineering
Engineering	Electrical engineering	Electronics engineering
Engineering	Electrical engineering	Electrical, electronics, communications engineering
Engineering	Mechanical engineering	Engineering mechanics
Engineering	Mechanical engineering	Mechanical engineering
Engineering	Materials/metallurgical engineering	Ceramic sciences
Engineering	Materials/metallurgical engineering	Materials science
Engineering	Materials/metallurgical engineering	Metallurgical engineering
Engineering	Other engineering	Agricultural engineering
Engineering	Other engineering	Bioengineering/biomedical engineering
Engineering	Other engineering	Engineering physics
Engineering	Other engineering	Engineering science
Engineering	Other engineering	Mining/mineral engineering
Engineering	Other engineering	Naval architecture/marine engineering
Engineering	Other engineering	Nuclear engineering
Engineering	Other engineering	Ocean engineering
Engineering	Other engineering	Operations research
Engineering	Other engineering	Systems engineering
Engineering	Other engineering	Textile engineering
Engineering	Other engineering	Engineering, general
Engineering	Other engineering	Engineering, other
Education	Education	Curriculum/instructions
Education	Education	Education administration/supervision
Education	Education	Education leadership
Education	Education	Education/instructional media design
Education	Education	Education statistics/research methods
Education	Education	Education assessment/testing/measures
Education	Education	Educational psychology

Table B.3 (continued).

Field	College Major	Discipline
Education	Education	School psychology
Education	Education	Social/philosophical foundations of education
Education	Education	Special education
Education	Education	Counseling education/counseling and guidance services
Education	Education	Education evaluation/research
Education	Education	Pre-elementary/early childhood teacher education
Education	Education	Elementary teacher education
Education	Education	Junior high teacher education
Education	Education	Secondary teacher education
Education	Education	Adult/continuing teacher education
Education	Education	Art education
Education	Education	Business education
Education	Education	English education
Education	Education	Foreign languages education
Education	Education	Physical education, health, recreation
Education	Education	Home economics education
Education	Education	Music education
Education	Education	Physical education/coaching
Education	Education	Reading education
Education	Education	Speech education
Education	Education	Trade/industrial education
Education	Education	Teacher education, specific academic/vocational
Education	Education	programs, other
Education	Education	Education, general
Education	Education	Education, other
Health	Medical sciences	Dentistry
Health	Medical sciences	Environmental health
Health	Medical sciences	Public health
Health	Medical sciences	Public health/epidemiology
Health	Medical sciences	Epidemiology
Health	Medical sciences	Medicine/surgery
Health	Medical sciences	Optometry/ophthalmology
Health	Medical sciences	Pharmacy
Health	Medical sciences	Veterinary science
Health	Other health sciences	Speech/language pathology and audiology
Health	Other health sciences	Health systems/services administration
Health	Other health sciences	Hospital administration
Health	Other health sciences	Nursing

Table B.3 (continued).

Field	College Major	Discipline
Health	Other health sciences	Rehabilitation/therapeutic services
Health	Other health sciences	Health sciences, general
Health	Other health sciences	Health sciences, other
Humanities	English/literature	Classics
Humanities	English/literature	Comparative literature
Humanities	English/literature	English/American literature
Humanities	English/literature	Literature, American
Humanities	English/literature	Literature, English
Humanities	English/literature	English language
Humanities	English/literature	Speech/rhetorical studies
Humanities	English/literature	Letters, general
Humanities	English/literature	Letters, other
Humanities	Foreign languages/literatures	French
Humanities	Foreign languages/literatures	German
Humanities	Foreign languages/literatures	Italian
Humanities	Foreign languages/literatures	Spanish
Humanities	Foreign languages/literatures	Russian
Humanities	Foreign languages/literatures	Slavic (other than Russian)
Humanities	Foreign languages/literatures	Chinese
Humanities	Foreign languages/literatures	Japanese
Humanities	Foreign languages/literatures	Hebrew
Humanities	Foreign languages/literatures	Arabic
Humanities	Foreign languages/literatures	Other languages/literature
Humanities	History	History, American
Humanities	History	History, European
Humanities	History	History, general
Humanities	History	History, other
Humanities	Religion/theology	Religious studies
Humanities	Religion/theology	Theology/ministries
Humanities	Other Humanities	Liberal arts/other humanities
Humanities	Other Humanities	Philosophy
Humanities	Other Humanities	Visual/performing arts
Professional Fields	Business/management	Agricultural business/management
Professional Fields	Business/management	Accounting
Professional Fields	Business/management	Banking/financial support services
Professional Fields	Business/management	Business administration/management
Professional Fields	Business/management	Business/managerial economics
Professional Fields	Business/management	Management information systems/business data processing

Table B.3 (continued).

Field	College Major	Discipline
Professional Fields	Business/management	Marketing management/research
Professional Fields	Business/management	Business statistics
Professional Fields	Business/management	Operations research
Professional Fields	Business/management	Organizational behavior
Professional Fields	Business/management	Business management/administrative services, general
Professional Fields	Business/management	Business management/administrative services, other
Professional Fields	Information fields	Communications research
Professional Fields	Information fields	Journalism
Professional Fields	Information fields	Mass communications
Professional Fields	Information fields	Radio/television
Professional Fields	Information fields	Communication theory
Professional Fields	Information fields	Communications, general
Professional Fields	Information fields	Communications, other
Professional Fields	Information fields	Library science
Professional Fields	Other Professional Fields	Architecture/related programs
Professional Fields	Other Professional Fields	Personal/culinary services
Professional Fields	Other Professional Fields	Engineering-related technologies
Professional Fields	Other Professional Fields	Home economics/family studies
Professional Fields	Other Professional Fields	Law and legal studies
Professional Fields	Other Professional Fields	Reserve officer training corps (ROTC)
Professional Fields	Other Professional Fields	Military technologies
Professional Fields	Other Professional Fields	Multi-/interdisciplinary studies
Professional Fields	Other Professional Fields	Parks/recreation/leisure/fitness
Professional Fields	Other Professional Fields	Basic skills
Professional Fields	Other Professional Fields	Citizenship activities
Professional Fields	Other Professional Fields	Health related knowledge/skills
Professional Fields	Other Professional Fields	Interpersonal/social skills
Professional Fields	Other Professional Fields	Personal awareness/self-improvement
Professional Fields	Other Professional Fields	Science technologies
Professional Fields	Other Professional Fields	Protective services
Professional Fields	Other Professional Fields	Public administration/social services professions
Professional Fields	Other Professional Fields	Construction trades
Professional Fields	Other Professional Fields	Mechanic/repair technologies
Professional Fields	Other Professional Fields	Precision production trades
Professional Fields	Other Professional Fields	Transportation/materials moving workers

APPENDIX C

METHODS FOR ANALYSIS OF NELS:88 DATA

C.1 DATA HANDLING

Working with the extensive volume of records and variables contained within the NELS:88 dataset required commensurately more time than a smaller dataset. Although SAS has features that permit the analyst to examine data at the individual record level, these were found to be less efficient than using Microsoft Access. Access was used in two main ways. First, it was a tool to quickly sort combinations of variables to determine the settings that would classify the data as desired. This supported the task of creating SAS code that examined combinations of variables, flagged the records that met specific criteria, created new variables to record the status, and sorted the records into different classes. Second, the query design capabilities of Access allowed key variables to be selected and presented in a table format. This in turn was used to quickly obtain the names of covariates and interaction terms in a format that could be directly copied into SAS programs more easily than manually typing the names.

Variables that were of potential interest in the classification and modeling process were selected within the NELS Electronic Codebook for Windows tool to create a variable "tag file." Then the Codebook was used to export the data for those variables to an Access database. The

Codebook also created the initial SAS programming code to import the data for these variables and their associated formatting into SAS. The ability of Access to sort records and select those that meet specific criteria made it much easier to examine the records, become familiar with the variable values, and draft the SAS code to manipulate the variables. While this could have been accomplished within SAS alone, the use of Access simplified and accelerated the task. In conducting complicated analysis there is no substitute for becoming deeply familiar with the raw data. The query building capability of Access was extremely beneficial in this analysis. Many different queries were constructed to examine subsets of records and variables sorted in a manner that highlighted any discrepancies in the classification process. The subsets of data provided in the different queries allowed the data to be examined in manageable portions as opposed to inspecting over 11,000 records at once. The SAS code was iteratively developed by creating draft code to classify the records, running the programs, creating temporary SAS datasets that contained the original and new classification variables, exporting the dataset to an Access database, examining the results in Access to see if each record was properly handled, and then repeating the process until no exceptions remained.

Many different models were created during this research and several were constructed multiple times with different random samples of fit and test data. Altering the SAS code to construct these models was simplified by using Access to generate lists of potential covariates for specific models. In particular, the code for models that explored different interaction terms was partially created by flagging variables found to be significant and writing queries that created the SAS code for the Interaction terms. This would not have saved much time over manually typing the interaction terms into SAS for smaller models, but for the extensive sets of covariates in this analysis the time savings was notable.

Microsoft Excel was also used to aid in creating graphs of the ROC Curves when the fitted models were applied to the test data. The predictions for each of the records at many different cutpoints were exported into spreadsheets that calculated the sensitivity, specificity, etc. and produced the graphs.

APPENDIX D

CODE FOR SAS PROGRAMS TO CLASSIFY RECORDS

D.1 SAS CODE FOR ORIGINAL CLASSIFYING SCHEME SORTING DATA BY DEGREE OUTCOME

This code first determined which records related to students that participated in all five waves of data collection. Then a series of flag variables were constructed to sort the records into the STEM, STEM-Related, Non-STEM, Sub-4 Yr Degree, and No Degree categories based upon the combination of majors and degrees earned. The text of the SAS code is presented here in a color-coded format that matches that used by SAS.

```
if (F4UNIV1 > 1024 & F4UNIV1 < 1041) then All5 = 1;
else if (F4UNIV1 > 1048 & F4UNIV1 < 1065) then All5 = 1;
else if (F4UNIV1 > 1075 & F4UNIV1 < 1085) then All5 = 1;
else if (F4UNIV1 > 1125 & F4UNIV1 < 1130) then All5 = 1;
else if (F4UNIV1 > 1132 & F4UNIV1 < 1144) then All5 = 1;
else All5 = 0;

* Create flag variables for STEM, STEM-Related, NonSTEM & No Degree ;
if(F4EMJ1D =112) then STEMD1 = 1;
else if (F4EMJ1D > 139 & F4EMJ1D < 145) then STEMD1 = 1;
else if (F4EMJ1D > 259 & F4EMJ1D < 272) then STEMD1 = 1;
else if (F4EMJ1D > 399 & F4EMJ1D < 404) then STEMD1 = 1;
else STEMD1 = 0;
```

```
if(F4EMJ2D =112) then STEMD2 = 1;
else if (F4EMJ2D > 139 & F4EMJ2D < 145) then STEMD2 = 1;
else if (F4EMJ2D > 259 & F4EMJ2D < 272) then STEMD2 = 1;
else if (F4EMJ2D > 399 & F4EMJ2D < 404) then STEMD2 = 1;
else STEMD2 = 0;

if (F4EMJ1D > 9 & F4EMJ1D < 32) then STEMrelD1 = 1;
else if (F4EMJ1D > 59 & F4EMJ1D < 63) then STEMrelD1 = 1;
else if (F4EMJ1D > 99 & F4EMJ1D < 112) then STEMrelD1 = 1;
else if (F4EMJ1D in(150, 170, 190, 194, 420, 450, 454, 471)) then STEMrelD1 = 1;
else if (F4EMJ1D > 174 & F4EMJ1D < 186) then STEMrelD1 = 1;
else if (F4EMJ1D > 300 & F4EMJ1D < 304) then STEMrelD1 = 1;
else STEMrelD1 = 0;

if (F4EMJ2D > 9 & F4EMJ2D < 32) then STEMrelD2 = 1;
else if (F4EMJ2D > 59 & F4EMJ2D < 63) then STEMrelD2 = 1;
else if (F4EMJ2D > 99 & F4EMJ2D < 112) then STEMrelD2 = 1;
else if (F4EMJ2D in(150, 170, 190, 194, 420, 450, 454, 471)) then STEMrelD2 = 1;
else if (F4EMJ2D > 174 & F4EMJ2D < 186) then STEMrelD2 = 1;
else if (F4EMJ2D > 300 & F4EMJ2D < 304) then STEMrelD2 = 1;
else STEMrelD2 = 0;

If (STEMD1 = 1 & F4EDGR1 > 2 OR STEMD2 = 1 & F4EDGR2 > 2) then STEM = 1 ;
* If either the 1st or 2nd degree was STEM & got a 4 yr degree the person is classified as STEM
;
else STEM = 0 ;

if (STEMrelD1 = 1 & F4EDGR1 > 2 & STEM = 0) then STEMrel = 1;
else if (STEMrelD2 = 1 & F4EDGR2 > 2 & STEM = 0) then STEMrel = 1 ;
* If either the 1st or 2nd degree was STEM-Rel and neither was STEM the person is classified as
STEM-rel ;
else STEMrel = 0 ;

if (STEM + STEMrel > 0) then NonSTEM = 0 ;
else if (F4EDGR1 > 2 & F4EMJ1D > 0 OR F4EDGR2 > 2 & F4EMJ2D >0) then NonSTEM = 1
;
* if the student isn't STEM or STEM-Rel, has a 4yr degree, and the "deg" is valid then
NonSTEM ;
else NonSTEM = 0 ;

if (STEM + STEMrel + NonSTEM > 0) then Sub4YrDeg = 0;
else If (F4EDGR1 > 0 & F4EDGR1 < 3 OR F4EDGR2 > 0& F4EDGR2 < 3) then Sub4YrDeg =
1;
* if the 1st or 2nd degree is at least a certificate or assoc. but no higher the person has a sub 4 yr
degree ;
```

```
else Sub4YrDeg = 0 ;

if(STEM + STEMrel + NonSTEM + Sub4YrDeg = 0 & F4EDGR1 > 0 & F4EMJ1D < 0 &
F4EDGR2 > 0 & F4EMJ2D < 0) then NoDegree = 1;
* if student is not STEM/STEMrel/NonSTEM/assoc. and a 1st or 2nd degree is listed but not
  verifiable then person is "no Degree" ;
else if (STEM + STEMrel + NonSTEM + Sub4YrDeg = 0 & F4EDGR1 > 0 & F4EMJ1D < 0 &
F4EDGR2 < 0 & F4EMJ2D < 0) then NoDegree = 1;
* if student is not STEM/STEMrel/NonSTEM/assoc. and a 1st degree is listed but not
  verifiable then person is "no Degree" ;
else if(STEM + STEMrel + NonSTEM + Sub4YrDeg = 0 & F4EDGR1 < 0 & F4EDGR2 < 0 )
then NoDegree = 1;
* if student is not STEM/STEMrel/NonSTEM/assoc. and no degree is listed then person is "no
Degree" ;
else NoDegree = 0 ;
* otherwise the person is assumed to have some kind of degree ;

if (STEM = 1) then Category = 4 ;
else if (STEMrel = 1) then Category = 3 ;
else if (NonSTEM = 1) then Category = 2 ;
else if (Sub4yrDeg = 1) then Category = 1 ;
else if (NoDegree = 1) then Category = 0 ;
else Category = 5 ;
```

D.2 SAS CODE FOR RECODING COVARIATES

The code to accomplish this task was very lengthy but repetitive for many of the variables. Therefore, only a few examples of the covariates are shown here.

```
IF (BY2XMPRO =  3) then  BY2XMPROro = 4 ;
else if (BY2XMPRO =  2) then  BY2XMPROro = 3 ;
else if (BY2XMPRO =  1) then  BY2XMPROro = 2 ;
else if (BY2XMPRO =  0) then  BY2XMPROro = 1 ;
else   BY2XMPROro = 0 ;

IF (BYFAMINC <  16) then  BYFAMINCro = BYFAMINC ;
else   BYFAMINCro = 0 ;
```

```
* Dummy variables for Sector of High School student plans to attend.
BYS14rPvREL = 1 if private religious, BYS14PvNRel = 1 if private non-religious
and both = 0 if public or not sure.  ;
IF (BYS14 =  2) then  BYS14rPvRel = 1 ;
else   BYS14rPvRel = 0 ;
IF (BYS14 =  3) then  BYS14rPvNRel = 1 ;
else   BYS14rPvNRel = 0 ;

IF (BYS34A >  0)& (BYS34A <  3) then  BYS34Arb = 0 ;
Else IF (BYS34A >  3)& (BYS34A <  8) then  BYS34Arb = 1 ;
else   BYS34Arb = 0 ;

IF (F4RACE2 = 1) then  F4RACE2rAI = 1 ;
else   F4RACE2rAI = 0 ;
if(F4RACE2 = 2 ) then  F4RACE2rAs = 1 ;
else   F4RACE2rAs = 0 ;
if(F4RACE2 = 3 ) then  F4RACE2rBl = 1 ;
else   F4RACE2rBl = 0 ;
if(F4RACE2 = 5 ) then  F4RACE2rHi = 1 ;
else   F4RACE2rHi = 0 ;
*  if F4RACE2 = 5 for White (not Hispanic) or 9 for missing then
all of the recoded F4RACE2r dummy variables = 0, the reference case ;
```

D.3 SAS CODE FOR CLASSIFYING DATA BY STEM TRACK DEPARTURE TYPE

This code created the revised classification scheme by first determining the timing of events such as high school graduation, college graduation, and dropping out of high school. Then a lengthy section of logic examined multiple records to determine the students' final educational outcome and the timing of the STEM track departure if this occurred. In several cases the records had to be inspected individually to judge the proper classification. Once a decision was made the records were individually coded into the classification by using the unique ID variable assigned to the students. The ID codes for individual students have been replaced in the code provided

here by "******" for privacy reasons. After the departure type was identified, additional programming logic was used to code the time of STEM track departure and whether the time was observed or censored.

```
* Create the decimal time versions of the F4HSGRDT, F4ED1, and F4ED2 dates.  ;
If(F4HSGRDT > 0 and F4HSGRmon = 01) then F4HSGRDTrn = F4HSGRyr + (0/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 0) then F4HSGRDTrn = F4HSGRyr + (0/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 02) then F4HSGRDTrn = F4HSGRyr + (1/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 03) then F4HSGRDTrn = F4HSGRyr + (2/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 04) then F4HSGRDTrn = F4HSGRyr + (3/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 05) then F4HSGRDTrn = F4HSGRyr + (4/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 06) then F4HSGRDTrn = F4HSGRyr + (5/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 07) then F4HSGRDTrn = F4HSGRyr + (6/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 08) then F4HSGRDTrn = F4HSGRyr + (7/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 09) then F4HSGRDTrn = F4HSGRyr + (8/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 10) then F4HSGRDTrn = F4HSGRyr + (9/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 11) then F4HSGRDTrn = F4HSGRyr + (10/12) ;
else if(F4HSGRDT > 0 and F4HSGRmon = 12) then F4HSGRDTrn = F4HSGRyr + (11/12) ;
else F4HSGRDTrn = F4HSGRDT ;

if(F4HSGRDTrn > 0) then F4HSGRDTrn = ROUND(F4HSGRDTrn,.01) ;

If(F4ED1 > 0 and F4ED1mon = 01) then F4ED1rn = F4ED1yr + (0/12) ;
else if(F4ED1 > 0 and F4ED1mon = 0) then F4ED1rn = F4ED1yr + (0/12) ;
else if(F4ED1 > 0 and F4ED1mon = 02) then F4ED1rn = F4ED1yr + (1/12) ;
else if(F4ED1 > 0 and F4ED1mon = 03) then F4ED1rn = F4ED1yr + (2/12) ;
else if(F4ED1 > 0 and F4ED1mon = 04) then F4ED1rn = F4ED1yr + (3/12) ;
else if(F4ED1 > 0 and F4ED1mon = 05) then F4ED1rn = F4ED1yr + (4/12) ;
else if(F4ED1 > 0 and F4ED1mon = 06) then F4ED1rn = F4ED1yr + (5/12) ;
else if(F4ED1 > 0 and F4ED1mon = 07) then F4ED1rn = F4ED1yr + (6/12) ;
else if(F4ED1 > 0 and F4ED1mon = 08) then F4ED1rn = F4ED1yr + (7/12) ;
else if(F4ED1 > 0 and F4ED1mon = 09) then F4ED1rn = F4ED1yr + (8/12) ;
else if(F4ED1 > 0 and F4ED1mon = 10) then F4ED1rn = F4ED1yr + (9/12) ;
else if(F4ED1 > 0 and F4ED1mon = 11) then F4ED1rn = F4ED1yr + (10/12) ;
else if(F4ED1 > 0 and F4ED1mon = 12) then F4ED1rn = F4ED1yr + (11/12) ;
else F4ED1rn = F4ED1 ;

if(F4ED1rn > 0) then F4ED1rn = ROUND(F4ED1rn,.01) ;


If(F4ED2 > 0 and F4ED2mon = 01) then F4ED2rn = F4ED2yr + (0/12) ;
else if(F4ED2 > 0 and F4ED2mon = 0) then F4ED2rn = F4ED2yr + (0/12) ;
```

```
else if(F4ED2 > 0 and F4ED2mon = 02) then F4ED2rn = F4ED2yr + (1/12) ;
else if(F4ED2 > 0 and F4ED2mon = 03) then F4ED2rn = F4ED2yr + (2/12) ;
else if(F4ED2 > 0 and F4ED2mon = 04) then F4ED2rn = F4ED2yr + (3/12) ;
else if(F4ED2 > 0 and F4ED2mon = 05) then F4ED2rn = F4ED2yr + (4/12) ;
else if(F4ED2 > 0 and F4ED2mon = 06) then F4ED2rn = F4ED2yr + (5/12) ;
else if(F4ED2 > 0 and F4ED2mon = 07) then F4ED2rn = F4ED2yr + (6/12) ;
else if(F4ED2 > 0 and F4ED2mon = 08) then F4ED2rn = F4ED2yr + (7/12) ;
else if(F4ED2 > 0 and F4ED2mon = 09) then F4ED2rn = F4ED2yr + (8/12) ;
else if(F4ED2 > 0 and F4ED2mon = 10) then F4ED2rn = F4ED2yr + (9/12) ;
else if(F4ED2 > 0 and F4ED2mon = 11) then F4ED2rn = F4ED2yr + (10/12) ;
else if(F4ED2 > 0 and F4ED2mon = 12) then F4ED2rn = F4ED2yr + (11/12) ;
else F4ED2rn = F4ED2 ;

if(F4ED2rn > 0) then F4ED2rn = ROUND(F4ED2rn,.01) ;

if(F2D6Y = 2) then F2D6yr = 1988 ;
else if(F2D6Y = 3) then F2D6yr = 1989 ;
else if(F2D6Y = 4) then F2D6yr = 1990 ;
else if(F2D6Y = 5) then F2D6yr = 1991 ;
else if(F2D6Y = 6) then F2D6yr = 1992 ;
else if(F2D6Y > 6) then F2D6yr = F2D6Y ;
else F2D6yr = 0 ;

If(F2D6Y > 1 and F2D6Y < 98 and F2D6M < 13) then F2D6Yrn = F2D6yr + ((F2D6M - 1)/12)
;
else if(F2D6Y > 1 and F2D6Y < 98 and F2D6M > 13 and F2D6M < 99)then F2D6Yrn = F2D6yr
+ (5/12) ;
else F2D6Yrn = 0 ;

if(F2D6Yrn > 0) then F2D6Yrn = ROUND(F2D6Yrn,.01) ;

if(F1D7YEAR = 1) then F1D7YEARyr = 1987 ;
else if(F1D7YEAR = 2) then F1D7YEARyr = 1988 ;
else if(F1D7YEAR = 3) then F1D7YEARyr = 1989 ;
else if(F1D7YEAR = 4) then F1D7YEARyr = 1990 ;
else if(F1D7YEAR > 4) then F1D7YEARyr = F1D7YEAR ;
else F1D7YEARyr = 0 ;

If(F1D7YEAR > 0 and F1D7YEAR < 8 and F1D7MNTH < 13) then F1D7Yrn = F1D7YEARyr
+ ((F1D7MNTH - 1)/12) ;
else if(F1D7YEAR > 0 and F1D7YEAR < 98 and F1D7MNTH > 13 and F1D7MNTH < 99)then
F1D7Yrn = F1D7YEARyr + (5/12) ;
else F1D7Yrn = 0 ;

if(F1D7Yrn > 0) then F1D7Yrn = ROUND(F1D7Yrn,.01) ;
```

```
* Identify which reported degree date happened sooner ;
if(F4ED1rn < = F4ED2rn and F4ED1rn > 0 and F4ED2rn > 0 )then D1bD2 = 1 ;
* D1 <= D2 and both valid ( > 0) so D1 happened before or =  D2 ;
else if (F4ED1rn > F4ED2rn and F4ED1rn > 0 and F4ED2rn < 0 ) then D1bD2 = 1 ;
* D1 > D2 but only D1 is valid so D1 is the one to use ;
else if (F4ED1rn > F4ED2rn and F4ED1rn > 0 and F4ED2rn > 0 ) then D1bD2 = 2 ;
* D1 > D2 and both are valid so D2 preceded D2 ;
else if (F4ED1rn < F4ED2rn and F4ED1rn < 0 and F4ED2rn > 0 ) then D1bD2 = 2 ;
* D1 < D2 but only D2 is valid so D2 is the one to use ;
else D1bD2 = 0 ;
* D1 and D2 both invalid (< 0) so neither is usable ;

* Departure Type code   ;

if(Category = 0 and ALLHDEG = 0) then Depart_Type = 1 ;
*  dropped out of high school and never got a diploma ;
else if(Category = 0 and ALLHDEG = 10) then Depart_Type = 7 ;
*  Move the Chem PhD student to STEM  ;

else if(Category = 0 and ALLHDEG = 1 and CREDRET = -8) then Depart_Type = 2 ;
*  graduated high school, never earned a college degree, and never claimed PSE attend.
(need to figure out who tried to get one.);
else if(Category = 0 and ALLHDEG = 1 and CREDRET > 3 and IN2000SC =0)then
Depart_Type = 3 ;
* graduated high school, attempted college, and dropped out with no degree ;
else if(Category = 0 and ALLHDEG = 1 and CREDRET > 3 and INSCHOOL >4 and
F4HSGRDT > 0)then Depart_Type = 2 ;
* graduated high school, may have taken college but not for a degree ;
else if(Category = 0 and ALLHDEG = 1 and CREDRET > 3 and INSCHOOL >4 and
F4HSGRDT < 0)then Depart_Type = 8 ;
* Allegedly graduated high school, may have taken college but not for a degree and cannot
confirm HS grad.
   Excluded from further analysis.  Applies to 8 records;
else if(Category = 0 and ALLHDEG > 0 and CREDRET > 3 and INSCHOOL >6) then
Depart_Type = 3 ;
* graduated high school, attempted college, and dropped out with no degree
  Note:  not in school in 2000 ;
else if(Category = 0 and ALLHDEG > 0 and CREDRET > 3 and INSCHOOL <5 ) then
Depart_Type = 4 ;
* graduated high school, attempted college, in school in 2000
  but study ended before degree earned.  ;
else if(Category = 0 and ALLHDEG = 1 and CREDRET < 0) then Depart_Type =  2 ;
* graduated high school and never attempted college ;

else if(Category = 0 and ALLHDEG = 2) then Depart_Type = 2 ;
*  graduated high school and pursued an unknown college degree
```

```
   Without a better way to figure out the major, depart time = HS grad.
   Check these students out to see if some had discernable majors;
else if(Category = 0 and ALLHDEG > 2 and ALLHDEG < 5) then Depart_Type = 3 ;
*  graduated high school and entered college, but no degree
   Without a better way to figure out what happened, depart time = HS grad.
   Check these students out to see if some had discernable majors and degrees;
else if(Category = 0 and ALLHDEG > 4 and ALLHDEG < 10) then Depart_Type = 8 ;
*  graduated high school and entered college, degree status conflicts between
  NOT and NOR.  Exclude from further analysis.  Applies to 155 records;

else if(Category = 4 and F4ED1 > 0 ) then Depart_Type = 7 ;
* no departure - earned a STEM degree ;
 else if(Category = 4 and F4ED1 < 0 ) then Depart_Type = 8 ;
*  Reported a STEM degree but have no valid graduation date.
   Applies to 3 cases that earned Non-STEM and STEM-Related degrees
   according to the PETS NOT file.;

else if(Category = 1 and ALLHDEG < 5 ) then Depart_Type = 5 ;
*  Sub 4 year deg (or no degree according to ALLHDEG).  ;
else if(Category = 1 and ALLHDEG = 5 ) then Depart_Type = 5 ;
*  Got a different degree ;
else if(Category = 1 and ALLHDEG > 4 and F4ED1 > 0) then Depart_Type = 5 ;
*  Graduated with a Sub 4 Year Degree ;
else if(Category = 1 and ALLHDEG > 4 and F4ED1 < 0) then Depart_Type = 3 ;
*  graduated high school and entered college, but no degree
   Without a better way to figure out what happened, depart time = HS grad.
   Check these students out to see if some had discernable majors and degrees;

else if(Category = 2 and ALLHDEG > 4) then Depart_Type = 6 ;
*  Got a NonSTEM degree ;
else if(Category = 2 and ALLHDEG = 4 and INSCHOOL = 9) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 4 and INSCHOOL < 5) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 4 and INSCHOOL = 8) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 3) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 2) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 2) then Depart_Type = 6 ;
*  Earned a different degree ;
else  if(Category  =  2  and  ALLHDEG  =  1  and  CREDRET  =  -1  and  F4HSGRDT  >0)  then
Depart_Type = 6 ;
*  Earned a different degree (relates to 2 students with no transcripts but reported degrees);
```

```
else if(Category = 2 and ALLHDEG = 1 and CREDRET > 0 and F4EDGR1 >0 or F4EDGR2
>0) then Depart_Type = 6 ;
*  Earned a different degree ;
else if(Category = 2 and ALLHDEG = 1 and CREDRET > 0 and F4EDGR1 <0 and F4EDGR2
<0 and F4HSGRDT >0 ) then Depart_Type = 2 ;
*  Graduated HS but no college (relates to 1 student who reported a degree & major,
   but had no transcript or valid reported college graduation date);
else if(Category = 2 and ALLHDEG = 0 and CREDRET = -1 and F4HSGRDT >0) then
Depart_Type = 6 ;
*  Earned a different degree (relates to 1 student with no transcripts but reported degree;
else if(Category = 2 and ALLHDEG = 0 and CREDRET = -1 and F4HSGRDT < 1) then
Depart_Type = 8 ;
*   Not discernable (relates to 1 student with no transcripts, a reported degree, but not verifible
HS grad date;

*else if(Category = 2 and ALLHDEG < 4) then Depart_Type = 3 ;
*  Graduated HS, entered college, but dropped out without a degree ;

else if(Category = 3 and ALLHDEG > 4) then Depart_Type = 6 ;
*  Got a STEMrel degree ;
else if(Category = 3 and ALLHDEG < 5 and F4ED1 > 0) then Depart_Type = 6 ;
*  Got a STEMrel degree ;
else if(Category = 3 and ALLHDEG < 5 and F4ED1 < 0 and F4EDGR1 < 3) then Depart_Type
= 6 ;
*  graduated high school and entered college, but no degree
    Without a better way to figure out what happened, depart time = HS grad.
    Check these students out to see if some had discernable majors and degrees;
*else if(Category = 3 and ALLHDEG < 5 and F4ED1 < 0 and F4EDGR1 = 3) then Depart_Type
= 6 ;
*  graduated high school, entered college, reported a degree,
   had valid PSBEG & PSEND dates, but had no valid grad date.
    Applies to 2 STEMrel classified students that had valid PSBEG & PSEND dates
     but not a valid graduation date. ;

else Depart_Type = 9 ;
*  Category mismatch - research ;

if(ID =****** or ID = ******) then Depart_Type = 8 ;
*  Reported earning a different degree (Non-STEM), but info. not reconcileable with PSE data.
   (Applies to 2 records  ID #'s ****** and ****** [withheld for privacy] ) Make Departure
Type 9 not 4;
if(ID = ******) then Depart_Type = 2 ;
*  Reported attempting college but no substantiation, just HS Graduation shown ;
if(ID in((******,
******,
******,))then Depart_Type = 3 ;
```

```
*  Reported earning a Sub 4yr degree, but no valid graduate date given
   or available from the PSE data.
   Changes the classification from Departure Type 4 to 3 ;

if(ID in(******,******))then Depart_Type = 2 ;
* Applies to 2 students classified as STEMrel based on reported majors
& degrees, but no valid graduation dates given either in NOR or PETS files ;

if(ID in((******,
******,
******,
) )then Depart_Type = 8 ;
*  Reported earning a Sub 4yr degree, but no valid graduate date given
   or available from the PSE data and not reconcilable with PSE info.
   Changes the classification from Departure Type 4 to 9 to exclude
   these records from further analysis;

If(ID in (******,
******,
******,
)) then Depart_Type = 8 ;
*  Applies to 14 records originally classified as HS dropouts but for which inconsistent
   data is present.  The F3EVDOST flag variable suggests they did not permanently drop out. ;

Observed = 0 ;

* Departure Time identification code  ;
if(Depart_Type = 1 and F2D6Yrn > 99) then Depart_Time = F2D6Yrn ;
else if (Depart_Type = 1 and F2D6Yrn <= 99 and F1D7Yrn > 99) then Depart_Time = F1D7Yrn
;
else if (Depart_Type = 1) then Depart_Time = 1991 ;
* Departure time for students classified as HS dropouts is based on 2 variables.
  If the last date reported to be in high school as of F2 is valid, that date is used.
  If the F2 date is not valid, but the last date reported to be in HS as of F1 is, that
    date is used.  If neither is valid, then the departure date is imputed to be 1991, the
    year between F1 and F2.  ;

else if(Depart_Type = 2 and PETSHSDT > 0) then Depart_Time = PETSHSDT ;
else if(Depart_Type = 2 and PETSHSDT <= 0 and F4HSGRDTrn > 0) then Depart_Time =
F4HSGRDTrn ;
* HS grads: set departure time based on the PETS NOT H.S. graduation date.
  All but 6 of the 2,367 records with Departure Type 2 have a valid PETSHSDT grad date.
  time is decimal with YYYY.mm  where mm = MM/12 with Jan. = 0, Dec = 11/12 ;

else if(Depart_Type = 3) then Depart_Time = PSEND ;
* College dropouts: set departure time based on the last date at which
```

```
  the student was reported   to be attending a post-secondary institution.
  Each of the 2,163 records had a   valid PSEND date.
  time is decimal with YYYY.mm  where mm = MM/12 ;

else if(Depart_Type = 4) then Depart_Time = 2001.00  ;
* Study ended before degree earned: set the depart time based on
  study end/censoring date of Dec. 31, 2000.   Time = 2001.00 ;

* Got a Sub 4 yr deg:
  1,704 records ;

else If(Depart_Type = 5  and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 1 and D1bD2 = 1)
  then Depart_Time = F4ED1rn ;
* both degrees are Sub4YrDeg and D1 is the earliest ( & D1 has a valid date)  ;
else if(Depart_Type = 5 and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 1 and D1bD2 = 2)
  then Depart_Time = F4ED2rn ;
* both degrees are Sub4YrDeg and D2 is the earliest ( & D2 has a valid date) ;
else if(Depart_Type = 5  and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 0 and D1bD2 > 0)
  then Depart_Time = F4ED1rn;
* D1 is Sub4YrDeg with a valid date and D2 isn't Sub4, use D1  ;
else if(Depart_Type = 5 and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 1 and D1bD2 > 0)
  then Depart_Time = F4ED2rn;
* D2 is Sub4YrDeg with a valid date and D1 isn't Sub4, use D2  ;
else If(Depart_Type = 5  and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 0 and D1bD2 = 1)
  then Depart_Time = F4ED1rn ;
* at least one degree in a 4 year subject but only got Certif or Assoc. and
   D1 had a valid date - about 73 records  ;
else if(Depart_Type = 5  and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 0 and D1bD2 = 1
        and DEGDAT2 > 0 and DEGDAT3 < 0) then Depart_Time = DEGDAT2;
* Should apply only to ???? that reported no valid degrees, but
  was classified as Sub4YrDeg based on the transcript data;
else if(Depart_Type = 5  and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 0 and D1bD2 = 1
        and DEGDAT2 < 0 and DEGDAT3 > 0) then Depart_Time = DEGDAT3;
* Should apply only to ???? that reported no valid degrees, but
  was classified as Sub4YrDeg based on the transcript data;
else if(Depart_Type = 5 and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 0
   and D1bD2 = 0 and DEGDAT2 > 0 and DEGDAT3 <= 0) then Depart_Time = DEGDAT2;
* Should apply only to records that reported valid Sub4YrDeg degrees w/ no valid degree dates,
but
  was also classified as Sub4YrDeg based on the transcript data   ;
else if(Depart_Type = 5 and Sub4YrDegD1 = 0 and Sub4YrDegD2 = 0
   and D1bD2 = 0 and DEGDAT2 <= 0 and DEGDAT3 > 0) then Depart_Time = DEGDAT3;
* Should apply only to records that reported valid Sub4YrDeg degrees w/ no valid degree dates,
but
  was also classified as Sub4YrDeg based on the transcript data   ;
else if(Depart_Type = 5 and Sub4YrDeg = 1 and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 0
```

```
 and D1bD2 = 0 and DEGDAT2 > 0 and DEGDAT3 <= 0) then Depart_Time = DEGDAT2;
* Should apply only to records that reported valid Sub4YrDeg degrees w/ no valid degree dates,
but
  was also classified as Sub4YrDeg based on the transcript data   ;
else if(Depart_Type = 5 and Sub4YrDeg = 1 and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 0
 and D1bD2 = 0 and DEGDAT2 <= 0 and DEGDAT3 > 0) then Depart_Time = DEGDAT3;
* Should apply only to records that reported valid Sub4YrDeg degrees w/ no valid degree dates,
but
  was also classified as Sub4YrDeg based on the transcript data   ;
* Got a STEMrel or NonSTEM deg:  set the depart time based on F4ED1, F4ED2, or DEGDAT4
or DEGDAT5.
    3,158 records ;
else if(Depart_Type = 5 and Sub4YrDeg = 1 and Sub4YrDegD1 = 1 and Sub4YrDegD2 = 0
 and D1bD2 = 0 and DEGDAT2 > 0 and DEGDAT3 > 0) then Depart_Time = DEGDAT2;
 *  should apply to just one record ID = 781418 that reported Sub 4 deg with no valid dates, but
had
     valid dates on transcript side for DEGDAT2 and DEGDAT3  ;

else If(Depart_Type = 6 and STEMrel = 1 and STEMrelD1 = 1 and STEMrelD2 = 1 and D1bD2
= 1) then Depart_Time = F4ED1rn ;
* both degrees are STEMrel and D1 is the earliest  ;
else if(Depart_Type = 6 and STEMrel = 1 and STEMrelD1 = 1 and STEMrelD2 = 1 and D1bD2
= 2) then Depart_Time = F4ED2rn ;
* both degrees are STEMrel and D2 is the earliest  ;
else  if(Depart_Type  =  6  and  STEMrel  =  1  and  STEMrelD1  =  1  and  STEMrelD2  =  0)  then
Depart_Time = F4ED1rn;
* D1 is STEMrel and D2 isn't use D1  ;
else  if(Depart_Type  =  6  and  STEMrel  =  1  and  STEMrelD1  =  0  and  STEMrelD2  =  1)  then
Depart_Time = F4ED2rn;
* D2 is STEMrel and D1 isn't use D2  ;
else  if(Depart_Type  =  6  and  STEMrel  =  1  and  STEMrelD1  =  0  and  STEMrelD2  =  0)  then
Depart_Time = DEGDAT4;
* Should apply only to ???? that reported no valid degrees, but
  was classified as STEMrel based on the transcript data;

*Got a NonSTEM degree ;

else If(Depart_Type = 6 and NonSTEM = 1 and NonSTEMD1 = 1 and NonSTEMD2 = 1 and
D1bD2 = 1) then Depart_Time = F4ED1rn ;
* both degrees are NonSTEM and D1 is the earliest  ;
else if(Depart_Type = 6 and NonSTEM = 1 and NonSTEMD1 = 1 and NonSTEMD2 = 1 and
D1bD2 = 2) then Depart_Time = F4ED2rn ;
* both degrees are NonSTEM and D2 is the earliest  ;
else if(Depart_Type = 6 and NonSTEM = 1 and NonSTEMD1 = 1 and NonSTEMD2 = 0 and
D1bD2 > 0) then Depart_Time = F4ED1rn;
* D1 is NonSTEM and D2 isn't use D1  ;
```

```
else if(Depart_Type = 6 and NonSTEM = 1 and NonSTEMD1 = 0 and NonSTEMD2 = 1 and
D1bD2 > 0) then Depart_Time = F4ED2rn;
* D2 is NonSTEM and D1 isn't use D2  ;
else if(Depart_Type = 6 and NonSTEM = 1 and NonSTEMD1 = 0 and NonSTEMD2 = 0 and
D1bD2 > 0) then Depart_Time = DEGDAT4;
* Should apply only to ???? that reported no valid degrees, but
  was classified as NonSTEM based on the transcript data;
else if(Depart_Type = 6 and NonSTEM = 1 and D1bD2 = 0) then Depart_Time = DEGDAT4;
* Should apply only to 4 records that reported valid NonSTEM degrees w/ no valid degree dates,
but
  was also classified as NonSTEM based on the transcript data  and B_NonSTEM = 1 ;

* Got a STEM deg:  set the depart time based on F4ED1, F4ED2, or DEGDAT4 or DEGDAT5.
  Then write the code based on   whether D1 or D2 was the degree in question
  and which came first or the associated DEGDAT4 date if there's not a valid
  F4ED1 or F4ED2 date. Departure time randomly right censored because STEM
  track departure never happened and event not observed
  736 records ;

else If(Depart_Type = 7 and STEMD1 = 1 and STEMD2 = 1 and D1bD2 = 1) then Depart_Time
= F4ED1rn ;
* both degrees are STEM and D1 is the earliest  ;
else if(Depart_Type = 7 and STEMD1 = 1 and STEMD2 = 1 and D1bD2 = 2) then Depart_Time
= F4ED2rn ;
* both degrees are STEM and D2 is the earliest  ;
else if(Depart_Type = 7 and STEMD1 = 1 and STEMD2 = 0) then Depart_Time = F4ED1rn;
* D1 is STEM and D2 isn't use D1  ;
else if(Depart_Type = 7 and STEMD1 = 0 and STEMD2 = 1) then Depart_Time = F4ED2rn;
* D2 is STEM and D1 isn't use D2  ;
else if(Depart_Type = 7 and STEMD1 = 0 and STEMD2 = 0) then Depart_Time = DEGDAT4;
* Should apply only to the Chemistry PhD that reported no valid degrees, but
  was classified as STEM based on the transcript data;

else if(Depart_Type = 8) then Depart_Time = 0 ;
* Excluded record:  set the depart time to 0   186 records. ;

else Depart_Time = 10 ;

if(Depart_Type = 7) then STEM_Outcome = 1 ;
else STEM_Outcome = 0 ;
* label the student outcomes as STEM or not ;

 if(ID = ******) then Depart_Time = DEGDAT4 ;
* hard code the Chemistry PHD's departure time to the date
  of the bachelor's degree.  Should be redundant line. ;
```

```
if(Depart_Type = 2) then  Observed = 1;
else if(Depart_Type = 3) then Observed = 1;
else if(Depart_Type = 4) then Observed = 0;
* Observed = 0 because the event of departing STEM track wasn't observed.
  There were 433 students in this status.  ;
else if(Depart_Type = 5) then Observed = 1;
else if(Depart_Type = 6) then Observed = 1;
else if(Depart_Type = 7) then Observed = 0;
* Observed = 0 because the event of departing STEM track wasn't observed.
  There were 736 students in this status.  ;
else if(Depart_Type = 8) then Observed = 0;
else if(Depart_Type = 1) then Observed = 1;
else Observed = 2 ;

Track_Time = Depart_Time - 1987.92;

if(BDAYrn > 0) then Educ_Dur = Depart_Time - BDAYrn ;
else Educ_Dur = 0 ;
* ;
if(Educ_Dur > 0) then Educ_Dur = ROUND(Educ_Dur,.01) ;
```

BIBLIOGRAPHY

[1] Bertsimas, Dimitris and John N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, MA, 1997.

[2] Montgomery, Douglas C., Elizabeth A. Peck, & G. Geoffrey Vining, *Introduction to Linear Regression Analysis* 3rd edition, John Wiley & Sons, Inc., New York, NY, 2001.

[3] Seber, George Arthur Frederick & Christopher J. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., New York, NY, 1989.

[4] Hosmer, David W. and Stanley Lemeshow, "Applied Logistic Regression 2nd edition," John Wiley & Sons, Inc., New York, NY, 2000

[5] Klein, John P. & Melvin L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* 2nd edition, Springer-Verlag New York, Inc., New York, NY, 2003.

[6] U.S. Department of Transportation, Surface Transportation Board, Code of Federal Regulations, Title 49, Chapter 10, Part 1201, http://www.access.gpo.gov/nara/cfr/waisidx_00/49cfrv7_00.html, October 2003.

[7] National Science Foundation, Division of Science Resources Statistics, *Graduate Students and Postdoctorates in Science and Engineering: Fall 2002*, NSF 05-310, Project Officers: Julia D. Oliver and Emilda B. Rivers (Arlington, VA 2004). (available from NSF website http://www.nsf.gov/statistics/nsf04318/).

[8] Commission on Professionals in Science and Technology (CPST), data derived from the American Association of Engineering Societies' (AAES) Engineering Workforce Commission report, *Engineering and Technology Enrollments*, *Fall 1990 through 2004*.

[9] National Science Foundation, Division of Science Resources Statistics, *Science and Engineering Degrees, by Race/Ethnicity of Recipients: 1992-2001*, NSF 04-318, Project Officers: Susan T. Hill and Jean M. Johnson (Arlington, VA 2004). (available from NSF website http://www.nsf.gov/statistics/nsf05310/).

[10] Astin, Alexander W., "Engineering Outcomes," *ASEE Prism*, September 1993, pp. 27-30.

[11] Berryman, S.E., *Who will do Science*? NY: The Rockefeller Foundation, 1983.

[12] National Center for Education Statistics, National Education Longitudinal Study of 1988, Project Officers: Peggy Quinn and Jeffrey T. Owings (Washington, DC), http://nces.ed.gov/surveys/nels88/index.asp.

[13] Hart, B. & T.R. Risely, (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: P.H. Brookes Co.

[14] Adams, Candace Renee and Kusum Singh, "Direct and Indirect Effects of School Learning Variables on the Academic Achievement of African American 10th Graders," *The Journal of Negro Education*, Vol. 67, No. 1, Winter 1998, pp. 48-66.

[15] Steinberg, Laurence, Susie D. Lamborn, Sanford M. Dornbusch, and Nancy Darling, "Impact of Parenting Practices on Adolescent Achievement: Authoritative Parenting, School Involvement, and Encouragement to Succeed," *Child Development*, Vol. 63, No. 5, October 1992, pp. 1266-1281.

[16] Darling-Hammond, Linda, "New Standards and Old Inequalities: School Reform and the Education of African American Students," *The Journal of Negro Education*, Vol. 69, No. 4, The School Reform Movement and the Education of African American Youth: A Retrospective Update, Autumn 2000, pp. 263-287.

[17] Good, Thomas L. and Sharon L. Nichols, "Expectancy Effects in the Classroom: A Special Focus on Improving the Reading Performance of Minority Students in First-Grade Classrooms," *Educational Psychologist,* Vol. 36, No. 2, Spring 2001, pp. 113–126.

[18] Steele, C. (1997). "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." in J.L. Eberhardt and S.T. Fiske (Eds.) *Confronting Racism: The Problem and the Response.* Thousand Oakes, CA: Sage, pp. 202–233.

[19] U.S. Department of Education, National Center for Education Statistics. *Educational Achievement and Black-White Inequality,* NCES 2001-061, by Jonathan Jacobson, Cara Olsen, Jennifer King Rice, Stephen Sweetland and John Ralph. Project Officer: John Ralph. Washington, DC: U.S. Government Printing Office, 2001.

[20] Astin, Alexander W., *What Matters in College?: Four Critical Years Revisited*, Jossey-Bass Publishers, San Francisco, 1993.

[21] Knapp, Michael S., "Between Systemic Reforms and the Mathematics and Science Classroom: The Dynamics of Innovation, Implementation, and Professional Learning," *Review of Educational Research*, Summer 1997, Vol. 67, No. 2, pp. 227-266.

[22] Ma, Xin and J. Douglas Willms, "Dropping out of Advanced Mathematics: How Much Do Students and Schools Contribute to the Problem?," *Educational Evaluation and Policy Analysis*, Vol. 21, No. 4, Winter 1999, pp. 365-383.

[23] Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variance," *The Quarterly Journal of Economics*, November 2000, pp. 1239–1285.

[24] Wreglinsky, Harold, "Finance Equalization and Within-School Equity: The Relationship between Education Spending and the Social Distribution of Achievement," *Educational Evaluation and Policy Analysis*, Vol. 20, No. 4, Winter 1998, pp. 269-283.

[25] Cohen, Peter A., James A. Kulik, and Chen-Lin C. Kulik, "Educational Outcomes of Tutoring: A Meta-analysis of Findings," *American Educational Research Journal*, Summer 1982, Vol. 19, No. 2, pp. 237-248.

[26] Good, Jennifer M., Glennelle Halpin, and Gerald Halpin, "A Promising Prospect for Minority Retention: Students Becoming Peer Mentors," *Journal of Negro Education*, Vol. 69, No. 4, Fall 2000, pp. 375-383.

[27] Porter, Andrew C., "National Standards and School Improvement in the 1990s: Issues and Promise," *American Journal of Education*, Vol 102, August 1994, pp. 421-449.

[28] Rowan, Brian, Fang-Shen Chiang, and Robert J. Miller, "Using Research on Employees' Performance to Study the Effects of Teachers on Students' Achievement," *Sociology of Education*, Vol. 70, No. 4, October 1997, pp. 256-284.

[29] Grogger, Jeffrey and Derek Neal, "Further Evidence on the Effects of Catholic Secondary Schooling," *Brookings-Wharton Papers on Urban Affairs: 2000*, pp. 151-201.

[30] Barnes, Robin D., "Black America and School Choice: Charting a New Course," *The Yale Law Journal*, Vol. 106, No. 8, Symposium: Group Conflict and the Constitution: Race, Sexuality, and Religion, June 1997, pp. 2375-2409.

[31] Singh, Kusum, Claire Vaught, and Ethel W. Mitchell, "Single-Sex Classes and Academic Achievement in Two Inner-City Schools," *The Journal of Negro Education*, Vol. 67, No. 2, Spring 1998, pp. 157-167.

[32] Baker, David P., Cornelius Riordan, and Maryellen Schaub, "The Effects of Sex-Grouped Schooling on Achievement: The Role of National Context," *Comparative Education Review*, Vol. 39, No. 4, November 1995, pp. 468-482.

[33] Astin, Alexander W., "Studying How College Affects Students: A Personal History of the CIRP," *About Campus*, July-August 2003, pp. 21-28.

[34] National Center for Education Statistics, National Longitudinal Study of the High School Class of 1972 (NLS-72), Project Officer: Aurora D'Amico (Washington, DC), http://nces.ed.gov/surveys/nls72/.

[35] National Center for Education Statistics, High School and Beyond (HS&B), Project Officer: Aurora D'Amico (Washington, DC), http://nces.ed.gov/surveys/hsb/.

[36] National Center for Education Statistics, The Education Longitudinal Study of 2002 (ELS:2002), Project Officer: Jeffrey T. Owings (Washington, DC), http://nces.ed.gov/surveys/els2002/.

[37] Gamoran, Adam and Eileen C. Hannigan, "Algebra for Everyone? Benefits of College-Preparatory Mathematics for Students with Diverse Abilities in Early Secondary School," *Educational Evaluation and Policy Analysis*, Vol. 22, No. 3. (Autumn, 2000), pp. 241-254.

[38] Sax, Linda J., "The Impact of College on Post-College Commitment to Science Careers: Gender Differences in a Nine-Year Follow-up of College Freshmen," *Proceedings*, *Annual Meeting of the Association for the Study of Higher Education*, Memphis, TN, November 1996.

[39] Smyth, Frederick L. and John J. McArdle, "Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice," *Research in Higher Education*, Vol. 45, No. 4, June 2004, pp. 353-381.

[40] Pascarella, Ernest T., John C. Smart, Corrina A. Ethington, and Michael T. Nettles, "The Influence of College on Self-Concept: A Consideration of Race and Gender Differences," *American Educational Research Journal*, Vol. 24, No. 1, Spring, 1987, pp. 49-77.

[41] Nicholls, Gillian M., Harvey Wolfe, Mary Besterfield-Sacre, Larry J. Shuman, and Siripen Larpkiattaworn, "A method for identifying variables for STEM intervention," *Journal of Engineering Education*, Vol. 96, No. 1, pp. 33-44, January 2007.

[42] Leslie, Larry L., Gregory T. McClure, and Ronald L. Oaxaca, "Women and Minorities in Science and Engineering: A Life Sequence Analysis," *The Journal of Higher Education*, Vol. 69, No. 3, May/June 1998, pp. 239-276.

[43] United States Department of Labor, Bureau of Labor Statistics, *National Longitudinal Survey of Youth 1979* (NLSY79), http://www.bls.gov/nls/nlsy79.htm.

[44] Astin, Alexander W., *What Matters in College?: Four Critical Years Revisited*, Jossey-Bass Publishers, San Francisco, 1993.

[45] Zhang, Guili, Timothy J. Anderson, Matthew W. Ohland, and Brian R. Thorndyke, "Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study," *Journal of Engineering Education*, Vol. 93, No. 4, October 2004, pp. 313-320.

[46] Besterfield-Sacre, Mary, Cynthia J. Atman, and Larry J. Shuman, "Characteristics of Freshman Engineering Students: Models for Determining Student Attrition in

Engineering," *Journal of Engineering Education*, Vol. 86, No. 2, April 1997, pp. 139-149.

[47] Besterfield-Sacre, Mary, Magaly Moreno, Larry J. Shuman, and Cynthia J. Atman, "Gender and Ethnicity Differences in Freshman Engineering Student Attitudes: A Cross-Institutional Study," *Journal of Engineering Education*, Vol. 90, No. 4, October 2001, pp. 477-489.

[48] Larpkiattaworn, Siripen, Obinna Muogboh, Mary Besterfield-Sacre, Larry J. Shuman, and Harvey Wolfe, "Special Considerations When Using Statistical Analysis in Engineering Education Assessment and Evaluation," *Journal of Engineering Education*, Vol. 92, No. 3, pp. 207-215, July 2003.

[49] Adelman, Cliff, "Women and Men of the Engineering Path: A model for Analyses of Undergraduate Careers," PLLI 98-8055, U.S. Department of Education, Office of Educational Research and Improvement, Washington, DC: Government Printing Office, 1998.

[50] Hintze, John M. & Benjamin Silberglitt, "A Longitudinal Examination of the Diagnostic Accuracy and Predictive Validity of R-CBM and High Stakes Testing", *School Psychology Review*, Vol. 34, No. 3, 2005, pp. 372-386.

[51] Brasier, Terry G. "The Effects of Parental Involvement on Students' Eighth and Tenth Grade College Aspirations: A Comparative Analysis," 2008, from an unpublished dissertation for Doctor of Education, North Carolina State University.

[52] U.S. Department of Education. National Center for Educational Statistics. *Entry and Persistence of Women and Minorities in College Science and Engineering Education.* NCES 2000-601, by Gary Huang, Nebiyu Taddese, and Elizabeth Walter. Project Officer, Samuel S. Peng. Washington, DC: 2000.

[53] Mensch, Barbara S. & Denise B. Kandel, "Dropping Out of High School and Drug Involvement", *Sociology of Education*, Vol. 61, April 1988, pp. 95-113.

[54] Civian, Janet Trabucco, "Examining Duration of Doctoral Study Using Proportional Hazards Models", unpublished dissertation, Harvard University Graduate School of Education, 1990.

[55] Willett, John B. & Judith D. Singer, (1991), "From Whether to When: New Methods for Studying Student Dropout and Teacher Attrition", Review of Educational Research, Winter, 1991, Vol. 61, No. 4, pp. 407-450.

[56] Vegas, Emiliana, Richard J. Murnane, & John B. Willett, "From High School to Teaching: Many Steps, Who Makes It?," Teachers College Record, Vol. 103, No. 3, June 2001, pp. 427-449.

[57] Fawcett, Tom, "An Introduction to ROC Analysis," Elsevier B.V, 2005, http://www.csee.usf.edu/~candamo/site/papers/ROCintro.pdf.

[58] Pepe, M.S., "The Statistical Evaluations of Medical Tests for Classification and Prediction," Oxford, UK: Oxford University Press, 2003.

[59] Lasko, Thomas A., Jui G. Bhagwat, Kelly H. Zou, and Lucila Ohno-Machado , "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, Vol. 38, Issue 5 Clinical Machine Learning, October 2005, pp. 404-415, PMID: 16198999.

[60] Hanley J.A. and B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* 1982;143(1):29-36.

[61] Obuchowski, N.A., "Receiver operating characteristic curves and their use in radiology," *Radiology* 2003; 229(1):3-8.

[62] Hosmer, David W. and Stanley Lemeshow, "Applied Logistic Regression 2nd edition," John Wiley & Sons, Inc., New York, NY, 2000, pgs. 160-164.

[63] Nicholls, Gillian M., Harvey Wolfe, Mary Besterfield-Sacre, Larry J. Shuman, and Siripen Larpkiattaworn, (2007), "A method for identifying variables for STEM intervention," *Journal of Engineering Education*, Vol. 96, No. 1, pp. 33-44, January 2007.

[64] Curtin, T.R., Ingels, S.J., Wu, S., and Heuer, R. (2002). National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual (NCES 2002-323). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[65] Ingels, Steven J., Leslie A. Scott, Judith T. Lindmark, Martin R. Frankel, & Sharon L. Myers, National Opinion Research Center (NORC) at the Univ. of Chicago, Shi-Chang Wu, Project Officer, NCES (April 1992), *National Education Longitudinal Study of 1988: First Follow-up: Student Component Data File User's Manual* vol. I. (NCES 92-030). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Education and Improvement. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=92030.

[66] Ingels, Steven J., Kathryn L. Dowd, John D. Baldridge, James L. Stipe, Virginia H. Bartot, Mrtin R. Frankel, National Opinion Research Center (NORC) at the Univ. of Chicago, Peggy Quinn, Project Officer, NCES (March 1995), *National Education Longitudinal Study of 1988: Second Follow-up: Student Component Data File User's Manual*. (NCES 94-374). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Education and Improvement. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=94374.

[67] Ingels, Steven J., Kathryn L. Dowd, John R. Taylor, Virginia H. Bartot, Mrtin R. Frankel, & Paul A. Pulliam, National Opinion Research Center (NORC) at the Univ. of Chicago, Peggy Quinn, Project Officer, NCES (March 1995), *National Education Longitudinal Study of 1988: Second Follow-up: Transcript Component Data File User's Manual* (NCES 95-377). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Education and Improvement. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=95377.

[68] Curtin, Thomas R., Steven J. Ingels, Shiying Wu, Ruth Heuer, (2002), *National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual* (NCES 2002-323). Washington, DC: U.S. Department of Education, National Center for Education Statistics, http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2002323.

[69] Haggerty, Catherine, Bernard Dugoni, Laura Reed, Ann Cederlund, & John R. Taylor, National Opinion Research Center (NORC) at the Univ. of Chicago, C. Dennis Carroll, Project Officer, NCES (March 1996), *National Education Longitudinal Study 1988-1994: Methodology Report*. (NCES 96-174). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Education and Improvement. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=96174.

[70] Curtin, Thomas R., Steven J. Ingels, Shiying Wu, Ruth Heuer, (2002), *National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual* (NCES 2002-323). Washington, DC: U.S. Department of Education, National Center for Education Statistics, http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2002323.

[71] Curtin, Thomas R., Steven J. Ingels, Shiying Wu, Ruth Heuer, (2002), *National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual* (NCES 2002-323). Washington, DC: U.S. Department of Education, National Center for Education Statistics, http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2002323

[72] Nicholls, Gillian M., Harvey Wolfe, Mary Besterfield-Sacre, Larry J. Shuman, and Siripen Larpkiattaworn, "A method for identifying variables for STEM intervention," *Journal of Engineering Education*, Vol. 96, No. 1, pp. 33-44, January 2007.

[73] Adelman, Cliff, "Women and Men of the Engineering Path: A model for Analyses of Undergraduate Careers," PLLI 98-8055, U.S. Department of Education, Office of Educational Research and Improvement, Washington, DC: Government Printing Office, 1998.

[74] Seymour, Elaine & Nancy M. Hewitt, *Talking About Leaving: Why Undergraduates Leave the Sciences*, 1997, Westview Press, Boulder, Colorado.

[75] Smyth, Frederick L. "Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice," 2000, from an unpublished thesis for Master of Arts in Psychology, University of Virginia.

[76] Smyth, Frederick L. and John J. McArdle, "Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice," *Research in Higher Education*, Vol. 45, No. 4, June 2004, pp. 353-381.

[77] College and Beyond Database, http://www.mellon.org/grant_programs/research.

[78] Andrew W. Mellon Foundation, founded 1969, http://www.mellon.org/.

[79] Bowen, William G. and Derek Bok, *The Shape of the River: Long-term consequences of considering race in college and university admissions*, Princeton University Press, Princeton, NJ, 1998.

[80] Astin, Alexander W. and Helen S. Astin, *Undergraduate Science Education: The impact of different college environments on the educational pipeline in the sciences: final report*, Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles, CA, 1992, # ED362404 (http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED362404&ERICExtSearch_SearchType_0=no&accno=ED362404).

[81] Hilton, Thomas L., and J. Hsia, D.G. Solorzano, and N.L. Benton, "Persistence in science of high-ability minority students", Princeton, NJ: Educational Testing Service, 1989.

[82] National Center for Education Statistics, Department of Education, Washington, D.C., http://nces.ed.gov/ .

[83] NSF Division of Science Resources Statistics (SRS), Science and Engineering Statistics http://www.nsf.gov/statistics/nsf08321/content.cfm?pub_id=3785&id=2 .

[84] Smyth, Frederick L. "Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice," 2000, from an unpublished thesis for Master of Arts in Psychology, University of Virginia.

[85] Nicholls, Gillian M., Harvey Wolfe, Mary Besterfield-Sacre, Larry J. Shuman, and Siripen Larpkiattaworn, (2007), "A method for identifying variables for STEM intervention," *Journal of Engineering Education*, Vol. 96, No. 1, pp. 33-44, January 2007.

[86] Klein, John P. & Melvin L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* 2nd edition, Springer-Verlag New York, Inc., New York, NY, 2003, pages 63-74.

[87] Willett, John B. & Judith D. Singer, (1991), "From Whether to When: New Methods for Studying Student Dropout and Teacher Attrition", Review of Educational Research, Winter, 1991, Vol. 61, No. 4, pp. 407-450.

[88] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 236-247.

[89] Klein, John P. & Melvin L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* 2nd edition, Springer-Verlag New York, Inc., New York, NY, 2003, pages 45-46.

[90] Klein, John P. & Melvin L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* 2nd edition, Springer-Verlag New York, Inc., New York, NY, 2003, pages 46-49.

[91] Cox, David R., (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, B34, 187-220.

[92] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 113-184.

[93] Klein, John P. & Melvin L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* 2nd edition, Springer-Verlag New York, Inc., New York, NY, 2003, pages 243-285.

[94] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 113-114.

[95] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 253-257.

[95] Welch, B.L., "The Significance of the Difference Between Two Means when the Population Variances are Unequal," Biometrika, Vol. 29, No. 3/4, pp. 350-362, February 1938.

[96] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 61-109.

[97] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 111-184.

[98] Allison, Paul D., "Survival Analysis Using SAS: A Practical Guide," SAS Institute Inc., Cary, NC, 1995, pages 29-60.

[99] Welch, B.L., "The Significance of the Difference Between Two Means when the Population Variances are Unequal," Biometrika, Vol. 29, No. 3/4, pp. 350-362, February 1938.

[100] Seymour, Elaine & Nancy M. Hewitt, *Talking About Leaving: Why Undergraduates Leave the Sciences*, 1997, Westview Press, Boulder, Colorado, Table 1.1, pg 16 & Appendix A, pg. 399.

[101] Smyth, Frederick L. "Ethnic and Gender Differences in Science Graduation at Selective Colleges with Implications for Admission Policy and College Choice," 2000, from an unpublished thesis for Master of Arts in Psychology, University of Virginia, Appendix A2.

[102] National Science Foundation, Division of Science Resources Statistics, *Classification of Programs*, NSF 07-307, Project Officer Maurya M. Green, Arlington, VA 2007. (available from NSF website http://www.nsf.gov/statistics/nsf07307/content.cfm?pub_id=3634&id=4)