2016

The impact of predictive analytics based advising on the selection and change of major among first-year, first-term students in engineering

Sylvester Charles Upah
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Educational Administration and Supervision Commons

Recommended Citation

https://lib.dr.iastate.edu/etd/15826

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
The impact of predictive analytics based advising on the selection and change of major among first-year, first-term students in engineering

by

Sylvester Charles Upah, Jr.

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Education (Educational Leadership)

Program of Study Committee:

Soko Starobin, Major Professor
Larry H. Ebbers
Janice N. Friedel
Steven A. Freeman
John P. Boysen

Iowa State University
Ames, Iowa
2016

Copyright © Sylvester Charles Upah, Jr., 2016. All rights reserved.
DEDICATION

To

my wife,

Karen,

and our daughters,

Kristin and Katie:

each of you have been a source

of love, encouragement, and inspiration.
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. vi

LIST OF TABLES .................................................................................................................... vii

ACKNOWLEDGMENTS ........................................................................................................... ix

ABSTRACT ................................................................................................................................. x

CHAPTER 1. INTRODUCTION .................................................................................................. 1
  Background of the Study ......................................................................................................... 1
  Problem .................................................................................................................................. 5
  Purpose ................................................................................................................................... 7
  Significance ............................................................................................................................. 8
  Research Questions ............................................................................................................... 10
  Research Design .................................................................................................................... 10
  Conceptual and Theoretical Frameworks .............................................................................. 13
    Self-efficacy theory ............................................................................................................. 14
    Hardness theory .................................................................................................................. 14
    Student integration theory ................................................................................................. 15
  Definition of terms ................................................................................................................. 17
  Summary ................................................................................................................................. 18

CHAPTER 2. LITERATURE REVIEW ......................................................................................... 20
  Theoretical Framework .......................................................................................................... 20
    Self-efficacy theory ............................................................................................................. 20
    Cognitive processes ........................................................................................................... 21
    Motivation processes ......................................................................................................... 21
    Affective processes ............................................................................................................ 22
    Selection processes ............................................................................................................ 23
    Hardness theory .................................................................................................................. 24
    Student integration theory ................................................................................................. 26
  Emergence of Analytics in Higher Education .................................................................... 30
    Academic analytics .............................................................................................................. 32
    Learning analytics .............................................................................................................. 33
    Usage .................................................................................................................................. 33
  Analytics Readiness .............................................................................................................. 34
    Organizational factors influencing data use ................................................................. 37
  Previous Research Methods to Predict Academic Outcomes ........................................... 40
  Summary ................................................................................................................................. 44
CHAPTER 3. METHODOLOGY

Methodological Approach ................................................................. 47
Research Design ................................................................................. 49
Data Sources ....................................................................................... 49
  Mapworks – validity ........................................................................ 50
  Mapworks – reliability ................................................................. 50
Data Access and Security of the Data .............................................. 51
Data Collection ................................................................................. 51
Variables ............................................................................................. 52
  Dependent ....................................................................................... 52
  Independent ..................................................................................... 53
  SIS .................................................................................................. 53
  Mapworks factors and variables .................................................. 54
Sample ................................................................................................. 56
Data Analysis Procedures ................................................................. 56

Research Question 1: What are the demographic characteristics of the subjects of this study? .............................................................. 57
Research Question 2: What are the correlational relationships between the independent variables with the dependent variables and what are the mean differences between the comparison and treatment groups? ...........................................(58
Research Question 3: To what extent did the independent variables and factors predict change of major and the selection of a program of study? ...........................................(59
Assumptions of logistic regression ...................................................... 61
Sample sizes for logistic regression ..................................................... 63
  Assessing and reporting the logistic regression models ................. 63
Research Question 4: What is the impact of receiving academic advising informed by predictive analytics on change of major and the selection of a program of study? .............................................................. 68
Assumptions of propensity score matching .............................................. 69
  Conditional independence .............................................................. 69
  Region of common support .......................................................... 70
  Sample size .................................................................................... 71
Steps in propensity score matching ..................................................... 72
  Computing the propensity score .................................................... 72
  Selecting a matching algorithm ..................................................... 73
  Checking for balance ..................................................................... 76
Reporting results and estimating program effect ............................... 76
Delimitations ....................................................................................... 78
Limitations ......................................................................................... 78
Ethical Considerations ....................................................................... 79
Summary ............................................................................................. 79
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>A preliminary model of institutional action</td>
<td>29</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Conceptual model of first-year, first-term engineering students</td>
<td>30</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Matching methods trade-offs in terms of bias and efficiency</td>
<td>75</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Density plot of propensity scores for Change of Major using SIS and Mapworks variables</td>
<td>102</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Density plot of propensity scores for Change of Major using SIS variables and Mapworks factors</td>
<td>104</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Density plot of propensity scores for Selection of Program using SIS and Mapworks variables</td>
<td>108</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Density plot of propensity scores for Selection of Program using SIS variables and Mapworks factors</td>
<td>111</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.</td>
<td>Dependent variables</td>
<td>52</td>
</tr>
<tr>
<td>Table 2.</td>
<td>Independent variables from SIS</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.</td>
<td>Independent variables and factors from Mapworks</td>
<td>55</td>
</tr>
<tr>
<td>Table 4.</td>
<td>Missing case analyses</td>
<td>61</td>
</tr>
<tr>
<td>Table 5.</td>
<td>Description of covariates</td>
<td>67</td>
</tr>
<tr>
<td>Table 6.</td>
<td>Results based on analysis of descriptive statistics</td>
<td>83</td>
</tr>
<tr>
<td>Table 7.</td>
<td>Correlations between variables (Mapworks variables, entire sample)</td>
<td>84</td>
</tr>
<tr>
<td>Table 8.</td>
<td>Correlations between variables (Mapworks variables, by group)</td>
<td>85</td>
</tr>
<tr>
<td>Table 9.</td>
<td>Correlations between variables (Mapworks factors, entire sample)</td>
<td>86</td>
</tr>
<tr>
<td>Table 10.</td>
<td>Correlations between variables (Mapworks factors, by group)</td>
<td>87</td>
</tr>
<tr>
<td>Table 11.</td>
<td>Mean differences between comparison and treatment groups</td>
<td>88</td>
</tr>
<tr>
<td>Table 12.</td>
<td>Change of major with significant SIS variables and Mapworks variables</td>
<td>91</td>
</tr>
<tr>
<td>Table 13.</td>
<td>Change of major with significant SIS variables and Mapworks factors</td>
<td>93</td>
</tr>
<tr>
<td>Table 14.</td>
<td>Selection of program with significant SIS variables and Mapworks variables</td>
<td>94</td>
</tr>
<tr>
<td>Table 15.</td>
<td>Selection of program with significant SIS variables and Mapworks factors</td>
<td>96</td>
</tr>
<tr>
<td>Table 16.</td>
<td>Probit regression to estimate propensity scores for Change of Major</td>
<td>101</td>
</tr>
<tr>
<td>Table 17.</td>
<td>ATT estimation with bootstrapped standard errors (50 replications)</td>
<td>102</td>
</tr>
<tr>
<td>Table 18.</td>
<td>Probit regression to estimate propensity scores for Change of Major</td>
<td>104</td>
</tr>
<tr>
<td>Table 19.</td>
<td>ATT estimation with bootstrapped standard errors (50 replications)</td>
<td>105</td>
</tr>
<tr>
<td>Table 20.</td>
<td>Probit regression to estimate propensity scores for Selection of Program</td>
<td>107</td>
</tr>
<tr>
<td>Table 21.</td>
<td>ATT estimation with bootstrapped standard errors (50 replications)</td>
<td>109</td>
</tr>
</tbody>
</table>
Table 22. Probit regression to estimate propensity scores for Selection of Program ....... 111

Table 23. ATT estimation with bootstrapped standard errors (50 replications) ............. 112
ACKNOWLEDGMENTS

I would like to express my most sincere thanks to the faculty on my committee who have guided and assisted me on this journey. First, I am most grateful for the guidance of my major professor, Dr. Soko Starobin, whose constant support and enthusiasm served as a source of my determination and perseverance. She has been a model and a mentor, and I will always be grateful to her. I also wish to thank my committee members who have supported me throughout my years in the program: Drs. Larry Ebbers, Janice Friedel, Pete Boysen, and Steve Freeman. Each provided me with learning opportunities and helpful advice along the way. A special thanks to Dr. Ebbers for initially encouraging me to enroll in the program.

Thank you to my editor, Pat Hahn, who edited this dissertation on a compressed timeline. Her willingness to respond quickly and help me meet target dates was greatly appreciated.

Finally, I would like to thank my family: my wife, Karen, who has supported and encouraged me in countless ways; my daughters, Kristin and Katie, who have supported me with their love and concern; and my parents who instilled the importance of education in my values. Finally, thanks to my great uncle’s son, James Balloun, who has been instrumental in supporting my goals as well as sponsoring multiple other professorships at Iowa State University.
ABSTRACT

The purpose of this study was to investigate the relationship between demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables on undergraduate outcomes (i.e., change of major and selection of program of study) as first-year, first-term students transition into the postsecondary environment. The students in this study were freshmen majoring in engineering at a large Midwest research-1 institution with no selected program of study. In the fall of 2015, students received academic advising informed by the use of predictive analytics (i.e., the treatment). The objective was to explore the variables which influence change of major and the selection of program of study and also the impact of academic advising informed by predictive analytics on change of major and the selection of a program of study. The independent variables were: age, gender, ethnicity, ACT composite score, ACT math score, high school rank, number of first-term credits attempted, number of first-term credits completed, first-term cumulative GPA, honors program membership, and learning community memberships. The environmental variables (i.e., level of commitment to completing degree, self-assessment of math skills, self-assessment of time management and planning, academic life satisfaction, academic self-efficacy that one can do well on challenging tasks and in hardest course), and environmental factors (i.e., academic self-efficacy, academic integration, social integration) were derived from the Mapworks® transition survey instrument. Logistic regression models were constructed to determine which variables were significant in predicting outcomes and to inform which variables were selected for use in propensity score analyses to determine the impact of the treatment on student outcomes.
CHAPTER 1. INTRODUCTION

Background of the Study

As recently as 1995, the United States ranked 1st in the world in four-year degree attainment among 25-34 year olds. Since that time, the U.S. has slipped dramatically. According to a 2014 report from the Organization for Economic Cooperation and Development (OECD), the U.S. ranked 12th among 36 countries. The report also identified a disturbing trend in educational attainment, with only 30% of the 25-64 year-old non-students having attained a higher level of education than their parents. In addition, the earning premium for tertiary-educated 25-65 year-olds declined as well with an earning premium of 86% in 2005 compared to 74% in 2012. Finally, the report alluded to the change in the funding model for higher education institutions that has taken place, pointing out that, as of 2011, 65% of the cost of tertiary education in the U.S. was attributed to private funding.

President Barack Obama has emphasized the need for an educated society as a means to an engaged democracy (Kanter, Ochoa, Nassif, & Chong, 2011). In 2013, President Obama signed an executive order to open and make machine-readable data the new default for government information. Making information about government operations more readily available and useful has been a central promise of a more efficient and transparent government. As a result, the Obama administration launched a number of Open Data Initiatives aimed at scaling up open data efforts across many areas, including education, with the goal of helping students make the best decisions in regards to their educational pursuits. An outcome has been the creation of a simple, straightforward college ratings system. The College Scorecard rating system was created, which centralizes pertinent government data from multiple sources (Department of Education, the Treasury Department, census data, and
other federal agency data) in one location and makes it simple for prospective students to search colleges based on criteria like cost, salaries of graduates, and graduation rates.

The use of data in this manner, namely, to identify patterns, trends, and provide insights to guide actions and decision-making aptly describes a relatively new field of study—analytics. A key prerequisite of analytics is the ability to collect and integrate of data from multiple, varied sources (i.e., data mining). Analytics has been shown to be a valuable asset in business and health-related fields to identify patterns, predict outcomes, and guide decision-making (Hersh, 2002; Ngai, Xiu, & Chau, 2009). Although the use of analytics has experienced considerable growth in the aforementioned fields, it has been claimed that the use of analytics in education is in its infancy (Tinto, 2014). Reinforcing this view, a recent EDUCAUSE survey revealed that a majority of higher education institutions are collecting data, but not using these data for predictive reasons or to guide decision-making (Bichsel, 2012). As more data has become readily available in recent years and more complex models and technologies have been developed, this has led to the introduction of predictive analytics. An area where predictive analytics have been applied in higher education is to the prediction of future academic performance of students based on past student performance data. However, it is not clear how the information gleaned from these analytics can best be used to guide student decision-making.

The extent to which students are able to formulate, adjust, and make decisions regarding their academic environments can be linked to the theoretical construct self-efficacy (Bandura, 1993; Zimmerman, Bandura, & Martinez-Pons, 1992), hardiness theory (Kobasa, 1979; Maddi, 2004), and academic integration theory (Tinto, 1975, 1993). Bandura (1993) focused on self-regulatory processes that influence human development and theorized that
the manner a person approaches goals, tasks and challenges is a representation of his or her
own self-efficacy. Self-efficacy manifests its influences in cognitive, motivational, affective
and selection processes, with each contributing to academic development and
accomplishments. Maddi (2004) operationalized hardiness as a psychological construct with
three sub-components: control, challenge, and commitment. Much like self-efficacy, the
hardiness constructs may be viewed as the ability to thrive under stressful conditions and
control one’s own destiny. Finally, Tinto (2014) emphasized the importance of goal
commitment, academic integration, and intellectual development. Taken together, these
theories illustrate the various perspectives from which a student may vision the choice of
major or selection of a program. When a student is committed to an academic goal (i.e.,
attainment of a degree) and is willing to persevere through challenges, he or she is likely to
succeed; however, not all students may have attained this level of efficacious intellectual
growth. In such cases novel advising tools, such as predictive analytics, can play a role and
shed insights that influence student decision-making and possibly enhance student success.

Seeking to gain a better understanding of the importance of choice of major or
selecting a program of study, Leppel (2001) examined the role of choice of major on college
persistence among undecided freshmen in business and engineering. Using a national dataset
and regression models, Leppel’s study revealed that students with undecided majors were
significantly less likely to persist to the second year of college, regardless of gender. Leppel
also noted that undecided majors also had significantly lower GPAs. Similarly, Kreysa
(2006) found that declaring a major increased the likelihood of persistence by 22%, leading
to the conclusion that students with well-defined career goals are able to choose a major from
the outset and more likely to be retained. Supporting the importance of deciding on a major,
Sandler (2000) found that persistence at the postsecondary level was improved by a student’s confidence in his or her ability to make appropriate career-related decisions beginning with the choice of major.

Social factors also have a tendency to play a role in the choice of major. For example, the “college experience” has been shown to influence choice of a major (Cohen & Hanno, 1993; Mauldin, Crain, & Mounce, 2000). Such experiences require time to develop over multiple semesters and may be inconsistent from person to person. Delaying decisions until later in a student’s college career can be significant in terms of the time of completion and financial liability. Environmental and personality alignment have also been studied as factors influencing choice of major (Malgwi, Howe, & Burnaby, 2005; Porter & Umbach, 2006). Measures such as these focus mainly on pre-college dispositions. Alternatively, as more data about an individual’s academic performance become known during their first semester, especially in comparison to those who have come before them, it may lead students to consider switching to other majors or digging in deeper to their selected program of study. Thus, predicting future academic performance can have a major effect on student decision-making.

The impact of perceived academic performance was shown to have a profound effect in a study conducted by Arcidiacono (2004), who asked students to estimate their cognitive ability in specific majors. Arcidiacono determined that choice of major was influenced mostly by how students perceived their ability to complete coursework in the major they had chosen. Although exploring the relationship between ability and coursework was useful, the results relied largely on self-reported responses and, therefore, were likely subjective insofar as participants interpreted their own perceptions.
While the attitudes and feelings perceived by students and developed in their social environments are part of the overall picture, the inclusion of academic performance data (revealed through predictive analytic dashboards and reports) is one of the few pieces of early evidence that can be used by first-term college students to evaluate their choice of major or program of study. The current study evaluated the impact that the use of predictive analytics has on the propensity of students to select a program study or change of major and their first-term academic performance. According to Tinto (2006):

Knowing why students leave do not tell us, at least not directly, why students persist. More importantly it does not tell institutions, at least not directly, what they can do to help students stay and succeed. In the world of action, what matters are not our theories per se, but how they help institutions address pressing practical issues of persistence. (p. 6)

The intent of this study was to provide an explanation of student outcomes (change of major or selection of program of study) in relationship to the demographics, pre-college academic characteristics, and first-term engagement and completion characteristics as well as environmental variables associated with first-year students in their first term of transition to higher education. A second intent was to provide insights and implications to administrators as to how student outcomes were impacted by the use of predictive analytics.

**Problem**

The selection of a college major is often viewed as one of the most important decisions an individual can make. Many freshmen enter postsecondary education with an undeclared major for various reasons. Some researchers have argued that it is not detrimental to delay on the decision of a major (Cueso, 2005). However, others have posited that attrition increases noticeably among undeclared students because of a delay in committing to goals (Pascarella & Terenzini, 1980; Cabrera, Castaneda, Nora, & Hengstler,
Even after a full initial year of academic experience, many students may still be unable to make a well-informed decision about their choice of a major and fallback on subjective, anecdotal information and personal preference (Malgwi, Howe, & Burnaby, 2005). Holland’s (1995, 1985) theory of careers, which relates a person’s choice of environment to preferred activities, interests, and competencies, has been often referenced in regards to the selection of a college major. Other researchers have shown that cultural and socio-environmental variables and factors influence the selection of a major, such as parental influence (Astin, 1993; Simpson, 2001), gender differences (Montmarquette, Cannings, & Mahseredjian, 2002; Turner & Bowen, 1999; Zafar, 2009), race (Porter & Umbach, 2006; Arcidiacono, Aucejo, & Spennier, 2012), and expected income (Berger, 1988; Boudarbat, 2008; Montmarquette, Cannings, & Mahseredjian, 2002).

The use of academic performance data, reflected in the form of predictive analytics, has the potential to play a considerable, positivistic role in the selection of major or program of study. Failure to perform well in first-year courses that are early markers for a particular field of study may contribute to a student changing majors several times. Repetitive, successive changes can lead to anxiety, failure in committing to the goal of obtaining a degree, and ultimately, to withdrawal. This situation creates academic advising challenges (Cuseo, 2005; Gordon, 2007). Thus, it is critical to monitor the early academic performance of students, particularly those undecided about a major. Predictive analytics can assist by providing a means to compare a student’s present academic performance with the performance of past students who have taken similar courses and pursued similar majors. Advisors using analytics can project future outcomes and proactively guide students to make
better decisions regarding their selection of a major or coursework pathways, improving a student’s likelihood of attaining a degree and consequent success in a given field. Due to the technological advances in the fields of educational data mining and use of analytics, the capacity to guide students using rational, evidence-based information is now a possibility.

A portion of the previous research has emphasized the importance of socio-environmental factors or while others have focused on academic performance indicators. Both areas are equally important. Hence, this study explored variables from both realms to more holistically determine which variables influence student outcomes in terms of choice of major and selection of program of study.

**Purpose**

The purpose of this study was to investigate the relationship between demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables and factors in predicting change of major and selection of program of study. A second goal was to measure the impact of academic advising informed by the use of predictive analytics to advise first-year, first-term freshmen majoring in engineering without a selected program of study in engineering. Understanding whether predictive analytics influences student tendencies in regards to change of major and selection of program of study is critical for leaders and policymakers. The results of this study will contribute to the body of research on the use of predictive analytics in higher education. This study employed a quasi-experimental research design using observational data and quantitative methods.
Significance

Research on the use of predictive analytics for the purpose of improving postsecondary student outcomes is important for several reasons. First, predictive analytics provide academic advisors with a tool that provides an evidence-based profile of a student performance. By using analytics, advisors can provide more timely support and guidance to their student advisees, especially during the first transitional semester into the postsecondary environment. Second, analytics can inform administrators by providing insights on whether incoming students are succeeding within a particular program of study or major, and make adjustments to the curriculum to improve retention and outcomes. Third, predictive analytics can provide institutional leaders with evidence to demonstrate to legislators and other external stakeholders the effectiveness of departments and programs within the institution for suggesting efficient pathways to students that can reduce time to degree attainment, thereby lowering institutional costs and student debt.

The institution in this study had previously recognized the importance of analytics through internally developed statistical models to predict academic performance. In addition, the Mapworks® transition survey instrument (Skyfactor, 2014) was being used to collect data on the support-need patterns of freshmen students in transition. As data collection and vendor-based analytics systems in higher education have matured in robustness and ease of use as well as their ability to monitor and predict, the academic progress of individual students has grown in importance. The institution in this study had purchased the use of a vendor-hosted predictive analytics systems. The analytics from this system delivers real-time information to administrators, advisors, and students through the use of browser-based dashboards and e-mail notifications. An institutional dashboard calculates the graduation
probability for each student. A program-based dashboard displays the distribution of at-risk students across academic programs. Finally, an advisor dashboard presents the academic progress and level of risk of each student.

Nevertheless, the impact of using predictive analytics within the institution was not entirely clear. There existed a need to understand and measure the impact of predictive analytics on student success by defining and investigating outcomes. This study posited that change of major and the selection of a program of study were two important measurements of student outcomes that analytics could impact. Through scholarly dissemination, this study will contribute to efforts to define policy and clarify actions that postsecondary leaders must consider when introducing predictive analytics into the academic environment that will result in positive consequences on persistence and student success.

Given the increase in the amount of data and the continued threat of declining state funds, institutions have begun to notice the value of utilizing their data archives (i.e., big-data) as a means to improve student success and enhance the effectiveness of the institution. Analytics is an applied use of such data to demonstrate institutional efficiency and effectiveness. Institutions anticipate the use of predictive analytics will have long-term benefits in areas such as improved completion and retention rates. Such outcomes not only reflect improved student outcomes but also have monetary implications for institutions. Other benefits might include the ability to make more timely decisions regarding where to reallocate resources to other higher-value efforts. While institutions have become increasingly data-driven, they should not lose sight of their mission to serve students and help them achieve their goals as they formulate their self-identities and make decisions during their first-term.
Research Questions

The following research questions guided this study. A quantitative approach was taken to explore relationships between student outcomes and variables that influence those outcomes as well as the impact of the use of predictive analytics on student outcomes:

1. What are the demographic characteristics of the subjects in this study?
2. What are the correlational relationships between the independent variables with the dependent variables and the mean differences between the comparison and treatment groups?
3. To what extent do the independent variables and factors predict change of major and the selection of a program of study?
4. What is the impact of receiving academic advising informed by predictive analytics on change of major and the selection of a program of study?

Research Design

This research utilized a quasi-experimental design with observational data collected spanning multiple years for the selected population. According to Shadish, Cook, and Campbell (2002), a “…key feature common to all experiments is to deliberately vary something so as to discover what happens to something else later—to discover the effects of presumed cause” (p. 3). Quasi-experimental designed studies endeavor to test causal hypotheses through the use of control groups and pretreatment measures leading to a counterfactual inference about the impact of a particular treatment. By definition, a quasi-experimental design lacks random assignment. Assignment to the treatment is the result of self-selection or at the discretion of an administrator (i.e., others decide which persons should receive which treatment). However, researchers who use quasi-experimental design still
retain control over selecting and scheduling measures, how nonrandom assignment is executed, and the creation of comparison or control groups by which a treatment group is compared. Another distinguishing feature of non-experimental designs is the use of observational data. Whereas traditional, cross-sectional studies usually collect data on all respondents at a single point in time, this is often difficult to accomplish in practice due to ethical or practical reasons. Shadish et al. (2002) also detailed a number of additional concerns that have been raised regarding conducting randomized experiments, including the limited generalizability of findings. In non-experimental designs such as this study, reliance is placed on measuring multiple, alternate models and statistically controlling for them.

With this perspective in mind, the observational data this study utilized included demographic, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental experiences of first-year, first-term students majoring in engineering with no selected program of study. This data enabled the researcher to investigate: (a) the relationships between the independent variables and change of major and selection of program of study; (b) how variables and factors predict change of major and selection of program of study; and (c) how the variables and factors can be used to measure the impact of academic advising informed by the use of predictive analytics.

The independent variables in this study were collected from two data sources at the institution where the students were enrolled. The first data source was the student information system (SIS) at the institution. The student information system maintains information on every student enrolled at the institution. The data are updated at the beginning of each academic semester to include new students admitted to the university. Data from previous years are archived. This data contained student demographic
characteristics, pre-college academic characteristics, and first-term academic engagement and completion characteristics.

The source for the second dataset was the response data collected from a survey (i.e. Mapworks) administered to first-year, first-term freshmen at the institution regarding their transition experiences and relationships. The Mapworks dataset uncovers variables that describe self-assessed feelings, beliefs, and attitudes regarding the social-environmental experiences of students as they transition through their first semester of matriculation. Participation rates in the Mapworks® transition survey have exceeded 80% since 2012. To form the comparison group in this study, the data from each data source were merged and matched based on a common key (i.e., university ID) for the 2012-2013, 2013-2014, and 2014-2015 academic years. For the treatment group, the same matching procedure was used for the 2015-2016 academic year.

The students selected for this study were first-year, first-term freshmen majoring in engineering without a selected program of study (i.e., undeclared engineering major). In order to control for confounding variables based on the previous academic experience of the subjects, the sample was limited to full-time students who were 18 years of age or younger. In addition, only students classified as first-year students in the fall and spring semesters were selected as case subjects for subsequent analyses. The dependent variables were change of major (i.e. whether a student transferred out of engineering altogether) and selection of program of study (i.e. whether a student selected a particular engineering program after their first semester).

Descriptive and inferential statistics were used to explore the relationships between the demographic, pre-college academic characteristics, first-term academic engagement and
completion characteristics, and environmental variables and factors regarding change of major and selection of program of study. Descriptive statistics were used to answer Research Question 1. Correlational analyses and t-tests were used to investigate Research Question 2 and reveal variable interrelationships and group differences. Logistic regression analyses were used to investigate Research Question 3 and explore the relationships and influence of independent variables on change of major and selection of program of study (i.e., the dependent variables). The analyses conducted on these questions informed the choice of variables used in the propensity score analyses. Logistic regression enables researchers to develop an understanding about the independent variables that predict an outcome and a goodness-of-fit of a model. Finally, propensity score analysis was an appropriate method to use to determine the impact of the treatment as posed in Research Question 4.

Conceptual and Theoretical Frameworks

Astin’s (1993) Input-Environment-Output provided a conceptual context for this study. As students develop their self-concept during their first postsecondary term, based on their previous K-12 academic career, they must make internal adjustments to the new academic environment in which they are immersed. In this study, the use of predictive analytics in academic advising was added to the academic environment. Bandura’s (1993) self-efficacy construct, Kobasa’s (1979) and Maddi’s (2004) hardiness framework, and Tinto’s (1975, 1987) student integration theory provided a theoretical foundation of the academic environment. These theories explain and describe the cognitive, academic, and social dynamics through which the students must navigate during their first-term as they acclimate to their academic environments and make decisions regarding their futures.
Self-efficacy theory

Educational psychologists and researchers such as Bandura (1993) and Eccles (1987) focus heavily on self-referent influences by which personal agency is exerted (Bandura, 1986). According to Bandura (1992), these influences affect the selection and construction of environments, and are mediated through cognitive processes. Bandura posited that, foremost among internal processes, is a person’s beliefs about their capabilities to exert control over their own level of functioning and over events that affect their lives. Bandura called this level of control self-efficacy. The extent to which a person perceives their self-efficacy in particular circumstances influences how they feel, think, motivate themselves, and behave. Self-efficacy is manifested in four internal processes: cognitive, motivational, affective, and selection.

Self-efficacy in college students is observed in how they respond to challenges, their conception of their ability, social comparison influences, framing of feedback, perceived controllability, cognized goals, self-reaction to goal progress and attainment, proactive control of motivation, anxieties, career choices, and choice of activities and environments.

Hardiness theory

According to Maddi and Khoshaba (2001), hardy individuals are those who construct meaning in their lives by recognizing that everything they do constitutes a decision. Decisions invariably involve pushing towards the future and the anticipated search for meaning. However, choosing the future may create stress and anxiety because of the unpredictable nature of things not yet experienced. To accept this anxiety requires an inner strength and determination (i.e., a manifestation of existentialist theory) to meet these challenges and succeed in these situations. Hardiness theorists submit that people who feel
committed, in control, and positively challenged by life circumstances have the tendency to perceive events or circumstances less stressfully, viewing them as manageable occurrences rather than overwhelming misfortunes (Khoshaba & Maddi, 1999).

When hardiness characteristics are present in a college student, these characteristics are reflected in the student’s ability to manage stressful circumstances such as taking examinations, managing difficult coursework and completing challenging projects on time, and turning these circumstances into opportunities for growth. Although not directly measured as hardiness factors, elements of these characteristics are reflected in the various questions within the Mapworks survey.

**Student integration theory**

Tinto (1993) posited that the college environment is divided into academic and social domains, noting that students should ideally feel a sense of belonging or membership in both arenas. He also introduced a model that emphasizes the need for congruency between student motivation and academic performance, viewing academic achievement as a reflection of the degree to which a student has been integrated into his or her academic and social domains. More importantly, focusing on the support aspect of achievement, Tinto (2005) posited, “The inability to obtain needed advice during the first year or at the point of changing majors can undermine motivation, increase the likelihood of departure, and for those who continue, result in increased time to degree completion” (p. 3). Although students may make credit progress, they do not make substantial degree-credit progress.” In his earlier research, Tinto (1975) emphasized the link between student goal-commitment and persistence, stating that when a student has more highly defined educational plans, educational expectations, or career aspirations; the student is more likely to remain in
college. Therefore, the lack of a selected program of study or a major should be given special attention. Students should receive support to develop a plan early in their academic career.

Focusing on the social domain, Berger and Milem (1999) submitted that students who do not experience high levels of institutional commitment or a sense of belonging are less likely to become involved and less likely to persist. Their research findings indicated that it is important to identify these students very early in their first year to be able to help them get involved in some aspect of academic or social life. The researchers’ study assessed student commitment by asking participants if they felt it was important to graduate and whether they were confident in their selection of the institution they were attending. Consistent with these findings, Jenkins and Weiss (2011) found that low-income students who did not select a program or major encountered academic problems more frequently. In their sample of first-time college students, the average course pass rate overall among students who did not “enter a concentration” was 49% as compared to a course pass rate of 87% for those who did select a program of study.

The adoption and implementation of predictive analytics by an institution demonstrates an institutional commitment that is central to improving persistence and student success. Yet, understanding the impact of such a new program is equally important to administrators and stakeholders. Logistic regression models were constructed to determine which variables and factors are significant in predicting the selection of a program of study or change of major (i.e., the dependent variables). Based on student information system data, the independent variables utilized were: age, gender, ethnicity, ACT composite score, ACT math score, high school rank, number of first-term credits attempted, number of first-term...
credits earned, learning community memberships, and honors program membership. In addition, environmental variables and factors related to academic integration, social integration, and academic self-efficacy were obtained through data collected from the Mapworks® transition survey instrument that all students are asked to complete early in their first term. Propensity score analyses were then conducted to measure the impact of the treatment (i.e., those students receiving academic advising informed by use of predictive analytics) on the outcomes (i.e., the dependent variables, selection of program of study, and change of major).

**Definition of Terms**

The following key terms deemed pertinent to this research were defined for use in this study:

*Analytics*: The science of collecting and examining data for patterns and converting those patterns into actionable insights that can be used to trigger alerts and inform decision-making processes and courses of action to plan for the future (Picciano, 2014).

*Analytics (Academic)*: Historical or real-time data in the form of indicators that reflect how a higher education institution and its units (colleges, schools, and departments) are performing (van Barneveld, Arnold, & Campbell, 2012).

*Analytics (Learning)*: The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Long & Siemens, 2011).

*Analytics (Predictive)*: A form of academic analytics that serves various levels of higher education and business serving as a connector between data collection, guiding intelligent action based on analysis, and thereby informing decision-making (van Barneveld, Arnold, & Campbell, 2012).
**Dashboards:** A visual display of big data in the form of indicators that allow the interpretation of performance (van Barneveld, Arnold, & Campbell, 2012).

**Data driven decision-making:** The systematic collection, analysis, and interpretation of data to inform practice and policy in educational settings (Mandinach, 2012).

**Educational data:** Data that enable educators to know more about the needs of student that can be codified to facilitate systematic analysis (Wayman, 2013).

**Educational data mining:** The extraction and development of computerized models to detect patterns in large collections of educational data systems and to convert the raw data into useful information that can potentially have a great impact on educational research and practice (Romero & Ventura, 2010).

**Summary**

The purpose of this study was to explore and develop statistical models that explain change of major and the selection of program of study. In addition, the study examined the impact of academic advising informed by the use of predictive analytics on change of major and selection of program of study.

Chapter 2 provides an outline of related research on theories related to the transitional experiences of college students. It begins with an overview of each theoretical framework: Bandura’s self-efficacy theory, Maddi’s hardiness framework, and Tinto’s integration theory. The overview is followed by a discussion of the emergence of analytics in higher education in response to the need for evidence-based decision-making and efforts to determine analytics readiness. It concludes with a discussion of previous efforts regarding prediction factors influencing academic performance.

Chapter 3 describes the methodology and methods used in designing and conducting the study. It includes the methodological approach, data sources, and data analysis procedures. The methods include: (1) the descriptive characteristics of first-year, first-term engineering freshmen with no selected program of study (i.e., undeclared engineering major);
(2) a correlational analysis of the variables and a $t$-test comparing the control and treatment groups; (3) a multivariate logistic regression to assess the predictive strength of various variables within a particular covariate; and (4) a propensity score analysis to measure the treatment effect (i.e. academic advising informed by use of predictive analytics) on student outcomes (i.e., change of major and selection of program of study).

Chapter 4 presents the results of this study using the methods described in Chapter 3. Chapter 5 concludes with an overview of the results, limitations, and implications for policy and practice, as well as recommendations for future research.
CHAPTER 2. LITERATURE REVIEW

The literature review focuses on: the four themes, the theoretical frameworks; the emergence of analytics in higher education in response to the need for evidence-based decision-making; efforts to determine analytics readiness, and a discussion of previous efforts regarding predictive variables and factors influencing academic performance. The first section focuses on theoretical frameworks related to the issue of student success and factors that are germane to student success. It includes Bandura’s self-efficacy theory, Maddi’s hardiness framework, and Tinto’s integration theory. The second section focuses on the emergence of analytics in higher education as a basis for improving postsecondary student success. It is followed by a review of the literature focused on evaluating analytics readiness within the postsecondary environment. The emphasis of the fourth section is on previous studies that have helped to shape and direct analytics research in education. Studies are described in terms of the data, methods, and variables used in order to provide insight for the methods selected in this study. A literature map for this chapter is provided in Appendix A.

Theoretical Framework

Self-efficacy theory

The basic premise of self-efficacy is that as strength in one’s perceived self-efficacy rises, the higher the goal challenges people set for themselves and the firmer their commitment (Bandura, 1991) in attaining them. According to Bandura, self-efficacy is rooted in four mental processes: cognitive, motivational efforts, affective reactions, and
selectivity of environments. These processes are particularly important as students enter their transitional period during their first-term in college and is discussed next.

**Cognitive processes**

Bandura (1991) asserted that people who have a high sense of efficacy visualize successful scenarios and outcomes that provide guides and supports for performance. When people doubt their efficacy, they tend to visualize failure, making it difficult to achieve. Those who regard ability as an attainable skill, seek challenges and view errors as opportunities to learn and, therefore, are not as easily stressed by difficulties. However, students often view their academic performance as a measure of their intellectual capacities and a judgment that they may lack intelligence. In such cases, they gravitate towards tasks that minimize errors and demonstrate proficiency rather than expanding their knowledge. In addition, in academics it is inevitable to be compared with others. Thus, self-esteem may be diminished as comparisons are made with other students based on grades. As Bandura pointed out, seeing oneself surpassed by others undermines self-efficacy and impairs performance attainment while seeing oneself progress strengthens efficacy. Additionally, feedback should focus on achievement progress, rather than on shortfalls to strengthen personal capabilities and improve self-esteem. In summary, Bandura recommended that educators should focus on portraying ability as an acquirable skill, deemphasizing social, competitive-like comparisons, and highlighting personal progress and accomplishments.

**Motivation processes**

Bandura (1975) asserted that self-efficacy plays a key role in motivation. People with high self-efficacy attribute failure to lack of sufficient effort whereas those with low self-efficacy regard failure as a reflection of their low ability, viewing outcomes in terms of
expectancy-value theory where certain behaviors will lead to certain outcomes. As a result, people formulate goals they believe will lead to their self-satisfaction when fulfilled. However, people are motivated by their internal beliefs of how likely goal attainment may be, based on whether they consider themselves to have the capabilities to achieve the desired outcome. Based on their level of self-efficacy, people may react differently to goal attainment by intensifying their efforts to achieve them or by readjusting their goals. Self-efficacy contributes to these motivational processes by determining which goals are set, the effort put forth, degree of perseverance, and the resiliency to setbacks. Finally, Bandura asserted those with a strong sense of self-efficacy are motivated by a continuous cycle consisting of “discrepancy production” (setting higher goals) and “discrepancy reduction” (working to achieve those goals), where performance is driven by the need to close the gap between goals and those not attained.

**Affective processes**

Bandura (1975) described affective processes as the emotional mediators of self-efficacy. These processes include the ability to exert control over stresses and anxieties. People who perceive themselves as having high self-efficacy are able to exert control over their emotions, manage threats, and reduce anxieties rather than succumbing to them. The inability to regulate or control thought processes related to stress is a key factor in producing avoidant behaviors and depression. In addition, Bandura stated that students with a low self-efficacy are unable to manage academic demands and are especially susceptible to achievement anxieties and depression. In contrast, students with higher self-efficacy believe in their capabilities to do well in subjects and are less likely to allow their academic performance to be impaired. Bandura stated that, in order to lessen academic anxieties,
students should be encouraged to develop stronger self-efficacy by promoting self-regulative skills to manage academic task demands and diminish thoughts that foster feelings of stress. In addition, he pointed out that teachers who lack efficacy, themselves, may manifest anxiety in the classroom, leading to the creation of conditions that weaken their own students’ self-efficacy.

**Selection processes**

The last process Bandura (1975) described is related to the selection of activities and environment. According to Bandura, people avoid activities and environments that exceed their capabilities, but tend to select activities and situations they judge which they believe they handle. Through such choices, people begin to embrace certain skills, interests, and social networks that affect direct the course of life. One example that Bandura pointed out is the choice of career. The higher a person’s self-efficacy the more career options they consider and the more they prepare themselves educationally for that career path and the more resiliency they have in staying with that choice in difficult or challenging times.

To summarize within the context self-efficacy, the importance of STEM (Science, Technology and Engineering) majors in society has been noted and associated with national economic needs and ability to compete globally (Kanematsu & Barry, 2016). Self-efficacy is a well-established theoretical construct that has spurred the development of social cognitive theory and been investigated as a construct contributing to the progress of undergraduate student populations in STEM majors (Pajares, 1996). Csikszentmihalyi (1990) stated the ideal level of self-efficacy should be slightly above ability in order to keep people engaged with challenging tasks and gaining experience. Throughout the research, a theme was that the development of self-regulatory skills should better enable students to control processes
related to their motivation, affect, and social function in academic environments. Self-efficacy variables were included in the current study, derived from the Mapworks® transition survey responses. This researcher explored the extent self-efficacy variables relate to student tendencies in regards to change of majors and selection of program with first-year, first-term freshmen in engineering.

**Hardiness theory**

Kobasa (1979) posited that hardiness could be described as a personality trait, where hardy persons tend to possess an internal locus of control that enable them to view stress and change as a challenge and a path to growth and achievement. The concept of hardiness has been shown to positively correlate to the persistence of students in higher education (Lifton, Seay, McCarly, Olive-Taylor, Seeger, & Bigbee, 2006) and academic success (Sheard, 2009; Sheard & Golby, 2007). Maddi (2004) operationalized the hardiness model by dividing it into three distinct constructs: commitment, control, and challenge. Maddi (2004) defined the construct of commitment as the level of “involvement with others and events of ongoing activities” (p. 290). Commitment can be interpreted to mean that individuals have a purpose and a feeling of being deeply involved in the decisions and activities in their life.

Maddi (2004) described the construct of control as “the sense that one has chosen and had influence over activities” (p. 290) and outcomes within one’s realm of authority. In the area of student development, Hodge, Baxter Magolda, and Haynes (2009) posited that students who are engaged within a domain eventually move away from authoritarian dependence to self-authorship, evaluating information critically to form judgments and collaborating with others as a way of acting intelligently. The objective (Hodge, et al., 2009) is to move students from an authoritarian model that “tells students what they need to know”
(p. 18) to a discovery model that encourages students to discover and synthesize new knowledge by taking greater ownership over their learning through self-directed efforts to more genuinely explore questions they may have about certain topics. Similar to Magolda’s (1999a & b) model of self-authorship and Tinto’s (1975) concept of intellectual development, students become active participants in controlling their own growth and development, continually seeking to transform underlying mental structures to construct meaning as their experiences become internalized.

Maddi’s (2004) last hardiness construct is challenge. Maddi stated that this construct focuses on the “process of continuing to learn from your experiences whether they are positive or negative, is developmentally fulfilling” (p. 290). The construct of challenge is obvious in education. Academic challenge is one of the five main benchmarks in Community College Survey of Student Engagement (McClenny, Marti, & Adkins, 2007) where it is defined as “the extent to which students engage in challenging mental activities and the quantity and rigor of their academic work” (p. 4). Surveys like the National Survey of Student Engagement and the Your First College Year Survey from the Higher Education Research Institute (HERI) at UCLA also contain several questions focusing on academic challenge.

Hardiness has been found to be critical to persistence in higher education and a better predictor of retention than either scores on the Scholastic Aptitude Test (SAT) or high school rank (Lifton, Seay, & Bushko, 2004). The constructs of hardiness as predictive factors on academic performance were investigated by Sheard (2009), whose findings revealed a significant positive correlation between commitment and academic achievement in areas such as GPA and dissertation scores.
As students are presented with academic performance evidence revealed through the use of predictive analytic dashboards, their capacity to make informed decisions regarding their future academic path is augmented, possibly enhancing their sense of commitment and control over their lives. Thus, the use of predictive analytics may stimulate the development of the control construct reflected in decisions such as a change of major or the selection of program of study, thereby fostering a deeper form of commitment to their degree.

In addition to the psychological constructs related to hardiness, it has been widely accepted that, as students encounter various academic and social aspects during their first year, they become more acclimated to the college environment (Astin, 1993; Tinto, 1993). The concept of academic integration is reviewed in the following section and provides a perspective on its role as students enter the postsecondary environment.

**Student integration theory**

A basic element of Tinto’s (1975) theoretical model is that of academic integration. Academic integration occurs when students become committed to the intellectual life within their college experience. Similarly, social integration occurs when students create relationships and connections outside of the classroom. Tinto posited that, while students must be integrated along both dimensions to increase their likelihood of persistence, they need not be equally integrated in each. Focusing on academic integration, Tinto stated that it can be measured in terms of both grade performance and intellectual development. Grade performance relates to meeting explicit standards of the academic system. Grades represent “an extrinsic form of reward of the person's participation in the college…utilized by persons as tangible resources for future educational and career mobility” (p. 104). Viewed from a self-efficacy perspective, grade performance is a form of goal attainment.
Another component of academic integration Tinto (1975) noted is that of intellectual development. Intellectual development is an internal, fundamental part of a student’s personal and academic development in which the student begins to accept and assimilate the norms of the academic system in which they are immersed. Expressed in terms of Maddi’s hardness framework, it is similar to the construct of commitment. While grade performance reflects the student’s ability, attributes, and achievements as evaluated by the institution's values and objectives, intellectual development can be seen as the student’s cognitive evaluation of the ability to succeed in the academic environment. Thus, academic integration is a critical factor in relation student persistence.

Testing the validity of Tinto’s model, Pascarella and Terenzini (1983) used a path analysis method to examine student persistence on a sample of 763 residential university freshmen. In agreement with Tinto’s expectations, their findings indicated a significant interaction between social and academic integration and between institutional and goal commitment. In terms of the magnitude of influence on persistence, Pascarella and Terenzini demonstrated that academic integration is more important for students with low levels of social integration, such as freshmen. For students such as these, where academic integration is a main indicator of their persistence, the performance feedback they receive from a predictive analytics system may be influential in validating their experience and identity.

Consistent with these findings is Grosset’s (1991) study of 449 students from a large, urban community college. Using discriminant analysis to compare first-term student persistence with non-persistence, Grosset determined that the quality of integration experiences, including advising experiences, is more important to persistence than the quantity of these experiences, and academic integration is more influential in the persistence
process than social integration. A noteworthy implication of the research was that academic advising informed by the use of predictive analytics may have contributed to the overall integration experience, and thereby, their level of self-efficacy and persistence.

Tinto’s (2005) views on the importance of integration experiences have remained steady and, yet, reflect a contemporary view of the connection between student success and the student’s first-year college experience as stated:

Involvement, or what has been frequently been described as academic and social integration, is a condition for student success (e.g., Astin, 1993; Tinto, 1993). Quite simply, the more students are academically and socially involved, the more likely are they to persist and graduate. This is especially true during the first year of university study when student membership is so tenuous yet so critical to subsequent learning and persistence. Involvement during that year serves as the foundation upon which subsequent affiliations and engagements are built. (p. 4)

Tinto (2005) challenged the institutions of today by asking, “What would it mean for institutions to take student success seriously” (p. 1)? He believed that research identified six conditions within institutions that are supportive of student success: commitment, expectations, support, feedback, involvement, and learning. Referring to feedback, he stated: “...monitoring and feedback is a condition for student success. Students are more likely to succeed in settings that provide faculty, staff, and students frequent feedback about their performance” (p. 4). Once again, this view supports the important role analytics can help provide institutions regarding student success.

Predictive analytics used in the context of academic advising intersects with the triad of the core elements (i.e., feedback, support, and involvement) of Tinto and Pusser’s (2006) institutional model (see Figure 1) of action for student success. Predictive analytics specifically contributes to these core elements by: (1) creating opportunities for feedback to occur through early warning notifications to faculty and advisors; (2) improving support
Figure 1. A preliminary model of institutional action through performance monitoring, leading to informed recommendations on next-steps such as supplemental instruction, and (3) increasing involvement and interactions with faculty and advisors. By enhancing engagement with more informed recommendations, students are more likely to take actions that lead to their persistence.

Applying Tinto’s (1975) integration framework to the context of analytics, and Astin’s Input-Environment-Output (I-E-O) model, a conceptual model emerges (Figure 2).
Key elements of the environment include academic self-efficacy, academic integration, and social integration. These are central factors in Tinto’s integration framework. These factors are a product of the support and involvement circles illustrated in Figure 1. First-term academic engagement characteristics are a result of the curriculum requirements and academic advising efforts expressed as number of credits attempted, completed and GPA. These elements reflected by the learning box in Figure 1. In the current study, the treatment—academic advising informed by predictive analytics—represents an attempt to boost the effectiveness of the triad of feedback-support-involvement in Figure 1 (i.e., the quality-of-effort box). The success box may be regarded in terms of outcomes, namely, students who do select a program of study generally do not change majors.

**Emergence of Analytics in Higher Education**

The premise of rational decision-making dates back to Weber (1947), who posited that organizational efficiencies could be derived from an unambiguous set of rules governing performance. However, Simon (1979) noted that the “complete knowledge” required to “compute the consequences” (p. 500) and make rational choices, was not easy to ascertain.
With the volume of data that is now available in higher education from sources such as student information systems, learning management systems and other services that collect data, the “full knowledge” Simon aspired to retrieve is now obtainable in higher education. Through appropriate implementation and use of data mining tools, the possibility for rational, evidence-based, data-driven decision-making based on analytics is achievable. According to Picciano (2014), the use of analytics as a basis for evidence-based decision-making reflects a “rational model directed by values and based on data” (p. 36) that aligns with the needs and values of institutional leaders and stakeholders.

Access to large volumes of data is required for analytics to demonstrate predictive capabilities. Big-data is a term often used to describe the capture, storage, and processing of large amounts of transactional data generated by an individual within a system. For example, student actions recorded in a learning management system (LMS) may generate thousands of transactions as multiple students interact through e-mails, course assessments, discussion forums, and blogs. Those data can then be examined for patterns. The goal of analytics is to identify patterns in large, multiple data sets that assign meaning to the patterns and predict outcomes which can be used to guide actions and decision-making.

Academic analytics and learning analytics are often referred to interchangeably in higher education; however, each form has its own distinguishable features. Academic analytics focuses on the improvement of resources, processes and workflows of the academic institution through the use of learner, academic, and institutional data (Campbell, DeBlois, & Oblinger, 2007). Learning analytics has more to do with the measurement and analysis of data about learners and the contexts in which learning can be made more effective (Anirban, 2014). A more detailed description of each of these types of analytics follows.
Academic analytics

Academic analytics systems utilize data institutional collected on students and faculty (Piety, Hickey, & Bishop, 2014). The availability of data from sources such as student information systems is essential (Wayman & Stringfield, 2006). Other researchers have also emphasized the importance of having real-time access to data to more accurately capture (i.e., data-mining) the performance of courses, programs, departments, and colleges (Campbell et al., 2007; Davenport & Harris, 2007; Ravishanker, 2011; van Barneveld, Arnold, & Campbell, 2012). Access to various institutional data sources enables scientists to construct complex models that, in turn, allow institutions to be compared objectively on metrics such as acceptance rates, progression measurements, and completion statistics. An important characteristic of these metrics is that they reflect the institution as a whole (Goldstein & Katz, 2005; Piety, Hickey, & Bishop, 2014). Describing the performance of an institution by using retention and completion rates has also been described as “macro level analytics” (MacNeill, Campbell, & Hawksey, 2014, p. 4).

Thus, the objectives of academic analytics for administrators and leaders in higher education include the discovery of issues to visualize relevant trends and project the impact of potential changes before actually implementing them. An additional goal of analytics is to provide administrators with the flexibility to disaggregate data to compare what is happening in different colleges and departments. For example, Siemens and Long (2011) posited that a goal of analytics is to help colleges and departments identify problematic marker courses so that more efficient academic pathways can be designed. In addition to institutional goals, another form analytics focuses on instructional facets of the academic environment.
Learning analytics

Siemens and Long (2011) defined learning analytics as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (p. 4). As instructors in higher education institutions have begun to use blended teaching approaches involving face-to-face and instructional technology more frequently, the context for learning has shifted to internet-based environments. This shift has increased the opportunity to capture of student behavior by recording the students’ actions in databases. The data collected on students from learning management systems has been shown to be useful in informing educators of at-risk students and the need for possible intervention (Gulbahar & Ilgaz, 2014; Picciano, 2014). In addition, data of this nature can be used to identify learning patterns and provide insights into the actions taken by learners who comprehend course concepts. Instructional activities that correlate to learning and comprehension can be used more closely as a basis for identifying students who may not be engaging in activities in the same ways (Siemens & Gasevic, 2012). Used in this way, learning analytics can inform faculty not only when but also where pedagogical improvements are needed.

Usage

The high volume and assortment of data sources, time constraints, and gaps in skills and knowledge experienced by faculty and administrators in higher education may lead to data overload. In such cases, data are simply ignored (Wayman, Cho, & Johnston, 2007; Snibbe, 2006). Analytics can improve this situation by requiring institutions to think pragmatically about their intended goals and then capture and analyze only the data that are germane for specific objectives. For example, for course-level objectives, data selected from
learning management systems and social networks can be used to identify at-risk students and provide an opening for engagement and intervention (Arnold & Pistilli, 2012; Dietz-Uhler & Hurn, 2013; Macfadyen & Dawson, 2010). At the departmental level, administrators can use analytics to construct models that reveal the need for changes in program elements and curriculum. These models can also be used to demonstrate the effect such changes may have on degree attainment, time-to-degree, and other student outcomes (Baepler & Murdoch, 2010; Siemens & Long, 2011; Shum & Ferguson, 2012). Finally, at the institutional level, analytical models based on data from student information systems can provide institutional administrators the means to assess the effectiveness of colleges, departments, and programs at a higher level in order to consider changes that may improve efficiency (Campbell et al., 2007; Head, 2010; Hung, Hsu, & Rice, 2012; Suryadi, 2007).

In summary, analytics can serve as a valuable tool to guide educators and administrators as they contemplate possible changes based on the “the capacity of the organization to absorb changes and resource constraints” (Kavanagh & Ashkanasy, 2006, p. 1). However, a key requirement for an analytics implementation to be successful is for the institutional culture to accept and embrace the system-wide use of data. Data literate leaders who can demonstrate the use of data are often advantageous in changing culture. These types of demonstrations can help promote data use readiness.

**Analytics Readiness**

Recent literature has expounded on the concept of readiness as a key indicator of the successful implementation and use of data analytics systems. Maltby (2011) examined analytics readiness from the perspective of challenges and barriers that inhibit adoption of analytics. After a comprehensive review of scholarly sources, he concluded that academic
and corporate organizations must examine the following areas before embarking on the use of analytics: privacy (i.e., how information about students is used), data security, legal rights (i.e., who owns the data), and human capital (i.e., whether sufficient personnel and skills are available). Finally, Maltby suggested that future studies of analytics should include cost-benefit outcomes in order to be more compelling.

Picciano (2014) examined the role of big data in conjunction with analytics readiness in postsecondary institutions and determined that the use of analytics created concerns in similar areas: incomplete data leading to biased conclusions (e.g., face-to-face interactions which are not recorded), privacy concerns (e.g., data profiling and collecting more data than needed for the purpose of student learning), lack of data scientists with the needed skills to use and interpret data, and insufficient security with increased risk of data exposure. While the findings of Maltby and Picciano allude to possible institutional challenges, Pressler (2014) applied the McKinsey 7S Model (Waterman, Peters, & Phillips, 1980) to develop a series of reflective, open-ended questions to assess organizational data readiness. By using this model, institutional leaders, administrators, faculty, and professional staff can participate in interviews and focus groups to identify vulnerabilities and inconsistencies in the use of data in such areas such as shared values, strategies, structures, systems, styles, staff, and skills. The responses are then reviewed for common themes to determine the organizational readiness.

Bichsel (2012) surveyed 231 postsecondary institutions to quantitatively identify common readiness factors in the use of analytics. An analysis of the survey responses revealed that participants believe institutions engaged in analytics must have leaders who support processes that augment the use of data to make decisions. The quality of the data
must also be good. In addition to appropriate reporting tools, adequate funding and resources, sufficient expertise and knowledge, a strong data governance model and IT infrastructure that can support analytics are required. Similarly, Baer and Norris (2013) surveyed 40 postsecondary institutions with established success and achievement in the use of analytics to improve student success. They produced a matrix connecting the stages of data analytics development with certain institutional characteristics. Their findings revealed that characteristics of an institution that most often contribute to a successful implementation include adequate technology, appropriate policies and practices, sufficient skills, a culture that embraces data-use, and a supportive leadership. Baer and Norris also identified gaps that inhibit the effective use of analytics in postsecondary institutions: the failure of an alignment with needs and available commercial software, unmet expectations in regard to capabilities, a lack of collaboration, and insufficient talent. Finally, in an effort to more accurately assess an institution’s analytics readiness, Arnold, Lonn, and Pistilli (2014) developed the Learning Analytics Readiness Instrument (LARI). The survey measures five factors in regards to analytics readiness: ability and skills, data collection and ownership, culture and processes, governance and infrastructure, and overall readiness perceptions.

While the aforementioned research attempted to identify common institutional challenges, the exploration of socio-cultural practices and processes within an institution can also shed light on the acceptance of data-use. Elias (2011) called attention to the importance of the organizational culture, and stated that “regardless of how good the analytics, to truly apply knowledge, institutions need committed leaders who will consciously build organizational capacity to improve performance and to change organizational culture and behavior” (p. 16). This statement by Elias alludes to the possible mismatch that may occur
between administrators who oversee an organization and those who operate within a culture. Subsequent research has revealed that data-use perceptions vary at different levels within an organization, and impact processes, practices, and the extent to which the data can be used to guide decision-making (Maxwell, Rotz, & Garcia, 2015). Therefore, it is essential to understand cultural perceptions related to data-use readiness.

**Organizational factors influencing data use**

Studies exploring the socio-cultural elements related to the development of data-driven educational environments have helped to uncover perceptions regarding data-use within institutions, especially in regards to traditional practices and processes that have existed and may need to be changed. As noted by Macfayden and Dawson (2012, p. 161), institutional leaders advocating for analytics “must delve into the socio-technical sphere” of those involved in the actual use of analytics and stimulate changes in established practices. The authors’ case study of an institution implementing learning analytics illustrates the importance of engaging the faculty at an institution that attempted to integrate learning analytics within their learning management system. The planning was dominated by technical concerns with little emphasis given to faculty involvement for the purpose of instructional improvements. As a result, the effectiveness of the analytics system was limited. Macfayden and Dawson posited that, while a lack of faculty involvement was an issue, there were numerous other related obstacles in the culture that resisted change. Among these obstacles were authoritative bodies that overemphasized consensus to make decisions, faculty control over traditional areas such as teaching and research, a preference to add resources rather than reallocate them, and a curriculum designed to uniformly educate the masses. When clarifying the role of administrators, Wayman and Cho (2014) recommended
that institutional leaders should establish processes that elicit feedback and embrace ideas about how new data systems might be used within the campus environment.

Norris, Baer, Leonard, Pugliese, and Lefrere (2008) explored factors related to data-use in higher education associated with the purpose to improve institutional performance. They posited that a culture that embraces and accepts change is fundamental when shifting to data-driven decision-making. The acceptance of change can be facilitated through the development of collaborative efforts that align institutional goals with those of colleges department, and program goals. To enhance collaboration, Greller and Drachsler (2012) proposed a framework that “should be used to design learning analytics services from an inclusive perspective” (p. 44). Their framework focused on such cultural factors that affect the adoption of analytics such as working with stakeholders, developing shared objectives, improving access to data, providing instruments and tools, identifying external constraints, and recognizing internal limitations.

The internal limitations to organization include resistance and lack of motivation as noted by Macfayden and Dawson. Greller and Drachsler (2012) asserted that such resistance is inherently related to anxieties regarding competence (i.e., a lack of skills in interpretative and critical evaluation in regard to data-use) and acceptance (i.e., a blunt rejection of empirical methods by certain stakeholders). Siemens and Long (2011) asserted that faculty, students, and administrators must clearly understand and articulate the improvements (i.e., changes) sought through data-use and remain actively involved in the process of enacting data-driven practices. Other researchers have noted the importance of shared goals and formative interactions centered on data-use (Baker, 2013; Blaich & Wise, 2010, Macfayden, Dawson, Pardo, & Gasevic, 2014). Wayman and Cho (2014) asserted that new data systems
should be situated within the existing environment in order to demonstrate that they are useful in solving problems encountered by educators in their current daily practices. A theme throughout the aforementioned research is the importance of small, collaborative teams working on shared problems that, when resolved, could have an immediate impact.

Researchers have also noted the need to focus on change as part of an analytics initiative. In this regard, institutions should form specific, strategic objectives that they believe need to be pursued (Daniel & Butson, 2014; Jenkins & Kerrigan, 2008; Norris & Baer, 2013) in order to lead to the desired changes. Macfayden and Dawson (2012) also asserted that change is predicated on “planning processes [that] create conditions that allow participants to think and feel positively about change” (p. 161). Hora, Bouwma-Gearhart, & Park’s (2014) investigated such institutional practices that motivate faculty to accept data-use to guide their decision-making. The results of their case study of three large, public universities revealed three central practices that facilitate data-use by faculty: providing structured opportunities for meaningful data collection and interpretation, providing adequate time for reflection about data, and encouraging a socio-cultural setting in which it is a norm to use data in daily practices. The authors stated that introducing new practices in relationship to a new data system, such as a learning analytics system, requires an understanding of the “integrative theory of change” (p. 24) whereby external and local leaders work with experienced educators to establish data-driven practices within the educational culture.

Jenkins and Kerrigan (2008) attempted to uncover data-use readiness factors that guide data-driven decision-making in higher education in a study of community colleges. Based on the results of a survey of faculty (N=2,478) and administrators (N=1,612) from
postsecondary institutions participating in Achieving the Dream, the authors suggested four composite indicators that promote data-use practices by faculty and administrators: (1) research student outcomes; (2) research the impact of race, ethnicity, and income; (3) participate in organized discussions on improving student success; and (4) stimulate teaching-related decisions. Jenkins and Kerrigan emphasized that building a “culture of evidence” (p. 43) that relies on data to make decisions is a complex process requiring a concerted effort from committed institutional leaders and support from departmental leaders to bring about changes in practice. This conclusion aligns with findings by others (Kavanagh & Ashkanasy, 2006; Rogers, 1995) who noted that significant changes in culture and practice through the introduction of data-use, require a lengthy timeline for processes to adapt and integrate data-use into workflows (e.g., Achieving the Dream was a five-year effort).

In summary, the findings presented in this section revealed that the objectives for using data and improving data-driven decision-making: (a) must be clear and strategic; and (b) must receive sufficient “buy-in” from key stakeholders, such as faculty, leaders, and those who interact with students. In regards to learning and instruction, such a buy-in may be difficult to achieve given the traditional nature of teaching and academic freedom in higher education as well as resistance to change. An implication is that the successful implementation of analytics initiatives in other areas of the institution, such as academic advising, may be more attainable and gain greater acceptance because outcomes are more focused on student success, which is a current issue of great importance.

**Previous Research Methods to Predict Academic Outcomes**

Astin’s (1993, 1999) theory of student involvement suggested that activities requiring a student to interact with his or her environment will improve retention and academic
performance. The use of predictive analytics can prompt such interactions. For example, an automated notification sent from an analytics system to a student and his or her advisor warning about academic performance issues provides an opportunity for an interaction with the student, such as a subsequent phone call or e-mail from the advisor. Campbell, DeBlois, and Oblinger (2007) noted that institutions should track and measure the impact of those interactions, such as whether students actually keep an advising appointment or follow through on the recommendations given. Analytics system provide a mechanism for tracking these kinds of interactions. While some analytics systems may use data from learning management systems to construct a predictive model and identify students who are potentially at-risk, student information systems data may also be used to construct predictive models.

Much of the early groundwork on predictive analytics within higher education focused on the construction of statistical models to predict students with issues of academic performance. Many of these models utilized logistic regression methods. The University of Alabama (Davis, Hardin, Bohannon, & Oglesby, 2007) was one of the pioneers in predictive models in higher education. Graduate students at the institution were given access to the data of enrolled freshmen from 1999, 2000 and 2001, and asked to develop predictive models of at-risk students. Using logistic regression, decision trees, and neural networks, the graduate students developed a single refined model with eight indicators of student performance: cumulative GPA, English course grades, math course grades, distance from campus to home, race, total earned hours, and ACT or ACT-converted SAT scores.

Another early pioneer was Sinclair Community College. Findings of a study by researchers at the institution revealed that GPA was a major predictor of student success in
marker courses (Burns & Cole, 2007). The analytics system they developed integrated data from three distinct sources, bearing a similarity to a big-data-like approach. The data sources were: (1) demographic and admissions data; (2) real-time course registration, grades, and financial aid status from the student information system; and (3) counselor risk-assessment notes and faculty-initiated alerts. The significant predictive variables in their model were: first-term GPA, first-term grades, placement-test referrals into two or more developmental courses, individual or family income level below the federal poverty level, full-time work status, and whether the student was an undecided major. Factors that were indicative of student success were: GPA, passing all developmental courses, deciding on a major and a career, resolving child-care and transportation issues, and attending class regularly (http://www.sinclair.edu/stservices/edu/pub/Generic%20SSP-EAP%20Pres.ppt).

In his doctoral dissertation at Purdue University, Campbell (2007) used factor analysis and logistic regression on student activity data collected from their learning management system. Six variables were found to be significant in determining academic performance: ACT or SAT score, overall GPA, an LMS usage composite score, an LMS assessment composite score, an LMS assignment composite score, and an LMS calendar composite score. This study resulted in the development of Course Signals (Arnold, 2010). The Course Signals system uses predictive models to identify students that may be struggling academically and in need of intervention. Research conducted between fall 2007 and fall 2009 revealed significant improvement on course completion and mastery of content.

Martinez (2001) developed a discriminant function analysis to help predict the placement of students in specific courses by using pre-college assessment data at the community college level. The findings revealed that high school GPA, age, gender, grade in
last math class, highest math class, ethnicity, major declared, and number of work hours were predictive of success in English courses. In addition, current credit hours, amount of financial aid, and program level were predictive of the likelihood of withdrawal.

A logistic regression was used by Allen and Robbins (2008) to predict persistence within a major. Using a sample of 47,914 students from across 25 four-year institutions, they found that a student’s first-year GPA ($p < .001$) and “major sureness” ($p < .001$) were significant indicators of major persistence. Surprisingly, they found that neither high school GPA nor ACT composite score contributed to the prediction of major persistence. However, a limitation of their analysis was the omission of students who entered as an undeclared major or undecided. This omission led to the exclusion of important subgroups, such as certain minority groups and students with lower GPAs or lower ACT scores. They asserted that, given the large number of students who enter as undecided, this group warrants further examination.

Barber and Sharkey (2012) reported on the creation of three different models at the University of Phoenix to predict levels of risk of a student failing a course. Multiple sources of data sources were used, including the learning management system, the financial aid system, and the student information system. The logistic regression model, whose outcome variable was an indicator of whether the student would pass the course, was accurate more than 90% of the time. The variables they used to develop the model included: prior credits earned, gender, age, ethnicity, marital status, number of dependents, employment status, household income, high school GPA, financial aid, Pell grant recipient, total student loans taken, financial status, ratio of credits earned to attempted, and military status. In contrast to previous studies, their findings revealed that such variables as gender, age, military status,
Pell grant receipt and financial aid were not significant (although there was a slight relationship between military status and degree level). The variables that revealed significant coefficients were: percentage of points earned in prior courses, credits earned, the number of online posts in forums, and cumulative points earned.

Finally, Jayaprakash, Moody, Lauria, Regan, and Baron (2014) produced four different predictive models to identify the risk of course failure based on student data from a mid-sized comprehensive liberal arts institution located in New York State. By applying logistic regression, they revealed that many of the regression coefficients they had selected were statistically significant. More specifically, the variables included in the regression were: regular student, part-time student, cumulative GPA, partial grades score, number of LMS sessions, freshman, sophomore, junior, senior, probation, regular standing, honors or dean’s list, and number of partial grades score. Only the freshman and junior class indicators were non-significant in predicting at-risk students. They noted that the number of LMS sessions decreased the expected probability of being at-risk. They also observed that regular students, as compared to online students, experienced a large reduction in the expected probability (.319) of being at-risk. Part-time students had a much higher probability (2.443) of being at-risk than full-time students. Sophomore students had a greater expected probability (1.340) of being at-risk than freshmen (.875) or juniors (1.023). Finally, students on probation and regular students were much more likely to be at-risk (25.5 and 8.4, respectively) than honors students.

**Summary**

Based on the research studies reviewed, it became evident that logistic regression is an accepted method for use in developing predictive models of academic outcomes.
Although the current study did not focus solely on the identification of at-risk students, many of the same variables were considered for first-year, first-term engineering students who had not selected a program of study since the delay to commit to a major or program of study introduces risk (i.e., failing to establish goals, creating low self-efficacy, disciplinary anomie). The goal of my study was to identify the first-year, first-term indicators that predicted change of major and the selection of program of study, and whether the use of predictive analytics impacted those outcomes.

The literature presented in this chapter revealed that a great deal of research has been conducted on the rationale for using analytics in higher education and the practices required to integrate analytics into the culture. In addition, the previous research connected to the construction of predictive models on student outcomes and persistence has been informative in helping to identify which variables tend to have significant impact on outcomes. In alignment with this previous research, the variables I selected as indicators for my study included: age, gender, ACT composite score, ACT math score, high school rank, number of first-term credits attempted, number of first-term credits completed, participation in a learning community, and participation in honors program. I also used Mapworks variables (i.e., commitment to completing a degree, self-assessment of math skills, self-assessment of time management and planning, academic life satisfaction, academic self-efficacy can do well on challenging problems and in hardest course) and factors (i.e., self-efficacy, academic integration, and social integration) from the Mapworks dataset.

The outcomes of interest (i.e., dependent variables) in my study were the selection of a program of study and change of major after the first-term. I also constructed a logistic regression model to determine which variables and factors were significantly predictive on
these outcomes. These variables, especially those that were significant, informed the subsequent selection of variables in the propensity score analysis to determine the impact of the treatment (i.e., use of predictive analytics by academic advisors) on the fall 2015 student cohort. The methodology is described in more detail in the following chapter.
CHAPTER 3. METHODOLOGY

This chapter describes the methodology used in this research study including data sources, data security, data collection, analysis procedures, research questions, limitations, and ethical considerations. A quantitative research approach tends to rely on empirical methods (Creswell, 2003), where properties are measured and counted. Empirical approaches are often linked to a philosophical viewpoint known as positivism where scientific research is conducted with objectivity and “objects in the world have meaning prior to, and independently of, any consciousness of them” (Crotty, 1998, p. 27). From this perspective, the world is viewed as scientific, systematic, well-organized, and where knowledge is certain, accurate, and attainable.

Methodological Approach

As social sciences surfaced during the 20th century, a modified epistemology was developed, labeled post-positivism. As with positivism, a belief in certitude remains a core component, but is less absolute. It is an approach that references “probability rather than certainty and seeks to approximate the truth rather than aspiring to grasp it in its totality” (Crotty, p. 29). In accordance with this statement, Popper (1959, 1969, 1972), regarded as one of the great philosophers and first thinkers of post-positivism, offered the following view: (a) observation takes place within the context of theory and is shaped by theory; (b) scientists engage in a continual process of conjecture to construct knowledge; (c) research seeks to develop statements that explain relationships probabilistically rather than with certainty; (d) falsification replaces verification (something is unable to be proved false); and (e) causal relationships are not viewed as unbending axioms but rather as a set of warranted statements that are modified or withdrawn in light of further investigation.
This research study was oriented from this philosophical perspective and quantitative methods were applied. A dataset was generated at the institution for subsequent use in correlational analyses, $t$-tests, logistic regressions, and propensity score analyses. The students in the dataset were students at the institution who were beneficiaries of a particular treatment. These students were first-year, first-term freshmen majoring in engineering who had not selected a program of study in engineering (i.e., undeclared engineering major). The dataset was constructed by extracting student records from the university’s student information system (SIS) database from the fall and spring semesters of 2012 through 2015. Then, the data from the SIS dataset was paired with student response records from the Mapworks® transition survey (Skyfactor, 2015) database coordinated by the Dean of Students and Department of Residence. The results of the statistical analyses were used to infer generalizations and provide implications for institutional leaders to consider in regards to academic advising informed by the use of predictive analytics in relation to first-year, first-term engineering student outcomes, as well as to provide guidance on the impact of using predictive analytics in regards to change of major and the selection of a program of study in engineering.

Hence, the purpose of this study was to obtain data from two datasets, one academic and one environmental, and to explore early indicators that predict change of major and selection of program of study among first-year, first-semester undeclared engineering students based on their demographic, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental Mapworks survey responses, particularly in regards to their academic self-efficacy, social integration, and academic integration. The goals of this study were to: (a) identify the relationships between
student characteristics from their academic profile and their Mapworks generated responses regarding their academic transition to the postsecondary and produce a model that would predict student outcomes such as change-of-major (i.e. disinclination) or selection of a program-of-study (i.e. perseverance); and (b) determine the impact of academic advising informed by predictive analytics on such outcomes.

**Research Design**

This study used a quasi-experimental design based on paired observational data from two separate datasets to determine which variables were explanatory in predicting academic outcomes and the impact of academic advising informed by the use of predictive analytics on the tendencies (i.e., change-of-major or selection of program of study) of first-year, first-term freshmen engineering students without a selected program of study (i.e., undeclared engineering major). Observational studies share similarities (i.e., the interpretation of cause-and-effect) with experimental design studies, but do not include the element of random assignment to treatment and control groups since it may be unethical, or unreasonable, to impose on the treatment participants. In this study, the treatment occurred in the fall 2015 (i.e., those receiving academic advising informed by the use of predictive analytics).

**Data Sources**

In order to explore the relationships and the impact of the treatment, data from the institution’s student information system (SIS) database was used as a source for demographic characteristics, pre-enrollment academic characteristics, and first-term academic engagement and completion characteristic germane to the outcomes of interest. In addition, the institution’s Mapworks dataset was used as a source for transitional, environmental data to measure perceptions of academic self-efficacy, academic integration, and social integration.
Using university ID numbers, the data records in the two databases were matched and de-identified by institutional staff in the Office of the Registrar and Department of Residence. In cases where SIS data records did not correspond to a Mapworks response, those cases were excluded. The student data spanned the 2012-2015 academic years. Data from the tenth day of class in the fall semester and on the tenth day of class in the following spring semester were used to construct the dependent variables: change-of-major and selection-of-program-of-study. The Mapworks® transition survey (Skyfactor, 2015) data was collected during the month of September each year.

**Mapworks – validity**

According to Fraenkel and Wallen (2006), the validity of a survey is defined as “the degree to which an instrument measures an intended hypothetical psychological construct, or non-observable trait” (p. 516). In 2005, EBI Mapworks partnered with Ball State University to develop the Mapworks® transition survey. The internal validity of survey questions has been improved upon over the years and, through factorial analysis, it has been confirmed that Mapworks does, indeed, measure relevant concepts in areas such as academic self-efficacy, peer connection, academic adjustment, sense of belonging, and overall evaluation of the university (Gansemer-Topf, Kollasch, & Sun, 2015).

**Mapworks – reliability**

Cronbach’s alpha (α) is used as a measure of internal consistency and reliability for survey constructs designed to measure the same construct (Streiner, 2003). Nunnally (1978; Nunnally & Bernstein, 1994) has recommended that reliability scores of 0.70 are acceptable in the early stages of research, 0.80 for basic research tools, and 0.90 or 0.95 for clinical purposes. In Mapworks the constructs of interest for this study achieved the 0.80 cut-off.
Based on 2014 analysis, the academic self-efficacy construct contained 3 items (α = .86), the academic integration construct contained 4 items (α = .87), and the social integration construct contained 3 items (α = .90).

**Data Access and Security of the Data**

For this research study the data was obtained in compliance with the procedures and policies of the IRB office at the institution. The dataset did not include any identifiable data such as names, addresses, etc. The data set was stored in a secured (i.e. password protected) cloud-based storage system. Only the researcher and major professor had access to this storage location. Statistical analysis was conducted using SPSS V20 and Stata/SE 13.1.

**Data Collection**

Two observational data sources were used in this study. First, the student information system (SIS) database at the institution (administered by the Office of the Registrar) contained information on every student enrolled. The data are updated at the beginning of every academic semester to include new students admitted to the university. Previous student data are archived. The data extracted for this study included: age, gender, ethnicity, classification year, ACT composite score, ACT math score, high school rank, number of first-term credits attempted, number of first-term credits earned, participation in honors programs, participation in learning communities, fall term major, fall term program of study, spring term major, and spring program of study. Second, data from the Mapworks dataset was provided through the Department of Residence. This data reflected the affective characteristics of students and their self-assessed perceptions on variables and factors such as academic integration, social integration, and academic self-efficacy.
Variables

This study focused on pre-treatment characteristics of first-year, first-term freshmen in engineering without a program of study and their self-reported integration experiences as reported through their participation in the Mapworks® transition survey (Skyfactor, 2015). Prior to performing the logistic regression analyses, correlations were performed to examine the relationships between the SIS variables and Mapworks variables and factors, and the outcomes of interest (see Appendix C).

Dependent

Table 1 lists the dependent variables used in this study. The dependent variables in this study were computed based on the data from the SIS dataset using the “tenth day of class” value for major and program of study. For the fall semester, all first-year, first-term freshmen in the sample chose engineering as a major but had not specified a program of study, or particular specialization of engineering such as electrical, aerospace, civil, etc. To determine if a student had changed their major (i.e., transferred from engineering for another college) or selected a program of study (i.e., picked a specialized program in engineering), their major and program were taken again on the tenth day of classes in the spring semester immediately following (i.e., F12->S13, F13->S14, F14->S15, F15->S16). The F15->S16 cohort was considered the treatment group because they were beneficiaries of academic

Table 1. Dependent variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
<th>Key</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of major</td>
<td>CHG_MAJ</td>
<td>1=yes, 0=no</td>
<td>RQ3, RQ4</td>
</tr>
<tr>
<td>Program of study selected</td>
<td>SEL_PGM</td>
<td>1=yes, 0=no</td>
<td>RQ3, RQ4</td>
</tr>
</tbody>
</table>
advising informed by the use of predictive analytics. The other cohort groups from previous years were pooled together to form the comparison group.

Each dependent variable represents a dichotomous outcome, in the logistic regression, represented as a binary value. These dependent variables were also utilized in the propensity score analysis as the outcome variable of interest to measure the impact of the treatment.

**Independent**

**SIS**

Table 2 lists the independent variables used in this study from the SIS dataset. Demographic variables were age (AGE) and gender (GENDER). Pre-college academic characteristics were ACT composite score (ACT_CMPST), ACT math score (ACT_MATH), and high school rank (HS_RANK). First-term engagement and completion variables were: first-term credits attempted (FT_CRED_ATT), first-term credits completed

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
<th>Type</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>AGE</td>
<td>Continuous</td>
<td>16 through 25</td>
</tr>
<tr>
<td>Gender</td>
<td>GENDER</td>
<td>Categorical</td>
<td>1=male, 2=female</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>ETHNICITY</td>
<td>Categorical</td>
<td>1 through 8</td>
</tr>
<tr>
<td>ACT composite score</td>
<td>ACT_CMPST</td>
<td>Continuous</td>
<td>14 through 36</td>
</tr>
<tr>
<td>ACT math score</td>
<td>ACT_MATH</td>
<td>Continuous</td>
<td>16 through 36</td>
</tr>
<tr>
<td>High school rank</td>
<td>HS_RANK</td>
<td>Continuous</td>
<td>27 through 99</td>
</tr>
<tr>
<td>First-term credits attempted</td>
<td>FT_CRED_ATT</td>
<td>Continuous</td>
<td>11 through 27</td>
</tr>
<tr>
<td>First-term credits completed</td>
<td>FT_CRED_COMP</td>
<td>Continuous</td>
<td>0 through 20</td>
</tr>
<tr>
<td>First-term cumulative GPA</td>
<td>FT_GPA</td>
<td>Continuous</td>
<td>0.0 through 4.0</td>
</tr>
<tr>
<td>Learning Community Memberships</td>
<td>LC_MEM</td>
<td>Continuous</td>
<td>0 through 3</td>
</tr>
<tr>
<td>Honors Program Membership</td>
<td>HP_MEM</td>
<td>Categorical</td>
<td>1=yes, 0=no</td>
</tr>
</tbody>
</table>
(FT_CRED_COMP), first-term GPA (FT_GPA), learning community memberships (LC_MEM), and honors program member (HP_MEM).

**Mapworks factors and variables**

Table 3 lists the independent factors and variables from the Mapworks dataset. When considering the factors, the academic integration factor (ACAD_INT) was comprised of four individual questions that clustered together which asked students about the degree they were keeping current with their academic work, the degree they were motivated to complete academic work, degree they were learning, and the degree they were satisfied with their academic life. The social integration factor (SOC_INT) was comprised of three individual questions that clustered together which asked students whether they felt they belonged at the institution, whether they were fitting in, and whether they were satisfied with their social life. The last factor, academic self-efficacy (ACAD_SE), was comprised of three questions that clustered together which asked students to what degree they did well on problems and task, to what degree they would do well in their hardest course, and to what degree they could persevere on class projects even when they were challenging.

In addition, the individual variables that correlated significantly with the outcome variables or were significant in the preliminary regression analyses were utilized in subsequent analyses. These variables included commitment to completing degree (DEG_COMM), self-assessment of math skills (SA_MATH), self-assessment of time management and planning (SA_PLAN), academic life satisfaction (SA_ACAD_LIFE), academic self-efficacy (can do well on problems (SE_PROBS_TASK) and in hardest course (SE_HARD_CRSE)), and institutional satisfaction (INST_SAT).
Table 3. Independent variables and factors from Mapworks

<table>
<thead>
<tr>
<th>Factor/Variable</th>
<th>Variables</th>
<th>Code</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academic integration (F)</strong></td>
<td>• Keeping current with academic work (Q152)</td>
<td>ACAD_INT</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td></td>
<td>• Motivated to complete academic work (Q153)</td>
<td></td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td></td>
<td>• Learning (Q154)</td>
<td></td>
<td>7 = Extremely</td>
</tr>
<tr>
<td></td>
<td>• Satisfied with academic life (Q155)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Social integration (F)</strong></td>
<td>• Do you belong here (Q156)</td>
<td>SOC_INT</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td></td>
<td>• Are you fitting in (Q157)</td>
<td></td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td></td>
<td>• Are you satisfied with your social life (Q158)</td>
<td></td>
<td>7 = Extremely</td>
</tr>
<tr>
<td><strong>Academic Self-efficacy (F)</strong></td>
<td>• Do well on problems and tasks (Q039)</td>
<td>ACAD_SE</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td></td>
<td>• Do well in hardest course (Q040)</td>
<td></td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td></td>
<td>• Persevere on class projects when there are challenges (Q041)</td>
<td></td>
<td>7 = Extremely</td>
</tr>
<tr>
<td><strong>Q002, Commitment to Completing a Degree (V)</strong></td>
<td></td>
<td>DEG_COMM</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td><strong>Q016, Self-Assessment: Math (V)</strong></td>
<td></td>
<td>SA_MATH</td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td><strong>Q019, Self-Assessment: Follows thru on what they say they will do (V)</strong></td>
<td></td>
<td>SA_FLW_THRU</td>
<td>7 = Extremely</td>
</tr>
<tr>
<td><strong>Q022, Self-Assessment: Plans Out Time (V)</strong></td>
<td></td>
<td>SA_PLAN</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td><strong>Q039, Academic Self-Efficacy: Degree to do well on problems and tasks</strong></td>
<td></td>
<td>SE_PROBS_TASK</td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td><strong>Q040, Academic Self-Efficacy: Degree to do well in hardest course (V)</strong></td>
<td></td>
<td>SE_HARD_CRSE</td>
<td>7 = Absolutely</td>
</tr>
<tr>
<td><strong>Q155, Self-Assessment: Degree satisfied with academic life (V)</strong></td>
<td></td>
<td>SA_ACAD_LIFE</td>
<td>1 = Not at all,</td>
</tr>
<tr>
<td><strong>Q159, Would choose ISU again (V)</strong></td>
<td></td>
<td>INST_SAT</td>
<td>4 = Moderately,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7 = Extremely</td>
</tr>
</tbody>
</table>
Sample

The students studied were enrolled at the institution during the years 2012 through 2016. Specifically, the population consisted of first-year, first-term freshmen majoring in engineering that had not selected program of study in engineering (i.e., an undeclared engineering major). The groups were first generated from the SIS dataset. Comparison group students were selected based on their data records on the tenth day of class in the fall and spring: F2012->S2013 (N=428), F2013->S2014 (N=531), and F2014->S2015 (N=430) for a total of 1,389 students. Again, using the tenth day of class in the fall and spring, the F2015->S2016 students were considered beneficiaries of the program or treatment (i.e., academic advising informed by the use of predictive analytics). This group was comprised of 489 students. Next, students were matched to their records in the Mapworks® transition survey (Skyfactor, 2015) dataset. This resulted in 1,422 cases in the sample with 1,185 students in the comparison group and 237 in the treatment group.

Data Analysis Procedures

Specific academic advising units at the institution were selected by university administrators to pilot predictive analytics software for use in academic advising. The particular advising unit this study focused advised first-term freshmen majoring in engineering who did not have a selected program of study (i.e., undeclared engineering major). The treatment effect of academic advising informed by the use of predictive analytics on the tendencies of students to change majors or select a program of study following their first semester was investigated through the use of inferential statistics in order to derive statements about the causal relationships between demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion
characteristics, and their environmental perceptions regarding their academic self-efficacy student integration experiences. The following sections present a description of the analyses used to answer each research question.

**Research Question 1: What are the demographic characteristics of the subjects of this study?**

Using descriptive statistics, this question investigated the demographic characteristics, pre-college academic characteristics, and first-term academic engagement and completion characteristics of first-year, first-term students majoring in engineering who had not selected a program of study. Data were collected on first-year, first-term engineering students without a selected program of study from the fall semesters of 2012 through 2015. Data from the SIS dataset included demographics such as age, gender, and ethnicity. The data also included pre-college academic performance measurements such as ACT composite score, ACT math score, and high school rank.

Finally, also included in the SIS dataset were first-term enrollment characteristics such as whether the student was an honors program member or a member of learning communities in addition to the college they were enrolled in, program of study, number of credits attempted, number of credits completed, and first-term GPA. In addition, spring semester variables germane to the study included the college they were enrolled in and the program of study, if selected. The reason for including these two variables from the spring semester was so that outcome variables could be determined (selection of program and change of major (i.e., transferred to another college other than engineering).

The Mapworks factors included academic self-efficacy, academic integration, and social integration. In addition, individual Mapworks response variables (i.e., commitment to completing degree, self-assessment of math skills, self-assessment of time management and
planning, academic life satisfaction, academic self-efficacy can do well on challenging problems and in hardest course) were examined for correlations with the outcomes variables as well. Those included commitment to completing degree, self-assessment of math skills, self-assessment of time management and planning, academic life satisfaction, academic self-efficacy (can do well on problems and in hardest course), and institutional satisfaction.

The purpose of this descriptive analysis was to compare the first-year, first-term students of the comparison group (F12->S13, F13->S14, F14->S15) with the treatment group (F15->S16). The fall of 2012 was selected as the starting year for data because the Mapworks® transition survey (Skyfactor, 2015) was changed that year and earlier years did not allow for the same factors to be computed.

**Research Question 2: What are the correlational relationships between the independent variables with the dependent variables and what are the mean differences between the comparison and treatment groups?**

It was important to determine which variables and factors may interrelate or have relationships with the outcome variables prior to the propensity score analysis. If it could be shown that the comparison and treatment groups exhibit similar relationships with outcomes in regards to the independent variables, this would help strengthen the principle of conditional independence (i.e., that assignment to one group or the other is independent of the potential outcomes if observable covariates are held constant). In order to evaluate group differences, *t*-tests were conducted on the independent variables and factors from the SIS and Mapworks datasets. In addition, correlational analyses were conducted to examine the strength and direction of relationships between variables and assess linearity (i.e., multicollinearity) between variables (De Veaux, Velleman, & Bock, 2012). A linear relationship should not exist between the independent variables.
Research Question 3: To what extent did the independent variables and factors predict change of major and the selection of a program of study?

For the third research question in this study the goal was to determine which, and to what extent, the variables from the SIS dataset (i.e., demographic, pre-college academic performance measurements, and first-term academic engagement and completion characteristics), and variables and factors (i.e., self-efficacy, academic integration, and social integration) from the Mapworks dataset predicted the selection of a program of study and change of major among first-year, first-term students majoring in engineering with no selected program of study. A regression is a statistical method that describes the relationship between an outcome (i.e. a dependent) variable and one or more explanatory (i.e., independent) variables. The outcome of a logistic regression is not a prediction of a Y-value as in linear regression, but rather the probability of belonging to one of two possible conditions of Y which can take on any value between 0 and 1.

In this study, a binary logistic regression using the forward- stepwise method was applied. Forward stepwise is useful in studies that are exploratory in nature rather than to test an established theory (Burns & Burns, 2008; Mertler & Vannatta, 2010; Peng, So, Stage, & John, 2002). Chen (2005) explained that, when using the forward-stepwise procedure, the explanatory variables will be included in the model one at a time based on the highest Wald or the likelihood-ratio statistics with the entry criterion ($p = .05$). As each new variable is added to the model, all of the existing explanatory variables in the model are evaluated for removal based on the Wald or likelihood-ratio test with the removal criterion ($p = .10$). When there are no more variables that meet the entry or the removal criteria, the algorithm of selecting significant variables ends, leaving the significant variables identified.
First, the dataset \((N=1422)\) was examined for outliers and missing data. As described by Mertler and Vannatta (2013), a preliminary regression was conducted using the variables high school rank, ACT composite, ACT math, and first-term credits attempted to compute the Mahalanobis Distance value. Those cases with a value greater than the chi-square of 18.467 were removed, reducing the sample to 1,366 cases. In order to remove possible confoundness associated with AP credits or other possible forms of previous postsecondary educational experience, the cases were then filtered on the basis of classification year code for the fall term and spring term (both equal to one to indicate students were freshmen the entire academic year) and age (i.e., less than or equal to 18 years of age). This resulted in a dataset of 855 students, with 706 in the control group and 149 in the treatment group.

Descriptive analyses revealed the number of missing cases per independent variable. The number of missing cases per independent variable was small, consistently less than 3.5%. To understand the impact of the removal of missing cases due to list-wise deletion in SPSS, four preliminary logistic regressions were conducted (i.e., SIS variables with Mapworks variables, and SIS variables with Mapworks factors) with each dependent variable (i.e., selection of program of study and change of major). A limited number of cases were removed in each variation of the logistic regression, ranging from 7% to 9% of the sample (Table 4). A reduction of 10% or more may lead to bias (Bennett, 2001), while others used 20% or more (e.g., Peng et al., 2006; Little & Rubin, 2002) as a guideline. The sample size is discussed in further detail after reviewing the assumptions of logistic regression in the next section.
Table 4. Missing case analyses (N=855)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Independents</th>
<th>Missing cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of Major</td>
<td>SIS variables with Mapworks factors</td>
<td>59 (6.9%)</td>
</tr>
<tr>
<td>Change of Major</td>
<td>SIS variables with Mapworks variables</td>
<td>76 (8.9%)</td>
</tr>
<tr>
<td>Selection of Program of Study</td>
<td>SIS variables with Mapworks factors</td>
<td>60 (7.0%)</td>
</tr>
<tr>
<td>Selection of Program of Study</td>
<td>SIS variables with Mapworks variables</td>
<td>77 (9.0%)</td>
</tr>
</tbody>
</table>

Assumptions of logistic regression

While logistic regression does not follow the same strict assumptions of linear regression in regards to the distribution of independent variables, there are certain assumptions that must be followed (Tabachnick & Fidell, 2013). First, a binary logistic regression requires the dependent variable to be dichotomous. In this study, both dependent outcome variables (i.e., selection of a program of study and change of major) were computed to a value of 0 (the outcome did not occur) or 1 (the outcome did occur).

A second assumption is that a logistic model should not be over-fit or under-fit. It is possible to include too many variables leading to a perfectly fit model, or to have too few variables leading to model lacks fit. Thus, only variables that contribute meaningfully to the outcome should be selected. Because this study was exploratory in nature, to determine which variables meaningfully contribute, a forward stepwise regression using likelihood ratio (LR) was applied. LR is considered the criterion least prone to error for estimating the logistic regression (UCLA Institute for Digital Research and Education, 2016; National Centre for Research Methods, 2016). Preliminary logistic regressions were conducted for each model using the full sample (N=855) to determine which independent variables significantly contributed to the selection of a program of study and change of major. The results of these preliminary logistic regressions are included in Appendix C.
Third, a logistic regression requires cases to be independent of one another and the groups to be mutually exclusive. This was true in this study since only one first-year, first-term cohort received the treatment (i.e., F15->S16).

Fourth, the model should have little or no multicollinearity between independent variables. Multicollinearity was examined based on the correlation matrix (Appendix B). A correlation of .70 or greater between two variables requires the removal of one of the variables (Kohler & Kreuter, 2009). For the SIS variables, correlations existed between ACT math and ACT composite, and between first-term credits completed and first-term GPA. For the Mapworks variables, correlations existed between self-efficacy to do well on problems and tasks and self-efficacy to do well in hardest course. The Mapworks variables self-efficacy to do well on problems and tasks and self-efficacy to do well in hardest course also correlated with the factor academic self-efficacy (sub-questions of that factor). Likewise, the Mapworks variable self-assessed satisfaction with academic life correlated with the factor academic integration (a sub-question of that factor).

Fifth, while logistic regressions do not require a linear relationship between dependent and independent variables, there should exist a linear relationship between the independent variables and the log odds (i.e., a one-unit change in an independent variable, should be in linearity with the log odds of the outcome variable). Finally, data points that heavily influence regression coefficients should be examined and removed. As recommended (Burns & Burns, 2008; Berman & Wang, 2011), I inspected the standardized residuals from the preliminary logistic regressions and removed 28 outlier cases at the .01 level with a ZResid $\geq \pm 2.58$. 
Sample sizes for logistic regression

For stepwise logistic regression, a larger ratio of observations to predictors is an accepted in order to improve the analysis and not overestimate the chances of an outcome. Researchers have suggested sample-size guidelines for logistic regression. Long (1997) recommended samples sizes of over 500 (or at least 10 observations per parameter) to achieve the strength needed to calculate the maximum likelihood estimation of the model (i.e., the value of the parameter(s) that give rise to the maximum likelihood of the probability of the outcome). Peng et al. (2002), and Marascuilo and Levin (1983) recommended at least 10 cases per independent variable. Lawley and Maxwell (1971) advised that samples contain at least 51 more cases than the number of variables under consideration. That is, $N - k - 1 \geq 50$, where $N$ is the sample size and $k$ is the number of independent variables. Thorndike (1978) formulated two general rules regarding sufficient sample size. The first rule states that sample size should be equal to 50 plus 10 times the number of independent variables for small datasets. The second rule states the sample size should be equal to 50, plus the square of the number of variables. In this study, there were 855 cases, and approximately 10 independent variables were used at any given time in the logistic regression models, making the sample size sufficiently large.

Assessing and reporting the logistic regression models

To assess the fit of the overall regression model and the significance of the independent variables, the results of the likelihood ratio and Wald tests were examined and interpreted. The likelihood ratio test compares the likelihood of the intercept only model to the likelihood of the model with the independent variables (Eliason, 1993; Hosmer & Lemeshow, 1989; Pampel, 2000). The purpose of the likelihood ratio test in logistic
regression is similar to the $F$-test in the linear regression model. By using it, investigators may conclude if any of the independent variables contribute to the probability of the binary outcome if the $p$-value is less than a predetermined significance level (i.e., .01, .05, or .001). A small value means the model fits the data well. In addition, the -$2\log$ Likelihood is often used to judge the significance of the explanatory variables as they are added to the model (Chen, 2005; Mertler & Vannatta, 2010). The -$2\log$ Likelihood (-2LL), a Chi-square statistic, indicates the probability of obtaining the binary outcome given the established parameter estimates and the fit of the estimated parameters fit on the data. Restated, the -$2\text{LL}$ assesses how likely the same results are given the parameter estimates with the value itself representing the sum of the probabilities associated with the predicted and actual outcomes for each case (Tabachnick & Fidell, 2013). A smaller value for the -$2\text{LL}$ indicates a better model fit. The -$2\text{LL}$ can also be used to compare models, as a basis for measuring the impact of adding additional explanatory variables. As additional explanatory variables are added, the -$2\text{LL}$ may decrease, reflecting a better fit of the model to the data. In addition, the Chi-square is estimated at each step as explanatory variables are added and represents a comparison between the two models from each step. A Chi-square at the 0.05 significance level at each step indicates the added variable is significantly better in predicting outcome than the constant-only model.

Pseudo $R^2$ values are usually reported in the same output table as the -$2\text{LL}$. The Cox & Snell and Nagelkerke are both used to indicate how useful the explanatory variables are in predicting the response variable and can be referred to as a measure of effect size. The maximum for the Cox & Snell $R^2$ is less than 1. The Nagelkerke $R^2$ is adjusted to range from 0 to 1. A higher $R^2$ reflects the model fits the data better, implying a larger effect size.
The Hosmer-Lemeshow (H-L) statistic is also used to assess the goodness-of-fit of a model. The test is similar to a Chi-square test. It partitions the observations into groups of approximately equal size (to avoid low expected frequencies) based on predicted probabilities. The test is calculated using the observed and expected counts for the outcome variable. If the actual number of observations are not significantly (> .05) different from those predicted by the model, the overall fit is considered good. Well-fitting models show non-significance on this test, indicating the mode prediction does not significantly differ from the observed. The H-L statistic works best with samples greater than 400 (Bewick, Cheeck, & Ball, 2005).

The accuracy of the model may also be assessed by examining the proportion of cases classified correctly. Included in the output from SPSS is the classification table which shows how many cases with the observed values of the dependent variable (0 or 1) were correctly predicted. The term “sensitivity” refers to the percentage of cases where the outcome did occur and were correctly predicted. The term “specificity” refers to the percentage of cases where the outcome did not occur were correctly predicted. A cutoff-value can be adjusted to improve classification. A value of 0.5 for the cutoff is basically a random flip of the coin. The value chosen for the cut-off can be changed if previous data exist to inform the decision. In summary, the classification table is useful in demonstrating the extent to which the model with its predictors classifies cases better than the constant-only model.

Similar to the classification table, the extent to which a model distinguishes outcomes is known as discrimination. Discrimination can be assessed through the use and interpretation of the area under the receiver operating characteristic (AUROC) curve (Bewick, Cheeck, & Ball, 2005). The AUROC curve plots sensitivity against specificity: the
higher the curve value, ranging between .45 and 1, the better the model (Bewick, Cheeck, & Ball, 2004). The classification table and the AUROC results were included in the discussion of the model.

The Wald statistic is commonly used to test the significance of the individual logistic regression coefficients (Hosmer & Lemeshow, 1989). The rationale for the Wald test in the logistic regression is to that of the t-test in the linear regression model. The significance value for a variable is assessed and if the value is less than .05, the null hypothesis is rejected and the explanatory variable does make a significant contribution to the prediction. If the regression coefficient (B) significantly differs from zero, the corresponding explanatory variable more significantly contributes to the probability of the outcome variable. In addition, the odds ratio (Exp(B)) is a measure of effect size. With respect to the explanatory variables, the odds ratio reflect the relative importance of the independent variable in terms of effect on the outcome (i.e. dependent) variable. Both of these values were reported in the study.

In addition, Peng et al. (2002) emphasized the necessity of including the standard error:

The absence of the standard error for the parameter estimate means that readers will not be able to assess the stability of these parameter estimates. The instability of parameter estimates can be caused by the low observation to predictor ratio, as we addressed previously. Missing this piece of information undermined the utility of the logistic regression results. Similarly, by not including the Y-intercept in an unstandardized equation, the authors of six articles offered incomplete information. Consequently, readers will not be able to revalidate the result with another sample or at another time/place. This, too, undermined the utility of logistic regression findings. (p. 281)

This study included the standard errors (S.E.) as well as the Y-intercept of the constant model in the results. In addition, as recommended by Mertler and Vannatta (2010), this study
included the regression coefficients of significant explanatory variables in the model along with their respective B values, Wald statistic, degrees of freedom, level of significance, and the odds ratio.

The following logistic regression models (i.e., additive form) were used in this study:

\[
\begin{align*}
\text{logit}(SEL_{PGM_i}) &= \alpha_i + D_i\beta_d + P_i\beta_p + F_i\beta_f + Mv_i\beta_m \\
\text{logit}(SEL_{PGM_i}) &= \alpha_i + D_i\beta_d + P_i\beta_p + F_i\beta_f + Mf_i\beta_m \\
\text{logit}(CHG_{MAJ_i}) &= \alpha_i + D_i\beta_d + P_i\beta_p + F_i\beta_f + Mv_i\beta_m \\
\text{logit}(CHG_{MAJ_i}) &= \alpha_i + D_i\beta_d + P_i\beta_p + F_i\beta_f + Mf_i\beta_m \\
\end{align*}
\]

where \( \alpha_i \) is the Y-intercept and \( \beta \) are the coefficients for each independent variable.

The covariates are described in Table 5. Variables that were meaningful in predicting the dependent outcomes for students in the comparison and treatment groups were subsequently included for use in the propensity score analysis.

Table 5. Description of covariates

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEL_PGM</td>
<td>Whether a student selects a program of study in engineering (binary outcome).</td>
</tr>
<tr>
<td>CHG_MAJ</td>
<td>Whether a student changes to an alternate major (binary outcome).</td>
</tr>
<tr>
<td>D</td>
<td>Student demographics: age, gender, and ethnicity</td>
</tr>
<tr>
<td>P</td>
<td>Pre-college academic performance measurements: ACT composite score, ACT math score, and high school rank.</td>
</tr>
<tr>
<td>F</td>
<td>First-term enrollment characteristics: learning community memberships, honors program member, number of first-term credits attempted, number of first-term credits completed, and fall term GPA.</td>
</tr>
<tr>
<td>Mf</td>
<td>Mapworks factors (academic integration, social integration, academic self-efficacy).</td>
</tr>
<tr>
<td>Mv</td>
<td>Mapworks variables (commitment to completing degree, self-assessment of math skills, self-assessment of time management and planning, academic life satisfaction, academic self-efficacy: can do well on problems, self-efficacy: can do well in hardest course, and institutional satisfaction)</td>
</tr>
</tbody>
</table>
Research Question 4: What is the impact of receiving academic advising informed by predictive analytics on change of major and the selection of a program of study?

The goal of the fourth research question in this study was to measure the impact of advising informed by the use of predictive analytics on first-year, first-term students majoring in engineering with no selected program. The outcomes measured for impact were change of majors and selection of a program of study. Pre-treatment variables were drawn from the demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables and factors from Mapworks.

Because the data were drawn from different points in time, in order to maintain “unconfoundedness” (Rosenbaum & Rubin, 1983), independent variables from the data that could have been influenced in some way by the treatment were not used. For example, after a student received academic advising based on predictive analytics, the number of credits completed could have been influenced since students may change their approach, so that variable was excluded from use. The results of the t-tests on the independent variables revealed differences between the treated beneficiaries (F15->S16 students) and the comparison group (F12->S13, F13->S14, F14->S15 students) and were interpreted for unconfoundness.

Propensity score matching (PSM) is an appropriate type of analysis for a quasi-experimental study utilizing observational data for two reasons. First, as with PSM studies, a non-random set of students (F14->S15 students) were chosen to receive the benefit of a treatment (i.e., advising informed by the use of predictive analytics). The second reason for using propensity scores is that measurements based purely on a logistic regression model could lead to estimates influenced by unobserved variables or factors correlated to the
outcome variable. These confounding variables are possible in studies that span across different years. Thus, the propensity score method requires the use of all observed variables if possible, namely, those known to measure the impact of the treatment.

PSM constructs a statistical comparison group based on a model of the probability (Pr) of participating in the treatment \( (T_i) \) derived from the observed covariates \( X_i \):

\[
P(X_i) = \Pr(T_i = 1|X_i)
\]

Where \( X_i \) represents the covariates from the SIS and Mapworks datasets, and \( T_i \) signifies whether the treatment (academic advising informed by predictive analytics) was received.

Although this approach does not solve the issue of unobservable variables entirely, it is an effective way to control for the observed heterogeneity in the type of student receiving the treatment (Rosenbaum & Rubin, 1983; Rubin, 1997). Thus, the propensity score is a number that “indicates the extent to which one person is similar to another along a collection of observed characteristics” (Agodini & Dynarski, 2004, p. 180). There are two assumptions related to PSM reviewed in the following sections.

**Assumptions of propensity score matching**

**Conditional independence.** Conditional independence implies that a given set of observable covariates used within the PSM should not be influenced by the treatment itself. Adhering to this assumption, enables a causal interpretation of the adjusted differences between the treatment and control group. Rosenbaum and Rubin (1983) referred to the independence between the covariates and the outcomes as “unconfoundedness.” Restated, the probability of participation in the program (i.e., the propensity score) should be based entirely on the observed pre-treatment characteristics and those not biased through self-
selection or unobservable treatment-related conditions. If unobserved characteristics exist which influence the effect of the program, then conditional independence cannot be assumed, and the PSM is not appropriate to use as a basis for causal interpretation.

In this study, the variables from the SIS that were selected to use in the PSM analysis were: gender (GENDER), ACT composite score (ACT_CMPST), ACT math score (ACT_MATH), high school rank (HS_RANK), first-term credits attempted (FT_CRED_ATT), honors program membership (HP_MEM), and learning community memberships (LC_MEM). The factors from the Mapworks data selected were: academic self-efficacy (ACAD_SE), academic integration (ACAD_INT), and social integration (SOC_INT). The variables selected from the Mapworks data were: commitment to completing degree (DEG_COMM), self-assessment of math skills (SA_MATH), self-assessment of degree of planning (SA_PLAN), self-efficacy in regards to ability to do well in hardest course (SE_HARD_CRSE), and self-assessment of satisfaction with academic life (SA_ACAD_LIFE). The Mapworks® transition survey (Skyfactor, 2015) was administered and results collected prior (i.e. in September) to the use of predictive analytics by academic advising (i.e. in October), thereby the treatment did not influence survey responses. In total, six PSM analyses were conducted using the SIS variables paired with either the Mapworks factors or with the Mapworks variables.

**Region of common support.** PSM requires an informed understanding of the covariates that influence participation in an intervention as well as a substantial overlap in the propensity scores between the treatment and comparison groups. This overlap is referred to as the region of common support. In order to determine the extent of the overlap in the region of common support between the treatment and comparison groups, a density plot of
the distribution of the propensity scores for the treatment and control groups is constructed and examined (Lechner, 2001). Density plots are included with results in Chapter 4.

**Sample size.** Like other statistical methods, the results from propensity score matching tend to improve with larger sample sizes since a larger pool leads to an improved balance of the observed covariates between treated and control groups (Rubin, 1997). Based on Monte Carlo simulations, Zhao (2004) determined that propensity score matching requires an N of at least 500 while other studies have found PSM to work sufficiently well with smaller samples (Dehejia & Wahba, 2002; Pirracchio, Resche-Rigon, & Chevret, 2012). A common view is to retain as many subjects as possible in the comparison group so as to increase the possibility of finding a closer match in the treatment group (Khandker, Koolwal, & Samad, 2010).

In addition, the number of observations per covariate impacts the model’s estimation strength (O’Connell & Amico, 2010). For categorical variables, researchers recommend a minimum sample size of 10 observations per predictor variable (Aldrich & Nelson, 1984). For continuous variables, 50 observations per predictor are recommended (Peduzzi, Concato, Holford, & Feinstein, 1996). Thus, researchers should be informed by the literature as well as their own preliminary logistic regression analyses to determine which subset of variables are most significant as predictors, and subsequently, for use in the propensity score analysis. Regardless of the sample size, a key factor is to check balance between groups with regards to the covariates used using t-tests or the standardized difference if small samples are used (Austin, 2009). Stata computes t-tests as part of its PSM calculations automatically.

The PSM utilizes a probit regression model to compute the propensity scores. The number of continuous variables used in this study was less than five and the number of
ordinal variables used was less than five. Therefore, the sample size (i.e., \( N=827 \) with outliers removed as determined by standardized residuals, \( Z_{\text{resid}} > 2.58 \)) satisfies the recommended sample size.

**Steps in propensity score matching**

PSM consists of four phases: (1) estimating the probability of a participant’s exposure to the treatment (i.e., the propensity score) for each participant in the sample; (2) selecting a matching algorithm to align treatment participants with comparison participants in order to construct a comparison; (3) checking for balance in the variables between the treatment and comparison groups; and (4) estimating the program effect and interpreting the results.

**Computing the propensity score.** Propensity scores are constructed using a logit or probit regression used to estimate the probability of a unit’s participation in the treatment condition based on a set of observable variables. In order for the propensity scores to correctly estimate the probability of participation, the covariates included in the propensity score estimation should be well thought out and meaningful. Including too many variables should be avoided to prevent over-specification of the model and amplified standard errors. Additionally, variables that may somehow have been influenced or affected by the treatment itself should be excluded. A scatterplot or \( t \)-test of the selected variables comparing the treatment group with the comparison group can be used to determine whether a covariate may have been influenced by the treatment. Subsequently, the logit or a probit regression is then applied to obtain the predicted probabilities (i.e., propensity scores). Once propensity scores are obtained, reflecting a balanced area of common supports, various matching algorithms can be selected to measure the impact of the treatment variable.
Selecting a matching algorithm. Once the propensity scores are estimated, the treatment group cases are matched with the comparison group cases based on their respective propensity scores. The choice of matching algorithm may vary and include:

- **Nearest neighbor:** Each treatment case is matched to a comparison case with the closest propensity score. Comparison cases for whom there are no treatment case matches with a sufficiently similar score are discarded from the sample; the same is true for treatment cases for which there is no similarly matched comparison case. However, matching may be modified to use a “with replacement” or “without replacement” approach. With the former, a comparison case can be used more than once to match a treatment case whereas in the latter, a comparison case is used only once for a match and then is ignored. Thus, matching involves a trade-off between bias (i.e. standard error) and variance (i.e., standard deviation). When using “with replacement” the average quality of matching increases (i.e., the precision) and bias decreases. Matching “with replacement” is of particular interest when the propensity score distribution is very different between the treatment and comparison group because this method reduces the number (i.e., size) of distinct non-participants used to construct the counterfactual outcome and thereby increases the variance of the estimator (Smith & Todd, 2005). When “without replacement” is selected, comparison cases are matched to one single treatment case. A problem related to matching “without replacement” is that estimates depend on the order in which observations are matched. Thus, when using “without replacement” it should be ensured that the ordering of matches is random. In general, the nearest-neighbor approach faces the risk of bad matches if the closest neighbor is far away. This is
avoided by setting a tolerance level (i.e., caliper) on the maximum propensity score distance akin to the following matching method.

- **Stratification:** This procedure partitions the region of common support into different intervals and calculates the impact of the treatment within each interval. Specifically, within each interval, the effect of the program is the mean difference in outcomes between treated and comparison cases. A weighted average of these interval impact estimates yields the overall program impact by calculating the number of participants in each interval as the weights. Cochrane and Chambers (1965) stated that five strata are usually enough to remove 95% of the bias associated with a single covariate. Imbens (2004) supported the use of five strata to remove most of the bias associated with all covariates.

- **Kernel:** For each treatment case, a weighted average of the outcome of all comparison cases is derived. The weights are based on the distance of the comparison case propensity score to that of the treatment case, with the highest weight given to those with scores closest to the treatment case. One major advantage of kernel matching is that lower variance is achieved because more information is used. A drawback mentioned is that observations may be included that are bad matches (Caliendo & Kopeinig, 2005).

- **Radius:** A maximum propensity score radius is established and then comparison cases within a given radius of a treatment case are matched to that treatment case. Dehejia and Wahba (2002) suggested that all of the comparison cases within the radius be used. Bad matches are avoided by adding a smaller caliper size. However, if fewer matches are located, the variance of the estimates increases. A benefit of the
radius approach is that it uses only as many comparison cases as there are available within the caliper and thereby allows the use of extra or fewer units when good matches are or not available. It has the advantage of using more than one nearest neighbor (i.e., oversampling), but avoids the risk of bad matches.

- **Trade-offs:** Trade-offs between bias and variance in the various matching methods are summarized by Caliendo and Kopeinig (2005) in Figure 3. Caliendo and Kopeinig (2005) recommended trying multiple matching approaches.

The performance of different matching estimators varies case-by-case and depends largely on the data structure at hand (Zhao, 2000). To give an example, if there are only a few control observations, it makes no sense to match without replacement. On the other hand, if there are a lot of comparable untreated individuals it might be worth using more than one nearest neighbour (either by oversampling or kernel matching) to gain more precision in estimates. (p. 12)

<table>
<thead>
<tr>
<th>Decision</th>
<th>Bias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest neighbor matching:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>multiple neighbours / single neighbor</td>
<td>(+) / (−)</td>
<td>(−) / (−)</td>
</tr>
<tr>
<td>with caliper / without caliper</td>
<td>(−) / (+)</td>
<td>(+) / (−)</td>
</tr>
<tr>
<td>Use of control individuals:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with replacement / without replacement</td>
<td>(−) / (+)</td>
<td>(+) / (−)</td>
</tr>
<tr>
<td>Choosing method:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN-matching / Radius-matching</td>
<td>(−) / (+)</td>
<td>(+) / (−)</td>
</tr>
<tr>
<td>KM or LLM / NN-methods</td>
<td>(+) / (−)</td>
<td>(−) / (−)</td>
</tr>
<tr>
<td>Bandwith choice with KM:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>small / large</td>
<td>(−) / (+)</td>
<td>(+) / (−)</td>
</tr>
</tbody>
</table>

KM: Kernel Matching, LLM: Local Linear Matching
NN: Nearest Neighbour
Increase: (+), Decrease (−)


Figure 3. Matching methods trade-offs in terms of bias and efficiency
This study utilized all of the matching methods for comparative purposes using the default settings (i.e., a radius of 0.1, bootstrap replications of 50) using the software application, Stata.

**Checking for Balance.** Once units are matched, the covariates of the treatment and comparison groups should not be significantly different. The matched units in the treatment and comparison groups are statistically compared using a \( t \)-test to compare the means of all covariates included in the propensity score in order to determine if the means are statistically similar in the treatment and comparison groups. If balance is not achieved, meaning the covariates are significantly different as revealed through \( t \)-tests, a different matching option or specification should be tried until the sample is balanced. Stata will include these results in the output by simply adding a flag to produce detailed output. These results are not usually reported, but achieving balance is required in order to calculate the effect of the treatment.

**Reporting results and estimating program effect**

Following the estimation of propensity scores, the selection of a matching algorithm and the achievement of balance, the impact of the intervention may be inferred by computing the differences in outcome between each treated unit and its neighbor or neighbors from the constructed comparison group. The difference in means of the subjects who participated in the intervention and those who did not can then be interpreted as the impact of the program.

Whereas a logistic regression estimates the average treatment effect (ATE) across the entire sample, a propensity score analysis instead computes the average treatment on the treated (ATT). The ATE is not as an insightful measurement of a program’s impact because it compares all individuals in the sample, who may differ substantially individually. The
ATT is preferred because it compares individuals in the treatment group with similar probabilities as those in the control group. In support of the use of ATT, the standard error is calculated and reported because the estimated variance of the treatment effect is influenced by factors such as the variance in the estimation of the propensity score, the imputation of the common support, and the matching technique used. The calculation of standard error is accomplished through a technique known as bootstrapping (Lechner, 2002; Caliendo and Kopeinig, 2008). Each bootstrap repetition incorporates the re-estimation of the results including the first steps of the estimation (propensity score, common support, etc.). Repeating the bootstrapping \( N \) times leads to \( N \) bootstrap samples and in our case \( N \) estimated average treatment effects. The distribution of these means approximate the sampling distribution and the standard error of the population mean. Bootstrapping can be time-consuming so it may not always be feasible. Bootstrapping in Stata occurs by specifying the `bootstrap` option and the number of repetitions. This option is available to the suite of `att` commands (i.e. `attnd, atts, attr, attk`). Hedges (1992) suggested a value of 2000 for the number of bootstrap repetitions to achieve statistical significance. However, others (Pattengale, Alipour, Bininda-Emonds, Moret, & Stamatakis, 2009) stated that 100 to 500 replications is sufficient to achieve values that correlate at better than 99.5\% for ML inference. Pragmatically speaking, the number of repetitions to select was described by Gould and Pitblado (2015) as follows:

In terms of the number of replications, there is no fixed answer such as “250” or “1,000” to the question. The right answer is that you should choose an infinite number of replications because, at a formal level, that is what the bootstrap requires. The key to the usefulness of the bootstrap is that it converges in terms of numbers of replications reasonably quickly, and so running a finite number of replications is good enough - assuming the number of replications chosen is large enough. The previous statement contains the key to choosing the right number of replications:
1. Choose a large but tolerable number of replications. Obtain the bootstrap estimates.
2. Change the random-number seed. Obtain the bootstrap estimates again, using the same number of replications.
3. Do the results change meaningfully? If so, the first number you chose was too small. Try a larger number. If results are similar enough, you probably have a large enough number. (p. 1)

For practical purposes of this study, the value of 50 was used for bootstrapping since higher values did not change the results meaningfully. In this study, the ATT was reported with bootstrapped standard errors.

**Delimitations**

This study attempted to determine the impact of predictive analytics on the selection of a program of study by first-term freshmen majoring in engineering. The impact on selection of program of study and change of major was investigated. The study did not purport to identify other outcomes due to the shortened span of the treatment applied to the subjects (i.e., first-year, first-term freshmen majoring in engineering who had not selected a program of study in engineering and who were beneficiaries of advising informed by predictive analytics). Thus, the context of this study was limited to this specific set of students. A decision was also made to not select students prior to fall 2012 due to significant changes made to the Mapworks survey instrument in 2012. The consistency of the Mapworks survey instrument from 2012 through 2015 enabled the researcher to use similar factors to be computed for use in this study.

**Limitations**

Because secondary datasets spanning multiple years were used in this study, only the most objective variables were selected that reflect and measure student outcomes. Although unobserved factors may influence the outcome variables as reported through use of logistic
regression, the use of propensity score matching mitigates those unobserved effects by balancing treated and control students.

In addition, every effort was made to include as many treatment cases (i.e., first-year, first-term freshmen majoring in engineering who had not selected program of study) in order to retain a large comparison pool for matching with the treatment cases in the propensity score analysis, thereby strengthening the results. Within the sample itself, a limitation was that the sample tended to be largely male (78%) and white (83.4%). In addition, this study focused on an academic discipline within a specific college. This was to ensure the first-year curriculum and advising experiences were consistent to minimize unobserved, confounding factors. Finally, because MAP-Works data is self-reported, some bias may be attributable to personal responses.

**Ethical Considerations**

Approval to use human subjects in this study was approved by the Institutional Review Board (see Appendix D). The researcher was aware of the confidential nature of the data, and complied with all policies and restrictions set forth regarding confidential information. University ID numbers were removed by institutional staff and replaced with non-identifiable pseudo-identifiers. No student data were reported that could potentially violate the anonymity of the individuals.

**Summary**

This chapter described the methodology for this study, including the philosophical paradigm from which the research was oriented as the basis for conducting empiric comparisons between the two groups in this study. The sample was comprised of first-year, first-term freshmen majoring in engineering who had not selected a specific program of study
in engineering. Research questions and data analysis methods were described. The results of this study may be used to inform higher education institutions with practical strategies and policy decision-making when considering how to implement and utilize resources to improve student success with first-year, first-term freshmen in engineering who have not selected a program of study as they transition to the postsecondary. The findings also have implications for academic advising as advisers work with students and seek to retain them by recommending certain programs of study or a change in major. A goal of the study was to inform practices and policies that lead to successful outcomes for students as they transition and immerse themselves in their new academic setting during their first term in postsecondary education. The next chapter will discuss the results and findings of the analyses conducted.
CHAPTER 4. RESULTS

This study focused on the relationships between the following: demographics characteristics (i.e., age, gender); pre-college academic characteristics (i.e., ACT composite score, ACT math score, high school rank); first-term academic engagement and completion characteristics (i.e., honors program membership; learning community memberships; number of first-term credits attempted and number of first-term credits earned; first-term cumulative GPA)—and, environmental variables (i.e., commitment to completing degree, self-assessment of math skills, self-assessment of time management and planning academic life satisfaction; academic self-efficacy can do well on challenging problems and in hardest course)—or, environmental factors (i.e., academic self-efficacy, academic integration and social integration). These characteristics and environmental variables were then used in the analyses to predict the outcomes of first-year, first-term freshmen who are majoring in engineering but had not yet selected a specific program of study within engineering. The outcomes of interest were change of major (i.e., leaving engineering and enrolling in a different college at the institution) and the selection of a program of study (i.e., selecting a specific engineering major). The results of this study may be used to contribute to the body of research on the use of predictive analytics in higher education and inform policy-makers on the implementation and use of analytics.

The impetus for conducting this research was to explore and investigate student success based on the following question: What variables and factors influence student tendencies regarding change of major and selection of a program of study? This question led to the formation of four research questions. Descriptive statistics were conducted to answer Research Question 1. Research Question 2 examined the correlational relationships between
the independent variables and the dependent outcomes. In addition, mean differences between the comparison and treatment groups were analyzed. Research Question 3 explored which variables and factors significantly predicted the odds of change of major and the selection of a program of study. Research Question 4 explored the impact of academic advising informed by the use of predictive analytics on change of major and the selection of a program of study. The results are presented based on each research question.

**Research Question 1: What are the demographic characteristics of the subjects in this study?**

As shown in Table 6, data from the demographic characteristics, pre-college academic characteristics, and first-term academic engagement and completion characteristics were used to investigate this question. Descriptive statistics were also conducted on the Mapworks variables and factors. The descriptive statistics were produced using SPSS.

Regarding gender, there were 645 males (78%) and 182 females (22%).

Regarding ethnicity, the sample was comprised of mainly white individuals (n=690, 83.4%). The other ethnic categories contained only a small number of cases (anywhere from 2 to 29).

Learning community (LC) memberships ranged from 0 to 3, representing the number of learning communities in which a student is a member. There were 11.2% not in an LC, 79.4% in one LC, 8.2% in two LCs, and 1.1% in three LCs.

**Research Question 2: What are the correlative relationships between the independent variables with the dependent variables and what are the mean differences between the comparison and treatment groups?**

Pearson correlations were conducted to explore the relationships between the variables. In addition, a $t$-test was conducted to investigate similarities and differences between the groups. Both analyses were conducted using SPSS. The correlative analyses
Table 6. Results based on analysis of descriptive statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>1</td>
<td>2</td>
<td>827</td>
<td>1.2201</td>
<td>0.4145</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>1</td>
<td>8</td>
<td>817</td>
<td>3.43</td>
<td>1.286</td>
</tr>
<tr>
<td>AGE</td>
<td>16</td>
<td>18</td>
<td>827</td>
<td>17.84</td>
<td>0.393</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>27</td>
<td>99</td>
<td>817</td>
<td>76.97</td>
<td>15.441</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>15</td>
<td>35</td>
<td>805</td>
<td>26.44</td>
<td>3.232</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>17</td>
<td>36</td>
<td>804</td>
<td>27.49</td>
<td>3.582</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>0</td>
<td>1</td>
<td>827</td>
<td>0.06</td>
<td>0.229</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>0</td>
<td>3</td>
<td>827</td>
<td>0.99</td>
<td>0.488</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>12</td>
<td>26</td>
<td>827</td>
<td>15.79</td>
<td>2.26</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>0</td>
<td>18</td>
<td>827</td>
<td>13.48</td>
<td>2.647</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0</td>
<td>4</td>
<td>827</td>
<td>2.8195</td>
<td>0.8459</td>
</tr>
<tr>
<td>DEG_COMM (Q002)</td>
<td>1</td>
<td>7</td>
<td>821</td>
<td>6.29</td>
<td>1.068</td>
</tr>
<tr>
<td>SA_MATH (Q016)</td>
<td>2</td>
<td>7</td>
<td>824</td>
<td>5.76</td>
<td>0.947</td>
</tr>
<tr>
<td>SA_FLW_THRU (Q019)</td>
<td>2</td>
<td>7</td>
<td>827</td>
<td>5.71</td>
<td>1.011</td>
</tr>
<tr>
<td>SA_PLAN (Q022)</td>
<td>1</td>
<td>7</td>
<td>823</td>
<td>4.94</td>
<td>1.451</td>
</tr>
<tr>
<td>SE_PROBS_TASK (Q039)</td>
<td>1</td>
<td>7</td>
<td>821</td>
<td>5.27</td>
<td>1.047</td>
</tr>
<tr>
<td>SE_HARD_CRSE (Q040)</td>
<td>1</td>
<td>7</td>
<td>819</td>
<td>4.77</td>
<td>1.198</td>
</tr>
<tr>
<td>SA_ACAD_LIFE (Q155)</td>
<td>1</td>
<td>7</td>
<td>799</td>
<td>5.29</td>
<td>1.339</td>
</tr>
<tr>
<td>INST_SAT (Q159)</td>
<td>1</td>
<td>7</td>
<td>798</td>
<td>5.94</td>
<td>1.207</td>
</tr>
<tr>
<td>ACAD_SE (F09)</td>
<td>1</td>
<td>7</td>
<td>821</td>
<td>5.1549</td>
<td>0.9559</td>
</tr>
<tr>
<td>ACAD_INT (F13)</td>
<td>2</td>
<td>7</td>
<td>802</td>
<td>5.5538</td>
<td>0.9838</td>
</tr>
<tr>
<td>SOC_INT (F14)</td>
<td>1</td>
<td>7</td>
<td>801</td>
<td>5.5499</td>
<td>1.2672</td>
</tr>
</tbody>
</table>

were conducted using the entire sample (i.e., combined) and then by group (i.e., comparison and treatment). To clarify, the first correlational analysis utilized Mapworks variables across the entire sample (Table 7) and then by group (Table 8). The second correlational analysis utilized Mapworks factors across the entire sample (Table 9) and then by group (Table 10).

As shown in Table 7, the first-term academic engagement and completion variables that significantly correlated positively with selection of program of study were first-term credits completed \((r = .159, p < .01)\) and first-term GPA \((r = .116, p < .01)\). The Mapworks variables of commitment to completing degree \((r = .166, p < .01)\), follows through on what they say \((r = .083, p < .01)\), self-assessment of degree of planning \((r = .129, p < .01)\), self-
<table>
<thead>
<tr>
<th>Variable</th>
<th>SEL_PGM</th>
<th>CHG_MAJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>-0.024</td>
<td>0.033</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>0.039</td>
<td>-0.018</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.078*</td>
<td>0.041</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>0.050</td>
<td>-0.054</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.053</td>
<td>-0.109**</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>-0.018</td>
<td>-0.114**</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>-0.048</td>
<td>-0.019</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>0.038</td>
<td>0.045</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>-0.023</td>
<td>0.092*</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>0.159**</td>
<td>-0.208**</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0.116**</td>
<td>-0.120**</td>
</tr>
<tr>
<td>DEG_COMM</td>
<td>0.166**</td>
<td>-0.075*</td>
</tr>
<tr>
<td>SA_MATH</td>
<td>0.057</td>
<td>-0.131**</td>
</tr>
<tr>
<td>SA_FLW_THRU</td>
<td>0.083*</td>
<td>-0.049</td>
</tr>
<tr>
<td>SA_PLAN</td>
<td>0.129**</td>
<td>-0.006</td>
</tr>
<tr>
<td>SE_PROBS_TASK</td>
<td>0.065</td>
<td>-0.120**</td>
</tr>
<tr>
<td>SE_HARD_CRSE</td>
<td>0.037</td>
<td>-0.146**</td>
</tr>
<tr>
<td>SA_ACAD_LIFE</td>
<td>0.140**</td>
<td>-0.135**</td>
</tr>
<tr>
<td>INST_SAT</td>
<td>0.115**</td>
<td>-0.094**</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001; N=755

Note: Shaded areas indicate results of interest related to this particular analysis.

assessed satisfaction with academic life \((r = .140, p < .01)\), and institutional satisfaction \((r = .115, p < .01)\) significantly correlated positively to the selection of program of study while the variable age weakly correlated negatively \((r = -.078, p < .05)\).

Pre-college academic variables that significantly correlated negatively with change of major were ACT composite \((r = -.109, p < .01)\) and ACT math \((r = -.114, p < .01)\). First-term academic engagement and completion variables that significantly correlated with change of major were first-term credits attempted \((r = .092, p < .05)\), first-term credits completed \((r = -.208, p < .01)\), and first-term GPA \((r = -.120, p < .01)\). Mapworks variables that significantly correlated (all negatively) with change of major were: commitment to completing degree \((r = -.075, p < .05)\), self-assessment of math skills \((r = -.131, p < .01)\), self-efficacy to do well on problems and tasks \((r = -.120, p < .01)\), self-efficacy to do well in
hardest course \((r = -0.146, p < 0.01)\), self-assessed satisfaction with academic life \((r = -0.135, p < 0.01)\), and institutional satisfaction \((r = -0.094, p < 0.01)\).

Correlations were also conducted by group (Table 8) with the same set of variables to determine group similarities or differences because the comparison group was substantially larger than treatment group, a common occurrence in quasi-experimental design studies. The group correlations were similar in regards to: first-term credits completed, first-term GPA, self-assessment of math skills, plans out time, and self-efficacy to do well in hardest course. The group correlations differed in regards to commitment to completing degree, self-efficacy to do well on problems and tasks, and institutional satisfaction.

Table 8. Correlations between variables (Mapworks variables, by group)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SEL_PGM</th>
<th>CHG_MAJ</th>
<th>SEL_PGM</th>
<th>CHG_MAJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>-0.022</td>
<td>0.026</td>
<td>-0.080</td>
<td>0.081</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>0.038</td>
<td>-0.002</td>
<td>0.036</td>
<td>-0.103</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.001</td>
<td>0.019</td>
<td>-0.077</td>
<td>0.117</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>0.016</td>
<td>-0.040</td>
<td>0.236*</td>
<td>-0.144</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.070</td>
<td>-0.107**</td>
<td>0.062</td>
<td>-0.128</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>-0.030</td>
<td>-0.116**</td>
<td>0.085</td>
<td>-0.108</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>-0.063</td>
<td>-0.025</td>
<td>-0.021</td>
<td>0.014</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>0.030</td>
<td>0.026</td>
<td>-0.011</td>
<td>0.175</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>-0.008</td>
<td>0.085*</td>
<td>0.065</td>
<td>0.140</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>0.138**</td>
<td>-0.211**</td>
<td>0.290**</td>
<td>-0.183</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0.090*</td>
<td>-0.115**</td>
<td>0.271**</td>
<td>-0.147</td>
</tr>
<tr>
<td>DEG_COMM</td>
<td>0.175**</td>
<td>-0.078*</td>
<td>0.125</td>
<td>-0.057</td>
</tr>
<tr>
<td>SA_MATH</td>
<td>0.028</td>
<td>-0.097*</td>
<td>0.249**</td>
<td>-0.344**</td>
</tr>
<tr>
<td>SA_FLW_THRU</td>
<td>0.059</td>
<td>-0.015</td>
<td>0.179</td>
<td>-0.274**</td>
</tr>
<tr>
<td>SA_PLAN</td>
<td>0.099*</td>
<td>0.006</td>
<td>0.271**</td>
<td>-0.066</td>
</tr>
<tr>
<td>SE_PROBS_TASK</td>
<td>0.025</td>
<td>-0.071</td>
<td>0.280**</td>
<td>-0.440**</td>
</tr>
<tr>
<td>SE_HARD_CRSE</td>
<td>0.012</td>
<td>-0.115**</td>
<td>0.167</td>
<td>-0.334**</td>
</tr>
<tr>
<td>SA_Acad_Life</td>
<td>0.125**</td>
<td>-0.116**</td>
<td>0.207*</td>
<td>-0.251**</td>
</tr>
<tr>
<td>INST_SAT</td>
<td>0.117**</td>
<td>-0.090*</td>
<td>0.081</td>
<td>-0.118</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001; N=641 (comparison); N=114 (treatment)
Note: Shaded areas indicate results of interest related to this particular analysis.
Finally, a noteworthy correlation was the variable self-assessed satisfaction with academic life which significantly correlated in both groups with selection of program of study (positively) and change of major (negatively).

Next, a correlational analysis using the Mapworks factors, rather than the Mapworks variables, was conducted on the entire sample (Table 9). The variables first-term credits completed ($r = .164, p < .01$) and first-term GPA (i.e., $r = .119, p < .01$) were significantly correlated (positively) with the selection of program of study. The Mapworks factors, academic integration ($r = .156, p < .01$) and social integration ($r = .121, p < .01$) were also significantly (positively) correlated with the selection of program of study. The variable, age, again correlated negatively ($r = -.076, p < .05$) with the selection of program of study.

The results for change of major indicate that ACT composite ($r = -.115, p < .01$) and ACT math ($r = -.118, p < .01$) both significantly correlated negatively. The variables first-

Table 9. Correlations between variables (Mapworks factors, entire sample)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SEL_PGM</th>
<th>CHG_MAJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>-0.018</td>
<td>0.029</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>0.043</td>
<td>-0.015</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.076*</td>
<td>0.034</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>0.047</td>
<td>-0.055</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.045</td>
<td>-0.115**</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>-0.009</td>
<td>-0.118**</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>-0.046</td>
<td>-0.019</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>0.042</td>
<td>0.044</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>-0.018</td>
<td>0.092*</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>0.164**</td>
<td>-0.211**</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0.119**</td>
<td>-0.120**</td>
</tr>
<tr>
<td>ACAD_SE</td>
<td>0.062</td>
<td>-0.163**</td>
</tr>
<tr>
<td>ACAD_INT</td>
<td>0.156**</td>
<td>-0.116**</td>
</tr>
<tr>
<td>SOC_INT</td>
<td>0.121**</td>
<td>-0.049</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$, *** $p < .001$; N=755
Note: Shaded areas indicate results of interest related to this particular analysis.
term credits completed \((r = -.211, p < .01)\) and first-term GPA \((r = -.120, p < .01)\)

significantly correlated negatively as well. The Mapworks factors academic self-efficacy \((r = -.163, p < .01)\) and academic integration \((r = -.116, p < .01)\) also significantly correlated negatively. Only first-term credits attempted \((r = .092, p < .05)\) correlated positively.

As mentioned previously, correlations were also conducted by group to examine group similarities or differences (Table 10). As shown previously in Table 9, ACT composite and ACT math significantly correlated (negatively) with change of major in the comparison group only. First-term GPA significantly correlated (positively) with the selection of program of study in both groups. Academic self-efficacy significantly correlated (negatively) with change of major in both groups. Finally, first-term credits completed and academic integration correlated with the selection of program of study (positively) and change of major (negatively) in both groups.

**Table 10. Correlations between variables (Mapworks factors, by group)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comparison</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEL_PGM</td>
<td>CHG_MAJ</td>
</tr>
<tr>
<td>GENDER</td>
<td>-0.015</td>
<td>0.024</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>0.040</td>
<td>0.004</td>
</tr>
<tr>
<td>AGE</td>
<td>0.008</td>
<td>0.020</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>0.019</td>
<td>-0.043</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.063</td>
<td>-0.108**</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>-0.022</td>
<td>-0.114**</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>-0.061</td>
<td>-0.024</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>0.032</td>
<td>0.027</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>-0.004</td>
<td>0.095*</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>0.141**</td>
<td>-0.206**</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0.095*</td>
<td>-0.115**</td>
</tr>
<tr>
<td>ACAD_SE</td>
<td>0.031</td>
<td>-0.115**</td>
</tr>
<tr>
<td>ACAD_INT</td>
<td>0.140**</td>
<td>-0.096*</td>
</tr>
<tr>
<td>SOC_INT</td>
<td>0.112**</td>
<td>-0.047</td>
</tr>
</tbody>
</table>

\* - \(p < .05\), \** - \(p < .01\), \*** - \(p < .001\); \(N=658\) (comparison); \(N=117\) (treatment)

Note: Shaded areas indicate results of interest related to this particular analysis.
The second part of this research question investigated differences between the comparison and treatment groups using a *t*-test. Results from this analysis are presented in Table 11. Regarding demographic differences, there was a significant difference in the variable gender at the *p* < .05 level between the comparison (M = 1.21, SD = 0.4045) and treatment (M = 1.31, SD = 0.4637) groups; *t* (798) = -2.499, *p* = 0.013 indicating the treatment group contained significantly more females. There was also a significant difference in age between the comparison (M = 17.98, SD = 0.156) and treatment (M =

Table 11. Mean differences between comparison and treatment groups

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comparison</th>
<th>Treatment</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M</td>
<td>SD</td>
<td>N</td>
</tr>
<tr>
<td>GENDER</td>
<td>680</td>
<td>1.21</td>
<td>0.4046</td>
<td>120</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>680</td>
<td>3.41</td>
<td>1.258</td>
<td>120</td>
</tr>
<tr>
<td>AGE</td>
<td>680</td>
<td>17.98</td>
<td>0.156</td>
<td>120</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>680</td>
<td>76.74</td>
<td>15.404</td>
<td>120</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>680</td>
<td>26.44</td>
<td>3.253</td>
<td>120</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>680</td>
<td>27.53</td>
<td>3.632</td>
<td>120</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>680</td>
<td>0.05</td>
<td>0.209</td>
<td>120</td>
</tr>
<tr>
<td>LC_MEM</td>
<td>680</td>
<td>0.96</td>
<td>0.456</td>
<td>120</td>
</tr>
<tr>
<td>FT_CRED_ATT</td>
<td>680</td>
<td>16.05</td>
<td>2.262</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>13.48</td>
<td>2.647</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>682</td>
<td>2.817</td>
<td>0.8489</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>6.32</td>
<td>1.067</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>5.77</td>
<td>0.937</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>5.66</td>
<td>1.016</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>4.92</td>
<td>1.449</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>5.26</td>
<td>1.046</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>4.77</td>
<td>1.196</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>5.27</td>
<td>1.342</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>658</td>
<td>5.92</td>
<td>1.228</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>677</td>
<td>5.1366</td>
<td>0.9542</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>677</td>
<td>5.5235</td>
<td>0.9941</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>677</td>
<td>5.5470</td>
<td>1.2555</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>702</td>
<td>0.55</td>
<td>0.498</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>702</td>
<td>0.10</td>
<td>0.298</td>
<td>125</td>
</tr>
</tbody>
</table>

Note: Shaded areas indicate results of interest related to this particular analysis.
17.06, SD = 0.416) groups; $t (798) = 43.163, p < .001$ indicating a slightly younger group
received the treatment.

Regarding first-term academic engagement and completion characteristics, there were
three significant differences. There was a significant difference at the $p < .05$ level for the
variable honors program membership between the comparison (M = 0.05, SD = 0.209) and
treatment (M = 0.10, SD = 0.301) groups; $t (798) = -2.443, p = .016$. There was also a
significant difference at the $p < .001$ level found for the variable learning community
memberships between the comparison (M = 0.96, SD = 0.456) and treatment (M = 1.23, SD
= 0.546) groups; $t (798) = -5.832, p < .001$. Finally, there was a significant difference at the
$p < .001$ level on the variable first-term credits attempted between the comparison (M =
16.05, SD = 2.262) and treatment (M = 14.38, SD = 1.348) groups; $t (798) = 7.866, p < .001$.

Regarding Mapworks variables, there was a significant difference found for the
variable follows through on what they say between the comparison (M = 5.66, SD = 1.016)
and treatment (M = 6.03, SD = 0.822) groups; $t (774 ) = -3.624, p < .001$. There was also a
significant difference at the $p < .05$ level on the variable self-efficacy to do well on problems
and tasks between the comparison (M = 5.26, SD = 1.046) and treatment (M = 5.47, SD =
0.967) groups; $t (774) = -2.076, p = .038$.

With respect to the Mapworks factors, there was a significant difference at the $p < .05$
level on the factor academic self-efficacy between the comparison (M = 5.1366, SD =
0.9542) and treatment (M = 5.3278, SD = 0.9574) groups; $t (796) = -2.029, p = .043$. There
was also a significant difference at the $p < .05$ level on the factor academic integration
between the comparison (M=5.5235, SD=0.9941) and treatment (M = 5.7293, SD = 0.9086)
groups; $t (796) = -2.124, p = .034$. 
Finally, in regards to differences on the dependent variable selection of program of study, there was a significant difference at the $p < .05$ level between the comparison ($M = 0.55, SD = 0.498$) and treatment ($M = 0.66, SD = 0.474$) groups; $t (825) = -2.378, p = .018$, indicating the treatment group selected a program of study more often.

**Research Question 3: To what extent did the independent variables and factors predict change of major and the selection of program of study?**

The main research question was deconstructed into four sub-questions. The purpose for the deconstruction was twofold. One, it was important to validate which, if any, of the Mapworks factors were significant. Two, the deconstruction of the factors enabled a closer examination of the variables were significant. Thus, to determine the extent to which the independent variables in the model predicted the probability of change of major and selection of program, logistic regressions were utilized. The first sub-question follows.

*a. To what extent do demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables predict change of major among first-year, first-term students majoring in engineering with no selected program of study?*

Informed by the previous analyses of correlations and $t$-tests, the independent variables investigated as potential predictors of change of major were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_CMPST
- First-term academic engagement and completion characteristics: FT_CRED_ATT, FT_CRED_COMP, FT_GPA, HP_MEM, LC_MEM
- Environmental variables: SA_MATH, SE_HARD_CRSE, SA_ACAD_LIFE
- Dependent variable: CHG_MAJ (change of major)

Forward stepwise logistic regression using likelihood ratio testing was conducted to determine which independent variables were predictors of change of major (Table 12).

Results indicated that the overall fit of the model with four predictors was moderate ($-2LL = $
Table 12. Change of major with significant SIS variables and Mapworks variables

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>$B$</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>$p$</th>
<th>Odds Ratio (Exp[$B$])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_ATT</td>
<td>0.219</td>
<td>0.055</td>
<td>15.791</td>
<td>1</td>
<td>0.000</td>
<td>1.245</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>-0.210</td>
<td>0.039</td>
<td>28.801</td>
<td>1</td>
<td>0.000</td>
<td>0.811</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.091</td>
<td>0.040</td>
<td>5.235</td>
<td>1</td>
<td>0.022</td>
<td>0.913</td>
</tr>
<tr>
<td>SE_HARD_CRSE</td>
<td>-0.418</td>
<td>0.104</td>
<td>16.124</td>
<td>1</td>
<td>0.000</td>
<td>0.658</td>
</tr>
<tr>
<td>Constant</td>
<td>1.188</td>
<td>1.322</td>
<td>0.808</td>
<td>1</td>
<td>0.369</td>
<td>3.281</td>
</tr>
</tbody>
</table>

420.595) but significantly better in discriminating change of major ($\chi^2(4, N=827) = 62.221, p < .001$) over the constant model, classifying 72.1% of all cases correctly, 61.6% of those who actually changed majors, and 73.2% of those who did not actually change majors as compared to the constant model which correctly accounted for only 10% of those who changed majors. The area under the receiver operating characteristic (AUROC) curve value was 0.760 indicating that on average, the model is interpreted as a fair test of sensitivity against specificity with a 76% probability that students were classified correctly. Confirming the goodness-of-fit, the Hosmer-Lemeshow statistic resulted in a Chi-square of $\chi^2(8, N=827) = 11.284, p = 0.186$. When significance exceeds the $p < .05$ threshold, the null hypothesis is then rejected (i.e., that there is no difference between observed and predicted values), thereby implying that this model’s estimated fit of the data is at an acceptable level.

As shown in Table 12, the significance of the Wald statistic for first-term credits attempted indicates that for every one-credit increase in first-term credits attempted, the odds of change of major significantly increase (i.e., an odds ratio of 1.25 times more likely to change majors). The significance of the Wald statistic for first-term credits completed indicate that for every one credit increase in first-terms credits completed, there is a negative effect on change of major, implying as credits completed increased, the odds of changing majors was decreased (i.e., after inverting the odds ratio, 1.23 times more likely not to
change majors). The significance of the *Wald* statistic for the pre-college enrollment variable of ACT composite score indicates that for each unit increase in ACT composite score, the odds of changing majors decreased (i.e. after inverting the odds ratio, 1.10 times more likely not to change majors). Regarding the Mapworks variables, the significance of the *Wald* statistic for academic self-efficacy regarding hardest course indicates that for each point increase in confidence, the odds of changing majors decreased (i.e. after inverting the odds ratio, 1.51 times more likely not to change majors). The odds ratio for each coefficient indicated a small effect size on the outcome. The Nagelkerke pseudo-$R^2$ of 0.167 reflects there an effect size of noteworthy interest, but slightly short of the generally accepted .2 level to indicate strength.

**b. To what extent do demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental factors predict change of major among first-year, first-term students majoring in engineering with no selected program of study?**

Informed by the previous analyses of correlations and *t*-tests, the independent variables investigated as potential predictors of change of major were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_CMPST
- First-term academic engagement and completion characteristics: FT_CRED_ATT, FT_CRED_COMP, FT_GPA, HP_MEM, LC_MEM
- Environmental factors: ACAD_SE, ACAD_INT, SOC_INT
- Dependent variable: CHG_MAJ (change of major)

Forward stepwise logistic regression using likelihood ratio testing was conducted to determine which independent variables were predictors of change of major (Table 13). The regression results indicated that the overall fit of the model with four predictors was moderate (-2LL = 422.851) but significantly better in distinguishing change of major ($\chi^2(4, N=827) = 61.352, p < .001$) over the constant model, classifying 72.6% of all cases correctly,
Table 13. Change of major with significant SIS variables and Mapworks factors

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_ATT</td>
<td>0.219</td>
<td>0.055</td>
<td>15.729</td>
<td>1</td>
<td>0.000</td>
<td>1.245</td>
</tr>
<tr>
<td>FT_CRED_COMP</td>
<td>-0.203</td>
<td>0.039</td>
<td>27.525</td>
<td>1</td>
<td>0.000</td>
<td>0.816</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.088</td>
<td>0.040</td>
<td>5.011</td>
<td>1</td>
<td>0.025</td>
<td>0.915</td>
</tr>
<tr>
<td>ACAD_SE</td>
<td>-0.505</td>
<td>0.130</td>
<td>15.174</td>
<td>1</td>
<td>0.000</td>
<td>0.604</td>
</tr>
<tr>
<td>Constant</td>
<td>1.641</td>
<td>1.355</td>
<td>1.468</td>
<td>1</td>
<td>0.226</td>
<td>5.162</td>
</tr>
</tbody>
</table>

64.4% of those who actually changed majors, and 73.4% of those who actually did not change majors as compared to the constant model which correctly accounted for only 10% of those who changed majors. The AUROC curve value was 0.769 indicating that on average, the model is interpreted as a fair test of sensitivity against specificity with a 76.9% probability that students were classified correctly. Confirming the goodness-of-fit, the Hosmer-Lemeshow statistic resulted in a Chi-square of \( \chi^2(8, N=827) = 6.608, p = 0.579. \) When significance exceeds the \( p<0.05 \) threshold of significance, the null hypothesis is then rejected (i.e. that there is no difference between observed and predicted values), thereby implying that this model’s estimated fit of the data is at an acceptable level.

As shown in Table 13, the significance of the Wald statistic for first-term credits attempted indicates that for every one credit increase in first-term credits attempted, the odds of change of major significantly increased (i.e. an odds ratio of 1.245 times more likely to change majors). The significance of the Wald statistic for first-term credits completed indicates that for every one credit completed, the odds of changing majors was decreased (i.e., after inverting the odds ratio, 1.23 times more likely not to change majors). The significance of the Wald statistic for the pre-college enrollment variable ACT composite score indicates that for each unit increase in ACT composite score, the odds of changing majors decreased (i.e. after inverting the odds ratio, 1.10 times more likely not to change
majors. Regarding the Mapworks factors, significance of the Wald statistic for academic self-efficacy indicates that for each point increase in academic self-efficacy, the odds of changing majors decreased (i.e., after inverting the odds ratio, 1.66 times more likely not to change majors). The odds ratio for each coefficient indicated a small effect size on the outcome. The Nagelkerke pseudo-$R^2$ of 0.164 reflects there an effect size of noteworthy interest, but slightly short of the generally accepted .2 level to indicate strength.

c. To what extent demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables predict selection of a program of study among first-year, first-term students majoring in engineering with no selected program of study?

Informed by the previous analysis of correlations and $t$-tests, the independent variables investigated as potential predictors of the selection of program of study were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_MATH
- First-term academic engagement and completion characteristics: FT_CRED_ATT, FT_CRED.COMP, FT_GPA, HP_MEM, LC_MEM
- Environmental variables: DEG_COMM, SA_PLAN, SA_ACAD_LIFE
- Dependent variable: SEL_PGM (selection of program of study)

Forward stepwise logistic regression using likelihood ratio testing was conducted to determine which independent variables were predictors of selection of program of study (Table 14). The regression results indicated that the overall model fit of the model was weak.

Table 14. Selection of program with significant SIS variables and Mapworks variables

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_COMP</td>
<td>0.146</td>
<td>0.033</td>
<td>19.894</td>
<td>1</td>
<td>0.000</td>
<td>1.157</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.438</td>
<td>0.202</td>
<td>4.691</td>
<td>1</td>
<td>0.030</td>
<td>0.645</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>-0.767</td>
<td>0.338</td>
<td>5.139</td>
<td>1</td>
<td>0.023</td>
<td>0.464</td>
</tr>
<tr>
<td>DEG_COMM</td>
<td>0.335</td>
<td>0.076</td>
<td>19.585</td>
<td>1</td>
<td>0.000</td>
<td>1.398</td>
</tr>
<tr>
<td>Constant</td>
<td>4.095</td>
<td>3.660</td>
<td>1.230</td>
<td>1</td>
<td>0.267</td>
<td>57.944</td>
</tr>
</tbody>
</table>
(-2LL = 996.534) but significantly better in discriminating selection of program ($\chi^2(4, N=827) = 51.156, p < .001$) over the constant model, classifying 61.8% of all cases correctly, 72.3% of those who actually selected a program of study correctly, and 47.7% of those who did not select a program correctly as compared to the constant model which correctly accounted for only 57.4% of those who selected a program and not any of those who did not select a program. The AUROC curve value was 0.635 indicating that on average, the model is interpreted as a weak, although not poor, test of sensitivity against specificity with a 63.5% probability that students were classified correctly. Confirming the goodness-of-fit, the Hosmer-Lemeshow statistic resulted in a Chi-square of $\chi^2(8, N=827) = 5.739, p = 0.676$. When significance exceeds the $p < .05$ threshold of significance, the null hypothesis is rejected (i.e. that there is no difference between observed and predicted values), thereby implying that this model’s estimated fit of the data is at an acceptable level.

As shown in Table 14, the significance of the Wald statistic for first-term credits completed indicates that for every one-credit increase in first-term credits completed, the odds of selecting a program of study increased (i.e., an odds ratio of 1.16 times more likely to select a program of study). The significance of the Wald statistic for age indicates that as age increased, the odds of selection of program of study decreased (i.e. after inverting the odds ratio, 1.55 times more likely not to select a program of study). The significance of the first-term enrollment variable of honors program membership indicates that participation in an honors program decreased the odds of selection of program of study (i.e., after inverting the odds ratio, 2.16 times more likely not to select a program of study). Regarding the Mapworks variables, the significance of the Wald statistic for commitment to completing degree indicates that for each point increase in degree commitment at the institution, the odds
of selection of program increased (i.e. an odds ratio of 1.40 times more likely to select a program of study). The odds ratio for each coefficient indicated a small effect size on the outcome. The Nagelkerke pseudo-$R^2$ of 0.087 reflects there a small effect size, short of the generally accepted .2 level to indicate strength.

d. To what extent do demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental factors predict selection of a program of study among first-year, first-term students majoring in engineering with no selected program of study?

Informed by the previous analysis of correlations and $t$-tests, the independent variables investigated as potential predictors of the selection of program of study were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_MATH
- First-term academic engagement and completion characteristics: FT_CRED_ATT, FT_CRED_COMP, FT_GPA, HP_MEM, LC_MEM
- Environmental factors: ACAD_SE, ACAD_INT, SOC_INT
- Dependent variable: SEL_PGM (selection of program of study)

Forward stepwise logistic regression using likelihood ratio testing was conducted to determine which independent variables were predictors of selection of program of study (Table 15). The regression results indicated that the overall model fit of the model was weak (-2LL = 1025.527) but significantly better in discriminating selection of program ($\chi^2(2, N=827) = 34.016, p < .001$) over the constant model, classifying 58.9% of all cases correctly, 67.1% of those who actually selected a program of study correctly, and 47.9% of those who actually did not select a program correctly as compared to the constant model which

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>$B$</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>$\rho$</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_COMP</td>
<td>0.115</td>
<td>0.031</td>
<td>13.688</td>
<td>1</td>
<td>0.000</td>
<td>1.122</td>
</tr>
<tr>
<td>ACAD_INT</td>
<td>0.273</td>
<td>0.077</td>
<td>12.475</td>
<td>1</td>
<td>0.000</td>
<td>1.314</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.783</td>
<td>0.554</td>
<td>25.195</td>
<td>1</td>
<td>0.000</td>
<td>0.062</td>
</tr>
</tbody>
</table>
correctly accounted for 57.2% of those who selected a program and not any of those who did not. The AUROC curve value was 0.611 indicating that on average, the model is interpreted as a weak, although not poor, test of sensitivity against specificity with a 61.1% probability that students were classified correctly. Confirming the goodness-of-fit, the Hosmer-Lemeshow statistic resulted in a Chi-square of $\chi^2(8, N=827) = 6.480, p = 0.594$. When significance exceeds the $p < .05$ threshold of significance, the null hypothesis is rejected (i.e., that there is no difference between observed and predicted values), thereby implying that this model’s estimated fit of the data is at an acceptable level.

As shown in Table 15, the significance of the Wald statistic for first-term credits completed indicates that for every one credit increase in first-term credits completed, the odds of selection of a program of study increased (i.e. an odds ratio of 1.22 times more likely to select a program of study). Regarding the Mapworks factors, the significance of the Wald statistic for academic integration indicates that as academic integration increased, the odds of selection of program increased (i.e. an odds ratio of 1.314 times more likely to select a program of study). The odds ratio for each coefficient indicated a small effect size on the outcome. The Nagelkerke pseudo-$R^2$ of 0.058 reflects there a small effect size, short of the generally accepted .2 level to indicate strength.

While the logistic regressions have shed useful insights on which variables are influential in predicting the odds of change of major or selection of program of study, this type of analysis does measure the impact of the treatment. In order to answer this question, Research Question 4 was investigated in the following section.
Research Question 4: What is the impact of receiving academic advising informed by predictive analytics on change of major and the selection of program of study?

The main research question was deconstructed into four sub-questions. The purpose for the deconstruction was twofold. First, it was important to validate which, if any, of the Mapworks factors were significant. Second, this was done to determine which variables that comprised the Mapworks factors were significant. Propensity score analyses were then conducted to determine the impact of the treatment.

The variables used to identify the treatment effect using propensity score analyses were selected on the basis of two assumptions. With respect to the first assumption, conditional independence only pre-treatment variables were considered, thereby eliminating end of first-term measurements such as GPA and credits completed. An important aspect of the conditional independence assumption is the emphasis on pre-treatment data thereby utilizing as many observed characteristics as possible to isolate the effect of the treatment on the outcome rather than from an unobserved characteristic. Another important aspect of this assumption relates to bias due to treatment participation based on self-selection or an economic incentive. However, in this study, the academic advising office used the analytics system to track all of the students in the sample, advising all students with academic issues accordingly using the analytics tools. There was not an issue with self-selection or incentive.

Regarding the second assumption of common support, treatment cases must be similar to non-treatment cases in terms of observed characteristics and unaffected by participation. The region of common support should be maintained, preserving the number of comparable control. Guided by these assumptions, the selection of pre-treatment variables was informed by the previous analyses (i.e. correlations, t-tests, and logistic regression) with
close attention given to significant relationships on the outcome variables since the treatment condition could have possibly influenced the outcome.

a. **What is the impact of advising informed by the use of predictive analytics on first-year, first-term students majoring in engineering with no selected program of study, on change of major, using demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables?**

Based on the previous analyses (i.e., correlations, t-tests, and logistic regressions) the unconfounded pre-treatment variables selected to investigate this sub-question were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_CMPST
- First-term academic engagement and completion characteristics: FT_CRED_ATT, HP_MEM, LC_MEM
- Environmental variables: DEG_COMM, SA_MATH, SA_FLW_THRU, SA_PLAN, SE_HARD_CRSE, SA_ACAD_LIFE

In addition, these variables were used to execute the propensity score analysis:

- Binary outcome variable: CHG_MAJ (change of major)
- Binary treatment variable: Recvd_Treatmt

After reviewing the correlation matrix (Appendix B) of the independent variables with respect to the treatment variable (i.e., academic advising informed by the use of predictive analytics) it was determined that the outcome variable (i.e. change of major) did not correlate with the treatment. Pre-treatment variable correlations were then examined to determine which variables to include in the construction of propensity scores. Although gender \((r = .093, p < .01)\) and age \((r = -.839, p < .001)\) correlated to the treatment, these variables could not have been influenced by the treatment and were selected for use. In addition, learning community memberships \((r = .171, p < .01)\), self-assessment of math skills \((r = -.081, p < .05)\), and follows through on what they say \((r = .082, p < .05)\) correlated with the treatment, but were unlikely to have been influenced by the treatment since the decision
to participate in learning communities would have been decided in advance of the treatment as were the responses to self-assessment of math skills and follows through on what they say. Although the variable first-term credits attempted \((r = -.258, p < .01)\) correlated to the treatment (and outcome), it would not have been influenced by the treatment since that credits attempted would have been decided prior to the treatment. The variable commitment to completing degree was selected due to its correlational strength \((r = -.073, p < .05)\) with the outcome, change of major (see Appendix B). The variable self-assessed satisfaction with academic life was also chosen due to a significant correlation with change of major as shown in Table 7 and Table 8.

The \(t\)-test results shown in Table 11 confirmed a significant difference between the groups with regards to age, gender, honors program membership, learning community memberships, follows through on what they say, and first-term credits attempted. Although the groups significantly differed on the variable self-efficacy to do well on problems and tasks, this variable also correlated significantly with self-efficacy to do well in hardest course \((r = .681, p < .01)\). Since self-efficacy to do well in hardest course and ACT composite were noted as a significant predictors of change of major in the logistic regression, the decision was made to use self-efficacy to do well in hardest course in the propensity score analysis along with ACT composite. Finally, high school rank and plans out time were selected because they were pre-treatment variables considered possibly to be informative.

In order to estimate the propensity scores, a probit regression model was constructed using the variables selected (Table 16). The value of McFadden's pseudo-\(R^2\) was 0.7284, indicating a moderately strong goodness-of-fit. Variables in the probit model of significance
Table 16. Probit regression to estimate propensity scores for Change of Major

| Received Treatment      | Coef.   | S.E.    | z     | P>|z|  | [95% Conf. Interval] |
|-------------------------|---------|---------|-------|------|----------------------|
| FT_CRED_ATT            | -0.4210 | 0.0881  | -4.78 | 0.000| -0.5938 -0.2483      |
| HS_RANK                | 0.0095  | 0.0074  | 1.28  | 0.200| -0.0050 0.0241       |
| ACT_CMPST              | 0.0152  | 0.0366  | 0.41  | 0.679| -0.0567 0.0870       |
| GENDER                 | -0.2603 | 0.2770  | -0.94 | 0.347| -0.0832 0.2826       |
| AGE                    | -3.1795 | 0.2400  | -13.25| 0.000| -3.6499 -2.7090      |
| HP_MEM                 | 1.0938  | 0.4836  | 2.26  | 0.024| 0.1460 2.0417        |
| LC_MEM                 | 0.2831  | 0.2193  | 1.29  | 0.197| -0.1467 0.7130       |
| DEG_COMM               | -0.1635 | 0.1039  | -1.57 | 0.115| -0.3671 0.0400       |
| SA_MATH                | -0.2792 | 0.1214  | -2.30 | 0.021| -0.5172 -0.0414      |
| SA_FLW_THRU            | 0.2135  | 0.1298  | 1.65  | 0.100| -0.0408 0.4679       |
| SA_PLAN                | 0.0827  | 0.0782  | 1.06  | 0.290| -0.0705 0.2359       |
| SE_HARD_CRSE           | 0.0057  | 0.1094  | 0.05  | 0.958| -0.2088 0.2202       |
| SA_ACAD_LIFE           | 0.0799  | 0.0972  | 0.82  | 0.411| -0.1107 0.2704       |
| Constant               | 60.6679 | 4.9987  | 12.14 | 0.000| 50.8706 70.4652      |

were first-term credits attempted ($p < .001$), age ($p < .001$), honors program membership ($p < .05$), and self-assessment of math skills ($p < .05$).

Based on the propensity scores produced by the probit regression, the region of common support was in the range [0.01109791, 1] with the optimal number of blocks set to five. A density plot of the propensity scores are in Figure 4. Next, $t$-tests were conducted for each variable in each block using the mean propensity scores to ensure the balancing property was satisfied between the comparison and treatment groups.

Finally, in order to measure the impact of the treatment, four matching methods were executed. The results are provided in Table 17. The number of treatment and control cases was fairly consistent among the matching methods. The Nearest Neighbor method only had 26 control cases used due to using the “matching with replacement” option.

The propensity score analysis using the four matching algorithms reveals a treatment impact (i.e., 7.8%, 7.1%, 7.8%, and 3.4%, respectively) at a high significance (i.e., $t= 3.078,$
Figure 4. Density plot of propensity scores for Change of Major using SIS and Mapworks variables

Table 17. ATT estimation with bootstrapped standard errors (50 replications)

<table>
<thead>
<tr>
<th>Matching Method</th>
<th>n treat.</th>
<th>n contr.</th>
<th>ATT</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>116</td>
<td>26</td>
<td>0.078</td>
<td>-0.0093</td>
<td>0.025</td>
<td>3.078</td>
</tr>
<tr>
<td>Stratification</td>
<td>114</td>
<td>251</td>
<td>0.071</td>
<td>-0.0059</td>
<td>0.023</td>
<td>3.040</td>
</tr>
<tr>
<td>Kernel</td>
<td>116</td>
<td>249</td>
<td>0.078</td>
<td>-0.0012</td>
<td>0.023</td>
<td>3.384</td>
</tr>
<tr>
<td>Radius</td>
<td>115</td>
<td>249</td>
<td>0.034</td>
<td>-0.0010</td>
<td>0.030</td>
<td>1.118</td>
</tr>
</tbody>
</table>

3.040, 3.384, 1.118, respectively) as a result of academic advising informed by the use of predictive analytics on change of major.

b. What is the impact of advising informed by the use of predictive analytics on first-year, first-term students majoring in engineering with no selected program of study, on change of major, using demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental factors?
Based on the previous analyses (i.e., correlations, $t$-tests, and logistic regressions) the unconfounded pre-treatment variables selected to investigate this sub-question were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK, ACT_CMPST
- First-term academic engagement and completion characteristics: FT_CRED_ATT, HP_MEM, LC_MEM
- Environmental factors: ACAD_SE, ACAD_INT, SOC_INT

In addition, these variables were used to execute the propensity score analysis:

- Binary outcome variable: CHG_MAJ (change of major)
- Binary treatment variable: Recvd_Treatmt

The only difference between this propensity score analysis and the previous sub-question was the use of the Mapworks factors rather than the Mapworks variables. Thus, the rationale for the selection of the demographic, pre-college academic characteristics, and first-term academic engagement and completion characteristics is not repeated. In regards to the factors, a weak positive correlation ($r = .072, p < .05$) between the factor academic integration and the treatment variable was noted. The $t$-tests results in Table 9 confirmed a significant difference between the groups with regards to age, gender, honors program membership, learning community memberships, academic self-efficacy, and academic integration. In addition, academic self-efficacy was a significant predictor of change of major in the logistic regression.

A probit regression model was constructed using the variables and factors selected to estimate the propensity scores (Table 18). The value of McFadden's pseudo-$R^2$ was 0.7085, indicating a moderately strong goodness-of-fit. Variables in the probit model of significance were first-term credits attempted ($p < .001$), age ($p < .001$), and honors program membership ($p < .05$). Learning community memberships was significant at the $p < .07$ level.
Table 18. Probit regression to estimate propensity scores for Change of Major

| Received Treatment | Coef.  | S.E.  | z      | P>|z| | [95% Conf. Interval] |
|--------------------|--------|-------|--------|------|----------------------|
| FT_CRED_ATT        | -0.3904| 0.0827| -4.72  | 0.000| -0.5525   -0.2283   |
| HS_RANK            | 0.0069 | 0.0068| 1.01   | 0.313| -0.0065   0.0202    |
| ACT_CMPST          | 0.0043 | 0.0342| 0.13   | 0.900| -0.0628   0.0714    |
| GENDER             | -0.2633| 0.2610| -1.01  | 0.313| -0.7748   0.2483    |
| AGE                | -3.0748| 0.2218| -13.86 | 0.000| -3.5096   -2.6400   |
| HP_MEM             | 0.9630 | 0.4484| 2.15   | 0.032| 0.0841    1.8420    |
| LC_MEM             | 0.3750 | 0.2052| 1.83   | 0.068| -0.0272   0.7771    |
| ACAD_SE            | 0.0409 | 0.1312| 0.31   | 0.755| -0.2163   0.2981    |
| ACAD_INT           | -0.0815| 0.1357| -0.60  | 0.548| -0.3474   0.1843    |
| SOC_INT            | 0.1138 | 0.0940| 1.21   | 0.226| -0.0705   0.2981    |
| Constant           | 57.9001| 4.5335| 12.77  | 0.000| 49.0146   66.7855   |

Based on the propensity scores produced by the probit regression, the region of common support was in the range [0.01281635, 0.99999998] with the number of blocks set to five. A density plot of the propensity scores are in Figure 5. Next, \( t \)-tests were conducted for each variable in each block using the mean propensity scores to ensure the balancing property was satisfied between the comparison and treatment groups.

![Density plot of propensity scores for Change of Major using SIS variables and Mapworks factors](image)
Finally, in order to measure the effect of the treatment, four matching methods were executed. The results are provided in Table 19. The number of treatment and control cases was fairly consistent among the matching methods. The Nearest Neighbor method only had 26 control cases used due to using the “matching with replacement” option.

Table 19. ATT estimation with bootstrapped standard errors (50 replications)

<table>
<thead>
<tr>
<th>Matching Method</th>
<th>n treat.</th>
<th>n contr.</th>
<th>ATT</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>117</td>
<td>26</td>
<td>0.068</td>
<td>-0.0056</td>
<td>0.028</td>
<td>2.441</td>
</tr>
<tr>
<td>Stratification</td>
<td>117</td>
<td>289</td>
<td>0.079</td>
<td>-0.0001</td>
<td>0.024</td>
<td>3.275</td>
</tr>
<tr>
<td>Kernel</td>
<td>117</td>
<td>289</td>
<td>0.078</td>
<td>-0.0016</td>
<td>0.026</td>
<td>3.025</td>
</tr>
<tr>
<td>Radius</td>
<td>117</td>
<td>285</td>
<td>0.040</td>
<td>-0.0023</td>
<td>0.031</td>
<td>1.303</td>
</tr>
</tbody>
</table>

The propensity score analysis using the four matching algorithms reveals a treatment impact (6.8%, 7.9%, 7.8%, and 4.0%, respectively) at a high significance ($t= 2.441, 3.275, 3.025, 1.303$, respectively) as a result of academic advising informed by the use of predictive analytics on change of major.

c. What is the impact of advising informed by the use of predictive analytics on first-year, first-term students majoring in engineering with no selected program of study, on selection of program, using demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables?

Based on the previous analyses (i.e., correlations, $t$-tests, and logistic regressions) the unconfounded pre-treatment variables selected to investigate this sub-question were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK
- First-term academic engagement and completion characteristics: FT_CRED_ATT, HP_MEM, LC_MEM
- Environmental variables: DEG_COMM, SA_MATH, SA_FLW_THRU, SA_PLAN, SE_PROBS_TASKS, SA_ACAD_LIFE

In addition, these variables were used to execute the propensity score analysis:
- Binary outcome variable: SEL_PGM (selection of program of study)
- Binary treatment variable: Recvd_Treatmt

After reviewing the correlation matrix in Appendix B of the independent variables with respect to the treatment variable (i.e., academic advising informed by the use of predictive analytics) it was determined that the outcome variable (i.e. selection of program of study) weakly correlated with the treatment ($r = .086, p < .05$). This is consistent with the $t$-test which also indicates a significant group difference on SEL_PGM regarding the treatment on the entire sample rather (i.e., ATE) than on the basis of propensity score analysis (i.e., ATT).

Pre-treatment variable correlations were then examined to determine which variables to include in the construction of propensity scores. Although gender ($r = .093, p < .01$) and age ($r = -.839, p < .001$) correlated to the treatment, these variables could not have been influenced by the treatment and were selected for use. In addition, learning community memberships ($r = .171, p < .01$), self-assessment of math skills ($r = -.081, p < .05$), and follows through on what they say ($r = .082, p < .05$) correlated with the treatment, but were unlikely to have been influenced by the treatment since the decision to participate in learning communities would have been decided in advance of the treatment and the responses to the Mapworks survey (i.e., related to self-assessment of math skills and follows through on what they say) were gathered prior to the treatment. First-term credits attempted ($r = .258, p < .01$) also correlated to the treatment (and outcome), but it could not have been influenced by the treatment since that number was decided on the tenth day of class of the semester. The variables commitment to completing degree ($r = .157, p < .01$) and plans out time ($r = .126, p < .01$) were selected due to their correlations with the outcome variable (see Appendix B).
The variable self-assessed satisfaction with academic life was chosen due to its significant correlation with the dependent variable, change of major, as shown in Table 7 and Table 8.

The t-tests results shown in Table 11 confirmed a significant difference between the groups regarding age, gender, honors program membership, learning community memberships, follows through on what they say, self-efficacy to do well on problems and tasks, and first-term credits attempted. In addition, commitment to completing degree, and honors program membership were noted as a significant predictors of selection of program in the logistic regression reinforcing the reason to use these variables in the propensity score analysis. Finally, high school rank and plans out time were selected based on t-tests results (Table 11) that while not significant ($p = .142$ and $p = .069$, respectively) were informative.

In order to estimate the propensity scores, a probit regression model was constructed using the variables selected (Table 20). The value of McFadden's pseudo-$R^2$ was 0.7237, indicating a moderately strong goodness-of-fit. Variables in the probit model of significance included first-term credits attempted ($p < .001$), age ($p < .001$), honors program membership ($p < .05$), and self-assessment of math skills ($p < .05$).

Table 20. Probit regression to estimate propensity scores for Selection of Program

| Received Treatment     | Coef.  | S.E.  | z      | P>|z|  | [95% Conf. Interval] |
|------------------------|--------|-------|--------|------|---------------------|
| FT_CRED_ATT            | -0.4314| 0.0872| -4.95  | 0.000| -0.6023 -0.2604     |
| HS_RANK                | 0.0123 | 0.0069| 1.77   | 0.076| -0.0013 0.0260      |
| GENDER                 | -0.2886| 0.2729|-1.06   | 0.290| -0.8236 0.2463      |
| AGE                    | -3.1452| 0.2357|-13.35  | 0.000| -3.6070 -2.6833     |
| HP_MEM                 | 1.1180 | 0.4512| 2.48   | 0.013| 0.2336 2.0023       |
| LC_MEM                 | 0.3157 | 0.2154| 1.47   | 0.143| -0.1065 0.7380      |
| DEG_COMM               | -0.1872| 0.0977|-1.91   | 0.056| -0.3787 0.0044      |
| SA_MATH                | -0.2679| 0.1184|-2.26   | 0.024| -0.4999 -0.0359     |
| SA_Flw_THRU            | 0.1798 | 0.1265| 1.42   | 0.155| -0.0681 0.4278      |
| SA_PLAN                | 0.0927 | 0.0752| 1.23   | 0.218| -0.0546 0.2400      |
| SE_PROBS_TASKS         | 0.0216 | 0.1206| 0.18   | 0.858| -0.2147 0.2579      |
| SA_ACAD_LIFE           | 0.1023 | 0.0912| 1.12   | 0.262| -0.0764 0.2810      |
| Constant               | 60.4181| 4.8807| 12.38  | 0.000| 50.8521 69.9841     |
Based on the propensity scores produced by the probit regression, the region of common support was in the range \([0.01102057, 1]\) with the optimal number of blocks set to five. A density plot of the propensity scores are in Figure 6. Next, \(t\)-tests were conducted for each variable in each block using the mean propensity scores to ensure the balancing property was satisfied between the comparison and treatment groups.

![Density plot of propensity scores for Selection of Program using SIS and Mapworks variables](image)

**Figure 6.** Density plot of propensity scores for Selection of Program using SIS and Mapworks variables

Finally, in order to measure the effect of the treatment, four matching methods were executed. The results are provided in Table 21. The number of treatment and control cases was fairly consistent among the matching methods. The Nearest Neighbor method only had 29 control cases used due to using the “matching with replacement” option.
Table 21. ATT estimation with bootstrapped standard errors (50 replications)

<table>
<thead>
<tr>
<th>Matching Method</th>
<th>n treat.</th>
<th>n contr.</th>
<th>ATT</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>120</td>
<td>29</td>
<td>-0.071</td>
<td>0.0946</td>
<td>0.249</td>
<td>-0.284</td>
</tr>
<tr>
<td>Stratification</td>
<td>118</td>
<td>253</td>
<td>0.091</td>
<td>0.0010</td>
<td>0.165</td>
<td>0.554</td>
</tr>
<tr>
<td>Kernel</td>
<td>120</td>
<td>251</td>
<td>0.078</td>
<td>0.0040</td>
<td>0.169</td>
<td>0.461</td>
</tr>
<tr>
<td>Radius</td>
<td>120</td>
<td>251</td>
<td>0.069</td>
<td>0.0016</td>
<td>0.066</td>
<td>1.037</td>
</tr>
</tbody>
</table>

The propensity score analysis using the four matching algorithms does not reveal a treatment impact (i.e., -7.1%, 9.1%, 7.8%, and 6.9%, respectively) at a significant level (i.e. \(t\) = -0.284, 0.554, 0.461, 1.037, respectively).

d. **What is the impact of advising informed by the use of predictive analytics on first-year, first-term students majoring in engineering with no selected program of study, on selection of program, using demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental factors?**

Based on the previous analyses (i.e. correlations, \(t\)-tests, and logistic regressions) the unconfounded pre-treatment variables selected to investigate this sub-question were:

- Demographic: GENDER, AGE
- Pre-college academic characteristics: HS_RANK
- First-term academic engagement and completion characteristics: FT_CRED_ATT, HP_MEM, LC_MEM
- Environmental factors: ACAD_SE, ACAD_INT, SOC_INT

In addition, these variables were used to execute the propensity score analysis:

- Binary outcome variable: SEL_PGM (selection of program of study)
- Binary treatment variable: Recvd_Treatmt

The only difference between this propensity score analysis and the previous sub-question was the use of the Mapworks factors rather than the Mapworks variables. Thus, the rationale for the selection of the demographic, pre-college academic characteristics, and first-term academic engagement and completion characteristics is not repeated. Regarding the
factors, a weak positive correlation \( r = .072, p < .05 \) between academic integration and the treatment variable was noted. In addition, a positive correlation was observed between the variables academic integration \( r = .161, p < .01 \) and social integration \( r = .116, p < .01 \) and the outcome variable.

The \( t \)-tests results shown in Table 9 confirmed a significant difference between the groups with regards to age, gender, honors program membership, and learning community memberships. With respect to the Mapworks factors, there was a significant difference at the \( p < .05 \) level on the factor academic self-efficacy between the comparison \( (M=5.1366, SD=0.9542) \) and treatment \( (M=5.3278, SD=0.9574) \) groups; \( t(796)=-2.029, p = .043 \). There was also a significant difference at the \( p < .05 \) level on the factor academic integration between the comparison \( (M=5.5235, SD=0.9941) \) and treatment \( (M=5.7293, SD=0.9086) \) groups; \( t(796)=-2.124, p = .034 \). In addition, academic integration was a significant predictor of change of major in the logistic regression.

In order to estimate the propensity scores, a probit regression model was constructed using the variables and factors selected. The value of McFadden's pseudo-\( R^2 \) was 0.7030, indicating a moderately strong goodness-of-fit. Variables in the probit model of significance included first-term credits attempted \( (p < .001) \), age \( (p < .001) \), and honors program membership \( (p < .05) \).

Based on the propensity scores produced by the probit regression, the region of common support was in the range \([0.01373774, 0.99999997]\) with the optimal number of blocks set to five. A density plot of the propensity scores are in Figure 7. Next, \( t \)-tests were conducted for each variable in each block using the mean propensity scores to ensure the balancing property was satisfied between the comparison and treatment groups.
Table 22. Probit regression to estimate propensity scores for Selection of Program

| Received Treatment | Coef.   | S.E.   | z      | P>|z|   | [95% Conf. Interval] |
|--------------------|---------|--------|--------|--------|----------------------|
| FT_CRED_ATT        | -0.3996 | 0.0817 | -4.89  | 0.000  | -0.5598 -0.2394      |
| HS_RANK            | 0.0087  | 0.0065 | 1.35   | 0.177  | -0.0040 0.0214       |
| GENDER             | -0.2784 | 0.2564 | -1.09  | 0.278  | -0.7810 0.2241       |
| AGE                | -3.0134 | 0.2134 | -14.12 | 0.000  | -3.4317 -2.5950      |
| HP_MEM             | 0.9660  | 0.4175 | 2.31   | 0.021  | 0.1478 1.7843        |
| LC_MEM             | 0.3921  | 0.2027 | 1.93   | 0.053  | -0.0052 0.7895       |
| ACAD_SE            | 0.0551  | 0.1271 | 0.43   | 0.665  | -0.1941 0.3043       |
| ACAD_INT           | -0.0379 | 0.1326 | -0.29  | 0.775  | -0.2979 0.2220       |
| SOC_INT            | 0.0961  | 0.0917 | 1.05   | 0.294  | -0.0836 0.2758       |
| Constant           | 56.7090 | 4.3070 | 13.17  | 0.000  | 48.2675 65.1505      |

Figure 7. Density plot of propensity scores for Selection of Program using SIS variables and Mapworks factors
Finally, in order to measure the effect of the treatment, four matching methods were executed. The results are provided in Table 23. The number of treatment and control cases was fairly consistent among the matching methods. The Nearest Neighbor method only had 27 control cases used due to using the “matching with replacement” option.

The propensity score analysis using the four matching algorithms does not reveal a consistent impact (i.e. -4.1%, 7.2%, 5.6%, and 10.7%, respectively) at a significant level (i.e. $t = -0.232, 0.428, 0.365, 1.801$, respectively) of the treatment on the outcome, although radius matching did begin to approach significance at the 1.96 level.

Table 23. ATT estimation with bootstrapped standard errors (50 replications)

<table>
<thead>
<tr>
<th>Matching Method</th>
<th>n treat.</th>
<th>n contr.</th>
<th>ATT</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>121</td>
<td>27</td>
<td>-0.041</td>
<td>0.0545</td>
<td>0.178</td>
<td>-0.232</td>
</tr>
<tr>
<td>Stratification</td>
<td>121</td>
<td>288</td>
<td>0.072</td>
<td>0.0086</td>
<td>0.168</td>
<td>0.428</td>
</tr>
<tr>
<td>Kernel</td>
<td>121</td>
<td>288</td>
<td>0.056</td>
<td>0.0228</td>
<td>0.154</td>
<td>0.365</td>
</tr>
<tr>
<td>Radius</td>
<td>121</td>
<td>284</td>
<td>0.107</td>
<td>-0.0097</td>
<td>0.059</td>
<td>1.801</td>
</tr>
</tbody>
</table>

**Summary**

The results revealed relationships between the independent variables and factors and change of major and the selection of a program study. Correlations that were significant in all cases (combined and by group) are especially noteworthy. For example, a variable from the first-term academic engagement and completion characteristics covariate that significantly correlated positively with the dependent variable selection of program of study in all cases was first-term credits completed. In addition, it significantly correlated negatively with change of major in the combined and comparison group analyses. Another variable from first-term academic engagement and completion characteristics covariate which correlated positively with selection in all cases was first-term GPA. In addition, it
significantly correlated negatively with change of major in the combined and comparison group analyses. Taken together, these correlations suggest that completing more credits and achieving a higher first-term GPA play influence the decision to select a program of study.

The Mapworks variables that significantly correlated negatively with change of major across all cases were self-assessment of math skills and self-efficacy to do well in hardest course, indicating the higher students self-assessed themselves on these two scales, the less likely they were to change majors. In addition, plans out time correlated positively with the selection of program of study, indicating if a student was a “planner”, as expected, they were more likely to select a program. The variable self-assessed satisfaction with academic life was noteworthy as it significantly correlated negatively with changed of major and positively with the selection of program of study in all cases. The Mapworks factors that were significant across all cases were academic self-efficacy (significantly negatively with change of major) and academic integration (significantly negatively with change of major and positively with selection of program of study). These last three indicate that, when students view their academic integration in a more efficacious, integrated manner, likelihood to change of major declines and the likelihood to selection of a program of study increases.

The results of the t-tests were notable as well. In terms of demographics, significant differences between the comparison and treatment groups were revealed on gender and age, with slightly more females and younger students in the treatment group. Examining the first-term academic engagement and completion covariate, it was also revealed the treatment group had more members in honors programs and learning communities as well as attempting fewer first-term credits. The treatment group was also significantly higher on the Mapworks variables of follows through on what they say and self-efficacy to do well on
problems and tasks. In addition, the treatment group was significantly higher on the Mapworks factors of academic self-efficacy and academic integration. Finally, the treatment group had a significantly higher mean on the selection of a program of study, but not on change of major.

The results of the logistic regressions revealed variables and factors that were predictors of the change of major and the selection of program of study. In regards to change of major, the variables first-term credits attempted, first-term credits completed, and ACT composite were significant predictors. An increased number of first-term credits attempted, the greater the odds of changing majors. In addition, as the number of first-term credits completed increased, the odds of changing majors were less likely. Finally, a higher overall ACT composite score also decreased the odds of changing majors.

Regarding Mapworks, the variable self-efficacy to do well in hardest course and the factor academic self-efficacy were indicators of change of major. As these two increased on their unit scales, the odds of changing majors decreased, indicating the importance of self-efficacy in doing well in a hard course and academically. Regarding the selection of program of study, the variable first-term credits completed was a significant predictor. As the number of completed credits increased, the odds of selecting a program of study increased. An interesting finding was that as age and honors program membership increased, this resulted in a decrease in the odds of selecting a program of study. Regarding the Mapworks variables, it was noted that as the variable commitment to completing degree increased, the odds of the selection a program of study increased. Finally, as the Mapworks factor academic integration increased, the odds of selecting a program of study increased.
Propensity score analyses were conducted using four matching methods (nearest neighbor, stratification, kernel, and radius) to measure the impact of the treatment. Regarding change of major, three of the four matching methods revealed a significant treatment effect of 7% to 8% on the average of the treatment on the treated (ATT). Regarding selection of a program of study, a significant treatment effect was not found; however, the radius matching method that utilized the Mapworks environmental factors was nearly significant. These findings contrast with the t-tests, wherein the entire sample was utilized and a difference on selection of program of study was found, but not on change of major. Propensity score analyses attempts to match similar treatment and comparison cases on the basis of the selected influential pre-treatment variables. Hence, the true impact of the treatment is revealed since subjects are matched on more similar characteristics. It was observed in the probit regressions in the propensity score analyses that learning community memberships was nearly significant in two of the analyses and commitment to completing degree was nearly significant in one. The effect of learning community memberships may be viewed as an academic integration experience and contributing to identity as an engineer. The near significance of commitment to completing degree is a reflection of the high level of self-efficacy these students possess early in the term. The implications of these findings are discussed in the following chapter.
CHAPTER 5. DISCUSSION

The purpose of this study was to investigate and describe the relationships between demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics, and environmental variables on student outcomes (i.e., change of major and the selection of program of study) of first-year, first-term students in engineering who are transitioning into their postsecondary academic careers. The central research questions that guided this study were:

a. What are the demographic characteristics of the subjects in this study?

b. What are the correlational relationships between the independent variables with the dependent variables and what are the mean differences between the comparison and treatment groups?

c. To what extent did the independent variables and factors predict change of major and the selection of a program of study?

d. What is the impact of receiving academic advising informed by predictive analytics on change of major and the selection of a program of study?

Data from the student information system (SIS) at the institution were used to collect demographic characteristics, pre-college academic characteristics, and first-term academic engagement and completion characteristics. Data from the institution’s Mapworks® transition survey instrument were used to obtain environmental variables and factors. This chapter provides a discussion of the findings, implications for policy and practice, and recommendations for future research.
Summary of the Findings

This study investigated how student academic characteristics and environmental variables at an institution predict and impact student outcomes (i.e. change of major and the selection of program of study). Select variables and factors from the data sources were investigated using Astin’s (1993) Input-Environment-Output (I-E-O) model as a conceptual framework for student integration, and Tinto and Pusser’s (2006) model of institutional action for student success. Logistic regression models were produced that explained to what extent the variables and factors contributed to change of major and the selection of program of study. This study also examined the impact of adding predictive analytics to the academic advising environment on the aforementioned student outcomes.

Change of major

Two logistic regression models were created to investigate the odds of change of major (i.e., probability). The first model utilized variables derived from the student information system (SIS) including demographic characteristics, pre-college academic characteristics, first-term academic engagement and completion characteristics. Environmental variables were derived from questions in the Mapworks® transition survey instrument. The second model utilized the same SIS variables in conjunction with Mapworks factors. The reason for the two models was not only to test the influence of the factors derived from the Mapworks survey that correspond to Tinto’s integration theory, but also to deconstruct the factors and determine if the specific variables which compose those factors were more significant as indicators of outcomes. The two models did not differ greatly in their capacity to predict the outcome (i.e., a Nagelkerke pseudo-$R^2$ of 0.167 based on Mapworks variables versus a pseudo-$R^2$ of 0.164 based on Mapworks factors). Compared
similarly, a successful case classification rate of 72.1% versus 72.6%, and an AUROC curve value of 0.760 versus 0.769. The logistic regression model that utilized the Mapworks variables did, however, reveal which of the specific variables most significantly contributed to the odds of changing major. Those variables were first-term credits attempted (positive), first-term credits completed (negative), ACT composite score (negative), and self-efficacy in hardest course (negative). The effect size of these variables tended to be small with the largest contributing variables being first-term credits attempted (i.e., an odds ratio of 1.245, indicating the more credits that were taken, the odds of changing majors increased) and self-efficacy in hardest course (i.e., an inverted odds ratio of 1.52, indicating that as a student’s confidence in their ability to do well in their hardest course increased, the odds of not changing majors increased). The second logistic regression model revealed that the same SIS variables and a single Mapworks factor predicted change of major. The factor of significance was academic self-efficacy in the second model (i.e. an inverted odds ratio of 1.66, indicating that as self-efficacy increased, the odds of not changing majors increased).

**Selection of a program of study**

Two logistic regression models were created to investigate the odds (i.e., probability) of selecting a program of study with the first model utilizing SIS variables in conjunction with environmental variables derived from questions in the Mapworks® transition survey instrument. The second model utilized the same SIS variables in conjunction with Mapworks factors. It was determined that the two models did not differ greatly in their capacity to predict the outcome (i.e., a Nagelkerke pseudo-$R^2$ of 0.087 based on Mapworks variables versus a pseudo-$R^2$ of 0.058 based on Mapworks factors). Compared similarly, a successful case classification rate of 61.8% versus 58.9%, and an AUROC curve value of 0.635 versus
The logistic regression model that utilized SIS and Mapworks variables did, however, reveal the variables that most significantly contributed to the odds of the selection of program of study were: first-term credits completed (positive), age (negative), honors program membership (negative), and commitment to completing a degree (positive). Although the effect size of these variables was small, the largest contributing variables were commitment to completing degree (1.398, indicating the greater the commitment to the degree at the institution, the odds of selecting a program of study increased) and honors program membership (i.e. an inverted odds ratio of 2.155, indicating that as membership in honors programs increased, the odds of not selecting a program of study decreased). The second logistic regression model revealed that the SIS variable first-term credits completed (positive) and the Mapworks factor academic integration (positive) were significant in predicting the odds of selecting a program of study (i.e. as they each increased, the odds of selecting a program of study increased).

**Connections with Contemporary Research**

Tinto and Pusser’s (2006) institutional action for student success framework, namely, the feedback-support-involvement triad can have a profound influence on self-efficacy, hardiness, and academic integration experiences which, in turn, lead to student outcomes. The next section discusses how the study findings relate to the triad of student success as engineering students integrate into their new academic environment.

**Self-efficacy**

As postsecondary students begin their journey in a field of study such as engineering, the concept of self-efficacy has been considered influential in relation to student’s ability to adapt and thrive (Lent, Brown, & Larkin, 1984). According to Bandura (1993), self-efficacy
is based on multiple self-regulatory processes (i.e., cognitive, motivation, affective, and selection). Similarly, the extent to which one anticipates to perform well academically has been shown to be a contributing factor in the decision-making processes of students in regards to change of major (Allen & Robbins, 2008; Eccles, 1987; Stinebrickner & Strinebrickner, 2011). Eliot and Turns (2011) posited that an individual’s ability to reconcile messages and feedback (i.e., sense-making) about one’s skills and knowledge can shape a student’s academic self-concept. In their study of engineering students, the authors contended that, as self-concept improves, self-identity as an engineer is strengthened. The implication is that a positive academic self-concept promotes academic growth. Academic self-concept is shaped by external and internal messages and reinforcing activities. Certain variables from this study have a role within the context of developing academic self-concept.

Variables such as self-assessed math skills, self-efficacy in hardest course, and self-efficacy to complete difficult problems and tasks reflect internal perceptions formed through feedback and involvement in the academic setting and may be considered indicators of a more resilient academic self-concept as an engineer. These same variables also significantly correlated (negatively) with change of major, indicating that, as self-efficacy in increased, students were less likely to abscond from an engineering major. Similarly, when reviewing the logistic regression models, the variable self-efficacy in hardest course and the factor academic self-efficacy were significant predictors of change of major in the logistic regression (i.e., as these predictors increased in unit, the odds of changing majors decreased). In addition, a noteworthy observation was the identification of ACT composite score as a significant negative predictor of change of major (i.e., as this variable increased, the odds of changing majors decreased). A possible explanation for this might be that the ACT
composite score is an early form of external feedback received by students, influencing their academic self-concept and affecting their perceptions of ability to academically succeed. Thus, student willingness to change majors is likely decreased by a belief in their academic ability based on ACT composite score. While the ACT score may function as a measure for admission purposes, it may serve to hinder students from considering the magnitude of subsequent feedback and messages from other sources during their first college semester in relation to academic performance.

In the context of this study, the decision to select a program of study may be viewed as an embracement of the academic environment in engineering, reifying self-identity as an engineer. The selection of a program of study indicates that efforts made in the feedback-support-involvement triad have amalgamated to the extent that the student has self-authored a positive view of their academic self-concept. The formation of academic self-concept was occurring through externally and internally framed activities, shown to provide the necessary scaffolding for students to define “self” as an engineer (Eliot & Turner, 2011). The selection of program of study is a demonstration of their belief in their values and skills as an engineer. In this study, variables that correlated positively with the selection of program of study were: first-term credits completed, first-term grade point average, commitment to completing a degree, plans out time, follows through on what they say, degree of satisfaction with academic life, institutional satisfaction, academic integration and, to a lesser extent, social integration. Based on the logistic regression models, the significant positive predictors of selection of program of study were first-term credits completed, commitment to completing a degree, and academic integration, thereby reinforcing the importance of these variables and factors. The self-assessed variables of follows through on what they say and plans out time
demonstrate an internal awareness by students of necessary values needed to success as an engineering student.

Reexamining these relationships in terms of academic self-concept and thereby, self-efficacy, external feedback akin to Tinto and Pusser’s (2006) framework was received through first-term credits completed and first-term GPA. In turn, this provided students with an internal cue in regards to potential academic performance and an understanding of what is required to succeed academically. The variables of academic life satisfaction and institutional satisfaction are self-reported affective variables based on external feedback received through involvement in the classroom and with faculty and peers. While still in the early stages of formation during a student’s first-term, these variables are influential in shaping the selection of program of study.

Finally, the factor of academic self-efficacy significantly correlated (negatively) with change of major and conversely with the selection of program (positively) in the treatment group. It was a significant predictor in decreasing the odds of changing majors, reinforcing the research on the formation of academic self-concept. In terms of Tinto and Pusser’s (2006) framework, the support provided by the institution (faculty, advisors, and peers) can play a vital role in the development of academic self-concept, ultimately leading to a higher level self-efficacy and perseverance within a chosen major.

**Hardiness**

The concept of hardiness is based on the proposition of the human tendency to thrive during periods of stress. In this study, the first-year, first-term transition experience may be viewed as a stressful period. Researchers have asserted that one’s level of hardiness positively correlates to the persistence of students in higher education (Lifton & Flanagan,
1995; Lifton, et al., 2004; Lifton, et al., 2006) and academic success (Sheard, 2009; Sheard & Golby; 2007). Maddi (2004) operationalized hardiness into three distinct components: commitment, control, and challenge. Commitment is regarded as one’s level of engagement with ongoing activities (i.e. people and events). In the context of this study, commitment was reflected in academic integration and social integration. When examining the correlation results (see Table 9) for the entire sample, the factor social integration significantly correlated positively with the selection of program of study. The factor academic integration correlated positively with the selection of program of study and negatively with change of major and was also a significant predictor in the logistic regression for the selection of program of study, thus, demonstrating commitment to further specialization in their discipline. While social integration was generally not significant in most of the analyses conducted, other variables that may contribute to a sense of social integration did display some significance. The honors program membership variable was consistently a significant predictor in the probit regression models constructed during the propensity score analyses for the selection of program of study. The learning community memberships variable missed significance by a small amount ($p = 0.053$) in the probit regression models for the selection of program of study. Finally, self-assessed satisfaction with academic life significantly correlated with selection of program (positively) and change of major (negatively). When viewed as a form of engagement in ongoing activities (i.e., the definition of commitment in the hardiness context), these variables tend to support the relevance of the commitment construct, tangential to the “involvement” support described in the Tinto and Pusser (2006) framework.
Similar to the self-efficacy processes of affect and selection, the hardiness construct of control was described by Maddi as the sense that one has authority and influence over activities and outcomes. The variables in this study which reflect the control construct are plans out time and follows through on what they say. The variable plans out time had a significant positive correlation with the selection of a program of study. Also, the variable of follows through on what they say exhibited a significant positive correlation with the selection of program of study overall and a significant negative correlation with change of major for the treatment group. These variables, in particular, reflect the extent of control exerted by a student. It also noteworthy to mention that the factor academic self-efficacy, a significant predictor of change of major in the logistic regression, was composed of questions that reflect a perceived control over outcomes such as ability to do well on challenging problems and tasks, ability to do well in hardest course, and ability to persevere on challenging projects. These variables also suggest a form of engagement or involvement as described in the Tinto and Pusser (2006) framework, albeit, initiated by the student.

The final hardiness construct described by Maddi (2004) is challenge. Challenge is described as the fulfillment attained in the process of continuing to learn from experiences. In this study, variables such as self-efficacy to complete difficult problems and tasks and self-efficacy of ability to do well in hardest course reflect the construct of challenge. The variable of self-efficacy to complete difficult problems and tasks significantly correlated with change of major (negatively). Conversely, self-efficacy to complete difficult problems and tasks correlated positively with selection of program in the treatment group. The variable of self-efficacy of ability to do well in hardest course significantly correlated (negatively) with change of major and was a significant predictor (negatively) of change of major, reflecting
that as confidence in academic ability to do well in their hardest course increased, the odds of changing majors decreased. These results suggest that, as engineering students perceived their hardest course as a challenge through which they could rise to and learn, they were less likely to change majors, implying that these variables can serve as indicators as to whether a student will choose to change their major or select a program of study. As denoted in the Tinto and Pusser (2006) framework, the “feedback” they receive in regards to academic challenging courses and tasks likely shape their academic self-concept in regards to their perceived ability to succeed.

**Student integration**

Tinto (1976) asserted that as academic integration (i.e., acclimating to courses and GPA performance) and social integration (i.e., interactions with faculty, club memberships, and study groups) increases, the likelihood of “disciplinary anomie” (i.e., changing majors or not selecting a program of study due to anxiety or alienation) should decrease. In support of Tinto’s assertion, recent research has affirmed that academic integration is associated with change of major. In a study of STEM majors, Xu (2016) found students were less likely to change majors if they felt positively about factors related to their academic program including teaching quality, faculty support, and academic advising while social engagement with peers and in campus activities was not significant in their persistence. Additionally, the models that Xu constructed revealed that demographic background characteristics did not have a significant relationship with changing majors. Meanwhile, Ferrare and Lee (2014) found that college GPA (i.e. a form of academic integration) and frequently engaging in study groups (i.e., a form of social integration in their study) significantly predicted persistence in STEM majors. They also found that a lower proportion of STEM credits in the
first year were predictive in switching out of STEM majors, indicating a lack of academic integration. These relationships were consistent in their blocked and full models, supporting the notion that a greater degree of academic and social integration reduces the likelihood of changing out of STEM disciplines.

While the role of academic integration has been well-established in the research and is supported by the findings in this study, it is apparent the influence of social integration is less clear. Social integration may vary and is likely more contextual. In STEM majors, this may include interactions with faculty or inclusion in study groups. In this study, social integration was a Mapworks factor derived from questions measuring sense of belonging, degree of fitting in, and satisfaction with social life that did not demonstrate a great deal of relation or influence to the outcomes. As mentioned, it is possible that like the Ferrare and Lee (2014) research, the variables of honors program membership and learning community memberships in this study more distinctly expose the essence of social integration for engineering students. In the probit models constructed by the propensity score analyses, it was noted that the variables honors program membership was significant and learning community memberships was nearly significant. In addition, honors program membership was a significant indicator in the logistic regression model that predicted the odds of the selection of a program of study; however, in a negative direction implying that honors program membership did not increase the odds of selecting a program of study. A possible explanation for this outcome is that while honors program membership may not be directly contributing to self-identity as an engineer, it may perhaps be reinforcing the self-efficacy process of selection whereby students are still contemplating the selection of their future
academic paths. Future researchers may wish to consider the process of selection as it relates to social integration in a particular discipline.

In summary, the Mapworks factor academic integration was revealed as having a significant influence in change of major and the selection of program of study for the engineering students. This factor also correlated significantly with selection of program (positively) and change of major (negatively) and was a significant predictor of the selection of a program of study (i.e., increasing the odds of selecting a major in a logistic regression model). While the factor social integration significantly correlated (positively) with selection of program in the overall sample, it was not a significant predictor in any of the logistic regression models. This may have occurred due to the way the questions that composed this factor were phrased or perceived by the engineering students in this study.

The influence of the factors academic integration and academic self-efficacy clearly reflect that the students in this study place a great deal of emphasis on their ability to stay current with work, their motivation to complete academic work, and feelings of learning. These factors subsequently shape student developmental perceptions of their academic self-concept and self-identity as engineers, and ultimately their decision to remain as an engineer or change majors. In the context of this study, the more academically aligned variables of self-assessed satisfaction with academic life and self-efficacy in hardest course and the factors academic integration and academic self-efficacy were more influential indicators of change of major and the selection of program of study than social integration or sense of belonging. An implication of these findings is that the application of predictive analytics should not only be used to determine which students are at-risk, but also how analytics should be applied to help students develop and receive the feedback, support, and
involvement they need to stay current and motivated to complete academic work, and thereby improve their feelings of learning and self-concept. With these findings in mind and their association to contemporary research, a discussion is warranted on the role of predictive analytics in the future direction of the academy.

**Directions for Policy and Practice**

The impetus to use predictive analytics in higher education continues to forge ahead. At the institutional level, motivation for using predictive analytics often center on improving the return on investment related to recruitment costs. Students can be identified who are likely to persist and are, therefore, more heavily recruited. Similarly, admission standards may be adjusted to increase or decrease enrollment in alignment with particular demographics. Once students are admitted, institutions are focused on operational efficiency and effectiveness by understanding which majors are showing growth through employment trends and, subsequently, hiring additional faculty and advisors to support those growth trends as well as possibly initiating capital improvement projects. Taken together, analytics support the decisions made by institutional leaders, enabling them to demonstrate success in key areas that accrediting agencies, the Federal government, and other important entities (e.g., governing bodies and state legislatures that allocate funding) are focused on.

Although there is a strong institutional motivation for using predictive analytics, important student-oriented reasons also exist. Among the reasons is the ability to identify individual students at-risk of not completing courses or programs and then to appropriately intervene (Siemens & Gasevic, 2012). Such efforts may enhance the academic integration needs of students, thereby improving their chances of success. From a pedagogical perspective, predictive analytics may improve student learning by providing insights about
what could be done to improve instruction by using additional or alternate learning materials or engaging with students directly (Dietz-Uhler & Hurn, 2013; Long & Siemens, 2011). Direct engagement might occur through a meeting with an academic advisor or individualized feedback with faculty, thereby enhancing their feelings of academic self-efficacy and support. These types of interactions are reflective of the feedback-support-involvement triad that Tinto and Pusser (2006) discussed in their model. When the use of analytics can directly benefit students in this manner and the institution to achieve its goals, the chances of acceptance within the culture are greatly improved.

While it may seem that analytics can only be perceived as advantageous, there are valid reasons to remain attentive to the consequences of becoming overly dependent on analytics. Recently, Scholes (2016) elaborated on the ethical concerns of analytics-based “screening”. Scholes cited other researchers (Schauer, 2003; Lippert-Rasmussen, 2007, 2011, 2014) who asserted there is no fundamental difference between assessing an individual in relation to group-based risk statistics as compared to using more individualized measures and indicators. However, she contended there are ethical concerns regarding the use of analytics with respect to limiting individual agency and autonomy. From this viewpoint, ethical concerns and challenges are important to deliberate when considering to use analytics.

**Ethical concerns and challenges**

Given that opportunities to collect educational data will continue to expand as Internet connected devices more formally permeate our daily lives, educational researchers and policy makers must carefully consider which portions of the volumes of data are ethical to amass, especially if they contain identifiable information, in the name of reducing student risk and institutional improvement. Developing overarching and cohesive policies in regards
to data mining and the sharing of data should be transparent to students. Informing students of how their data are shared, and with whom, is essential. As analytical tools and techniques are developed and continue to attract institutional funding, educational practitioners, policymakers and researchers should proceed with thoughtful consideration about the choice of which data is shared and how it is secured by commercial predictive analytics systems, especially those systems that are hosted off-site by vendors. According to Wang (2014), there exists a “fine line between using data to facilitate student learning and using data to deny the credit of students’ further development” (p. 3). As Scholes (2016) has cautioned, institutional policymakers must carefully consider whether providing archived student data to predictive analytic systems will lead to consequences, such as loss of private data or denying certain segments of the population admission or pigeonholing students into certain academic majors while other segments retain their agency and freedom to pursue their interests.

As policy regarding student data privacy continues to be defined, the demand for additional data sources has increased. Mattingly, Rice, and Berge (2012) concluded in their review of practices on the collection, analysis, and reporting of data as predictors of student success, that the availability of ample data is instrumental in guiding departmental processes and curriculum decisions. The tracking of the daily activities of students is possible in many ways: classroom attendance, classroom participation, online coursework completion, social network interactions, and the use of ancillary services such as recreational facilities, dining centers, and the library. A term that has been used recently to describe this type of tracking is “dataveillance” (Manca, Caviglione, & Raffaghelli, 2016). A potential result of dataveillance is that students may be profiled and classified without knowing or having had an opportunity to react or respond. The possibility of coincidental associations leading to
incorrect categorizations is possible. As stated by Mayer-Schönberger and Cukier (2013), correlation does not imply causation; nevertheless, correlation does provide a “glimpse of the important variables that we then use in experiments to investigate causality” (p. 66). Thus, inferences drawn based on correlated data should be scrutinized and subjected to deep, complex analysis before decisions are made that might, in some way, affect students’ ability to choose and/or breach their autonomy. Analysis of this nature will require institutions to employ data scientists with advanced data literacy skills.

In addition to an adequate supply of data scientists to conduct analyses and an IT infrastructure to support the secure storage data and retrieval of data, institutions that intend to connect programs and services with analytics must give some forethought to which metrics will reflect the outcomes they are interested in obtaining. When considering student outcomes, an institution will often turn to obvious measurements such as retention or completion rates. With respect to the use of predictive analytics, the impact on student success may not be as straightforward or measured immediately. Thus, the current study attempted to go beyond common forms of measurement that require numerous semesters and, possibly, many years of data to demonstrate an effect. A methodology was used that relies on observational data for evaluating the impact of predictive analytics on change of major and the selection of a program of study for students who received academic advising informed by predictive analytics.

A specific technological area of concern related to analytics, as noted by Thorn, Meyer, and Gamoran (2007), is the issue of segregated data silos thereby leading to incomplete conclusions regarding student outcomes. As systems have been constructed throughout the years, forethought may not have been given how to establish connections
across data sources to address complex educational problems. As technology has advanced, a solution to segregated data systems has been the establishment of centralized data warehouses where data is imported, cleaned, and standardized leading to stronger connections. Kitchin and Lauriault (2015) asserted that valuable insights are obtainable by linking smaller data datasets with larger institutional data sets such as those found in the data warehouses or enterprise resource planning (ERP) systems. In accordance with this viewpoint, this study attempted to connect data from two distinct data systems—a Mapworks dataset and a larger student information system dataset—so that a more complex statistical method (i.e., propensity score analysis) could be used to answer a more complex question (i.e., how to measure the impact of a new program such as predictive analytics in academic advising). It is through the linkage of multiple datasets that researchers can investigate causal relationships that explain and validate theory on a larger scale leading to more generalizable results. Researchers must be bold and identify new opportunities to use new data sources to construct holistic, inclusive models that explain student actions more completely. In addition, practitioners must be willing to collaborate with researchers and make data accessible in a centralized, secure manner.

The statewide longitudinal database systems (SLDS) developed in recent years provide an exemplary model for higher education institutions in this regard. As noted by Starobin and Upah (2015), a major contribution as well as strength of the SLDS system has been an increased awareness and understanding of the milestones and indicators that lead to student success. The ability to track students longitudinally throughout their entire academic lifetimes enables educators to identify barriers to their success. The comprehensive tracking of student progress over their academic careers requires the development of policy and
practices that effectively guide students, staff, and faculty on data use, without infringing upon student privacy and confidentiality (Moore, Offenstein, & Shulock, 2009; Leinbach & Jenkins, 2008).

**Importance of faculty engagement with analytics**

Although a great deal of what students do in terms of learning can be garnered from the online data they produce and the analytics which use that data, a critical next step is to actually intervene and interact with the students. While institutional support of student success may be reflected by the acquisition of an analytics system, integrating that tool into the academic culture and practice of those that directly interact with students is complex and requires advocacy and commitment from leaders. Tinto (2006) asserted:

> …many faculty believe to be the root causes of attrition, namely the lack of skills and motivation they might observe, that they would not have a retention “problem” if the admission office only admitted more qualified students. Consequently, they would argue, at least privately, that student retention is appropriately the job of student affairs professionals, and in particular, those who work in the area of developmental education. (p. 9)

In response to this sentiment, Tinto contends that student retention and, thereby, student success, is directly linked to the early postsecondary education that students receive through their interactions with faculty. Thus, faculty are an integral part of the support-feedback-involvement triad that Tinto and Pusser referred to in their model. Findings of research conducted by Ramos-Sánchez and Nichols (2007) suggest that faculty members often have more direct contact with students than do academic advisors, and they are more likely to encounter a student having academic or adjustment difficulties.

This study demonstrated the importance of self-efficacy and academic integration in relation to the change of major in particular. By working with academic advisors and using tools such as analytics, faculty are able to make instructional adjustments as well as
appropriate referrals, and provide constructive feedback and integrative experiences as students form their academic self-concept in their chosen field of study. As the collection and integration of student data continues to expand and complex analytical models are developed and exposed through dashboards and reports, faculty will likely be expected to be able to utilize analytics to help them refine their instructional approaches, improve learning, and demonstrate effectiveness when interacting with students.

Specifically, faculty may engage more directly with the use of data and analytics in the area of curriculum planning and development. As tools and software associated with curriculum analytics mature, programs may be evaluated on indicators such as graduation rates and time to degree, learning assessments and outcomes, and post-graduation employability. Strategically, curriculum analytics can generate faculty interest and involvement by producing more effective and efficient course pathways – eliminating courses that are no longer relevant or by adding courses that increase the value of the degree. Taking curriculum practice to a higher level, some researchers have also proposed a new line of faculty development that would include the scholarship of curriculum practice (SoCP) which would emphasize methodologically rigorous and systematic approaches to curriculum development along with peer review and public dissemination of curriculum inquiries (Hubball, Gold, Mighty, & Britnell, 2007; Richlin & Cox, 2004).

**Recommendations**

Recommendations for the future use of analytics are contextualized based on the level of the stakeholders involved. These levels are positioned as follows: state (mega), institutional (macro), college or departmental (meso), and student (micro).
At the state (mega) level, there is large potential to discover new insights by facilitating cross-institutional use of data from all levels of the PK-20 learning spectrum. The statewide longitudinal data systems (SLDS) provides one data source. The ability to link SLDS data with higher education data would provide a unique research opportunity into investigating the academic and social elements that influence student outcomes. The linking of these rich datasets across institutions would provide valuable, holistic guidance to state decision-makers in regards to funding, resource alignment, and educational policymaking.

At the institutional (macro) level analytics are focused on strategic issues that spread across the entire institution for optimizing effectiveness and efficiency. In regards to student outcomes, increasing retention and completion are viewed as two of the most important metrics that are routinely used and monitored. Improving retention of at-risk students by even a few percentage points can provide an immediate return on investment and justify the acquisition analytic applications. Other institutional analytics may focus on student recruitment, academic advising, financial decisions, process improvements, and improved customer services and satisfaction. These types of analytics help leaders demonstrate to governing bodies that institutions are making evidence-based decisions and not wasting taxpayer funds. A recommendation for institutional leaders is to not allow analytics to overshadow the development of a student’s individual self-efficacy and agency in the name of efficiency.

Analytics at the departmental (meso) level can be used to support curriculum decisions and also to improve learning. Regarding curriculum, analytics can be used to monitor course demand, enrollment patterns, and the gauge the complexity of marker courses as reflected by the academic performance of students, thereby influencing the sequence of
courses and the need for additional faculty. In addition, analytics can be used for more pedagogical purposes, such as monitoring learning outcomes and informing faculty where students are having difficulty, thereby leading to opportunities for feedback, support, and involvement with students to improve their academic self-concept, increase their self-efficacy, and create deeper academic integration experiences. A philosophical issue that may intrude on academic freedom is whether faculty are obligated to use the information provided by analytics to improve student learning. A recommendation for departmental leaders is to provide data-use support and training to enable the integration of analytics into departmental curriculum. As faculty become familiar and increase their use of analytics they will be able to improve their instructional efforts, especially in key marker courses.

At the student (micro) level, analytics can serve learners by providing students with feedback regarding their understanding of concepts. If taken to the next step, adaptive learning environments could be added to the instructional environment (i.e., via the learning management system, for example). In this way, learners will receive more immediate and individualized instruction on particularly difficult concepts that impede their progress. A recommendation, however, is to not allow students to interpret their comparative academic performance with others, as reflected through an analytics dashboard, in isolation. When left to their own interpretation, students may become overwhelmed and prematurely withdraw due to reduced self-efficacy. Therefore, policy must be mindful of what information is exposed to students directly versus what is shared through developmental meetings with advisors and faculty.
Limitations

As mentioned by Salkind (2003), a challenge in using the quasi-experimental design rests in the selection of pre-treatment independent variables. When using observational data, demographic data such as gender, ethnicity, and age are not characteristics the researcher can control as in experimentally designed studies. There are choices to consider with regards to maintaining a sufficient sample size and upholding unconfoundness while simultaneously facilitating the inclusion of variables that are informative to the analysis. Variables that associate with the dependent variable (i.e., the outcome) but differ between the groups are viewed as particularly informative for use as matching variables. As asserted by Salkind (2003), the researcher should anticipate differences will exist and use analyses such as $t$-tests to determine the extent of the differences to inform and help interpret subsequent analyses.

In addition, the decision was made in this study to select only students classified as freshmen in both the fall and spring terms who were also 18 years of age and younger as being representative of a genuine first-year, first-term freshmen experience. The outcome of this decision reinforced an existing limitation in terms of the subjects being overwhelming white, thereby limiting how these results might apply to a more diverse subpopulation or underrepresented minorities.

In contrast to a quasi-experimentally designed study, a traditional experimental study is more likely done in shorter period of time and conditions are more closely controlled, including the delivery of the treatment. However, in quasi-experimental designs, the temporal nature of the observational data may introduce unobserved factors. Therefore, researchers must try to include all pre-treatment data that is possible and pertinent. Thus, an opportunity for future research would be to link to additional data sets that shed light on
models of student success. However, in a traditional university data environment, where data exists in distinct legacy silos, this may be a challenging obstacle to overcome. As institutional-wide centralized data warehouses that collect, merge, and expose data to researchers through a unified interface continue to mature, this evolution will play a crucial role in providing the necessary datasets researchers need to develop more complex models.

This study focused on the transitional experiences of first-year, first-term engineering students at a very early stage of their postsecondary careers. A possible limitation is that student responses to the transitional survey may have not have been fully formed at the beginning of the semester. Future researchers may find it useful to interpose additional transitional surveys at the end of the first semester or end of the first year to obtain more deeply formed perceptions regarding integration and self-efficacy. A survey instrument could also be constructed that focuses on disciplinary-specific experiences in the use of the predictive analytics software and how it impacts students. As more is known about the logging capabilities of the predictive analytics software and the reports that are available, it may be possible to obtain additional data that more fully describe the details of the interactions between students and advisors. Such data may reveal measures that more accurately describe the intensity of the use of the predictive analytics. Potential measures that would be insightful to obtain might include the frequency of interactions between students and advisors, length of those meetings, characteristics (i.e., e-mail, face-to-face, etc.) and purposes of those meetings (i.e., planning, check-ups, academic warnings, etc.). Finally, it may be interesting to determine if the advisors perceived purpose for using of the software was received in the same way by the students and achieved the intended outcomes.
In addition, because this study focused on a particular subset of students at a single institution who majored in engineering where they received similar advising treatment experiences, the generalizability of this study is limited. Inferences about causal relationships from a specific study need to be validated in additional settings and conditions. The ability to establish this type of external validity demonstrates how well conclusions generalize to other people in other settings substantiates causal inferences. This study revealed there was an impact created by the treatment associated with the use of predictive analytics in academic advising at a specific institution with a specific subset of students. In order to do this, one option would be to study the same types of students (i.e., first-year, first-term engineer students) at other similar research institutions where similar data are available. Finally, a possibility for future research is to consider conducting this research longitudinally across multiple years with the same students to determine how predictive analytics and self-identity (reflected as change of major or selection of program) continue to evolve.

An alternate statistical analysis suggestion for future researchers is to consider using a multinomial logistic regression. Akin to binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership (Starkweather & Moske, 2011; Tabachnick & Fidell, 2013). This type of regression method is useful when the dependent variable is strictly categorical and falls into one of a set of unordered categories. In this study, a researcher may wish to expand the dependent variable into a set of mutually exclusive categories reflective of additional outcomes of interest such as: withdrew, changed majors, selected program of study, or remained undeclared.
Given the vast amount of observational data growth in education, methods such as propensity score analysis should increasingly be emphasized. Through advanced data mining efforts and data warehouses, valuable data continues to be captured and stored, allowing researchers to construct and test more complex models. Due to the ethical challenges associated with experimental research designs on human subjects and obtaining sufficiently large sample sizes, utilizing observational data along with statistical methods such as propensity score matching is an approach which should be increasingly pursued. Scholars and researchers will merely need an interface that allows them to obtain data in a safe, de-identified, secure manner so they can construct and test more complex models and measure the impact of environmental modifications on student outcomes.

**Conclusion**

The recent literature on the emergence of analytics in education and the empirical research conducted on transitional factors that influence the selection and change of major of students in higher education provide a foundation for this study. From that grounded perspective, this investigation explored research questions surrounding variables that influenced first-year, first-term engineering student tendencies in regards to the selection of a program of study and change of major. The results support the significance of self-efficacy and academic integration as important elements in the formation of academic self-concept as an engineering student, reflected in student outcomes (i.e., the selection of program of study and change of major). Particularly noteworthy were the variables students emphasized as a part of their developing self-concept: remaining current with academic work, learning satisfaction, satisfaction with academic life, doing well on problems and tasks and in their hardest course, and persevering on class projects with challenges. It was also evident that,
while self-efficacy and academic integration play a role in student decision-making, there was an impact associated with the use of predictive analytics in academic advising, particularly on change of major.

Although the sample was comprised of first-year, first-term undeclared engineering students from a single major research university, the issues faced by these students are not unlike those encountered by students in engineering or STEM fields at other universities. By linking data from multiple data sources, a more holistic understanding of the critical variables that effect student decision-making. The findings of this study demonstrated that joining data sources from disjoint systems is viable and can lead to new understandings of the influences that predict and impact the selection of a program of study and change of major. It is anticipated that this study will encourage future researchers to construct more complex models that contribute to the development of policies and practice which, in turn, lead to improved student success in higher education.
APPENDIX A. LITERATURE MAP
### APPENDIX B. CORRELATION MATRIX

<table>
<thead>
<tr>
<th></th>
<th>GENDER</th>
<th>ETHNICITY</th>
<th>HS_RANK</th>
<th>ACT_CMPST</th>
<th>ACT_MATH</th>
<th>HP_MEM</th>
<th>LS</th>
<th>ACT_CRES_ATT</th>
<th>FT_CRES_ATT</th>
<th>FT_CRES_COMP</th>
<th>FT_GPA</th>
<th>DEG_COMM</th>
<th>SA_MATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>1</td>
<td>-0.03</td>
<td>0.12**</td>
<td>-0.08**</td>
<td>0.08**</td>
<td>0.047</td>
<td>0.07**</td>
<td>-0.13**</td>
<td>0.28**</td>
<td>0.079</td>
<td>0.37**</td>
<td>0.037</td>
<td>0.017</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>-0.09%</td>
<td>1</td>
<td>0.025</td>
<td>-0.03</td>
<td>0.023</td>
<td>0.018</td>
<td>-0.028</td>
<td>-0.09%</td>
<td>-0.08%</td>
<td>0.01</td>
<td>-0.02%</td>
<td>-0.057</td>
<td>0.017</td>
</tr>
<tr>
<td>HS_RANK</td>
<td>210**</td>
<td>0.25%</td>
<td>1</td>
<td>0.288**</td>
<td>-0.277**</td>
<td>-0.208</td>
<td>-0.086</td>
<td>0.080**</td>
<td>0.554**</td>
<td>0.314**</td>
<td>0.407</td>
<td>1.15**</td>
<td>0.14**</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-0.05%</td>
<td>-0.2%</td>
<td>1</td>
<td>0.712**</td>
<td>0.707**</td>
<td>0.022</td>
<td>-0.016</td>
<td>0.077**</td>
<td>0.35**</td>
<td>0.25**</td>
<td>0.097</td>
<td>0.047</td>
<td>0.005</td>
</tr>
<tr>
<td>ACT_MATH</td>
<td>-0.06%</td>
<td>-0.26%</td>
<td>0.277**</td>
<td>0.712**</td>
<td>0.711**</td>
<td>0.051</td>
<td>-0.025</td>
<td>0.077**</td>
<td>0.49**</td>
<td>0.28**</td>
<td>0.056</td>
<td>-0.012</td>
<td>0.117</td>
</tr>
<tr>
<td>HP_MEM</td>
<td>0.048</td>
<td>0.018</td>
<td>0.208**</td>
<td>0.30%</td>
<td>0.211**</td>
<td>1</td>
<td>0.065</td>
<td>0.165**</td>
<td>2.63**</td>
<td>1.94**</td>
<td>0.101</td>
<td>0.033</td>
<td>0.096</td>
</tr>
<tr>
<td>LS</td>
<td>-13%</td>
<td>-0.5%</td>
<td>-0.66%</td>
<td>-0.032</td>
<td>0.051</td>
<td>0.065</td>
<td>1</td>
<td>-0.212**</td>
<td>2.17**</td>
<td>0.055</td>
<td>0.005</td>
<td>0.037</td>
<td>0.004</td>
</tr>
<tr>
<td>ACT_CRES_ATT</td>
<td>0.24%</td>
<td>0.09%</td>
<td>0.08%</td>
<td>-0.036</td>
<td>-0.025</td>
<td>1.65**</td>
<td>-0.231**</td>
<td>1</td>
<td>0.055</td>
<td>0.038</td>
<td>0.056</td>
<td>0.009</td>
<td>-0.058</td>
</tr>
<tr>
<td>FT_CRES_ATT</td>
<td>-0.05%</td>
<td>-0.26%</td>
<td>0.07%</td>
<td>0.137**</td>
<td>0.147**</td>
<td>0.34%</td>
<td>0.217**</td>
<td>0.058</td>
<td>0.241**</td>
<td>1.11**</td>
<td>0.095</td>
<td>0.036</td>
<td>0.195</td>
</tr>
<tr>
<td>FT_CRES_COMP</td>
<td>0.115</td>
<td>-0.0%</td>
<td>0.81**</td>
<td>0.728**</td>
<td>0.797**</td>
<td>1</td>
<td>0.056</td>
<td>0.087</td>
<td>2.41**</td>
<td>1.511**</td>
<td>0.026</td>
<td>0.218</td>
<td>0.098</td>
</tr>
<tr>
<td>FT_GPA</td>
<td>0.124</td>
<td>-0.02%</td>
<td>0.495**</td>
<td>0.255**</td>
<td>0.255**</td>
<td>0.312</td>
<td>0.017</td>
<td>0.035</td>
<td>0.084**</td>
<td>0.611**</td>
<td>0.021</td>
<td>0.229</td>
<td>0.098</td>
</tr>
<tr>
<td>DEG_COMM</td>
<td>0.027</td>
<td>-0.05%</td>
<td>0.041</td>
<td>-0.087</td>
<td>0.012</td>
<td>-0.039</td>
<td>0.004</td>
<td>0.026</td>
<td>0.002</td>
<td>0.026</td>
<td>1</td>
<td>0.185</td>
<td>0.005</td>
</tr>
<tr>
<td>SA_MATH</td>
<td>0.017</td>
<td>-0.05%</td>
<td>0.183**</td>
<td>0.244**</td>
<td>0.417**</td>
<td>0.036</td>
<td>0.046</td>
<td>0.059</td>
<td>0.095</td>
<td>0.085</td>
<td>0.518</td>
<td>2.295</td>
<td>0.001</td>
</tr>
<tr>
<td>SA_TW_TW</td>
<td>0.061</td>
<td>-0.01%</td>
<td>0.068</td>
<td>0.044</td>
<td>0.062**</td>
<td>0.043</td>
<td>0.083</td>
<td>0.01</td>
<td>0.012</td>
<td>0.012</td>
<td>0.156</td>
<td>1.175</td>
<td>0.001</td>
</tr>
<tr>
<td>SA_PLAN</td>
<td>-0.175%</td>
<td>-0.09%</td>
<td>0.094**</td>
<td>-0.092**</td>
<td>0.002</td>
<td>0.083</td>
<td>0.059</td>
<td>0.024</td>
<td>0.179**</td>
<td>0.868**</td>
<td>0.232</td>
<td>0.078</td>
<td>0.035</td>
</tr>
<tr>
<td>SE_PRBS_TASKS</td>
<td>-0.105%</td>
<td>0.015</td>
<td>0.082</td>
<td>0.081**</td>
<td>0.017**</td>
<td>0.008</td>
<td>0.067</td>
<td>0.008</td>
<td>0.298</td>
<td>0.998</td>
<td>0.041</td>
<td>0.372</td>
<td>0.019</td>
</tr>
<tr>
<td>SE_HARD_CRES</td>
<td>-0.067</td>
<td>0.09%</td>
<td>0.012</td>
<td>0.065**</td>
<td>0.029**</td>
<td>0.018</td>
<td>0.085</td>
<td>0.035</td>
<td>0.199**</td>
<td>0.199**</td>
<td>0.029</td>
<td>0.165</td>
<td>0.032</td>
</tr>
<tr>
<td>SA_ACAD_LIFE</td>
<td>0.001</td>
<td>0.05%</td>
<td>0.03%</td>
<td>0.095**</td>
<td>0.131**</td>
<td>-0.023</td>
<td>-0.047</td>
<td>-0.04</td>
<td>0.022</td>
<td>0.249</td>
<td>0.365</td>
<td>0.019</td>
<td>0.003</td>
</tr>
<tr>
<td>INST_SAT</td>
<td>0.068</td>
<td>0.025%</td>
<td>0.045</td>
<td>0.006</td>
<td>0.003</td>
<td>0.038</td>
<td>0.02</td>
<td>0.031</td>
<td>0.035</td>
<td>0.073</td>
<td>0.02</td>
<td>0.466</td>
<td>0.019</td>
</tr>
<tr>
<td>ACAD_SE</td>
<td>-0.053</td>
<td>0.037%</td>
<td>0.043</td>
<td>0.115**</td>
<td>0.066</td>
<td>0.062</td>
<td>0.021</td>
<td>0.033</td>
<td>0.110**</td>
<td>0.118</td>
<td>0.251</td>
<td>3.511**</td>
<td>0.001</td>
</tr>
<tr>
<td>ACAD_INT</td>
<td>0.033</td>
<td>0.07%</td>
<td>0.024</td>
<td>0.002</td>
<td>0.025**</td>
<td>0.009</td>
<td>0.016</td>
<td>0.009</td>
<td>0.207**</td>
<td>0.025</td>
<td>0.195</td>
<td>0.320**</td>
<td>0.001</td>
</tr>
<tr>
<td>SOC_INT</td>
<td>0.093</td>
<td>0.013%</td>
<td>0.004</td>
<td>-0.034</td>
<td>-0.031</td>
<td>0.017</td>
<td>0.063</td>
<td>0.003</td>
<td>0.019</td>
<td>0.073</td>
<td>0.021</td>
<td>0.014</td>
<td>0.001</td>
</tr>
<tr>
<td>REL_GLM</td>
<td>-0.019</td>
<td>0.037%</td>
<td>0.096</td>
<td>-0.048</td>
<td>-0.029</td>
<td>0.049</td>
<td>0.024</td>
<td>0.02</td>
<td>-0.189**</td>
<td>0.131**</td>
<td>0.075</td>
<td>0.057</td>
<td>0.005</td>
</tr>
<tr>
<td>CHG_MA1</td>
<td>0.028</td>
<td>-0.02%</td>
<td>0.059</td>
<td>-0.105**</td>
<td>-0.115**</td>
<td>-0.081</td>
<td>0.042</td>
<td>0.046</td>
<td>0.089**</td>
<td>0.213**</td>
<td>0.121</td>
<td>-0.073</td>
<td>-0.024</td>
</tr>
<tr>
<td>Revd_Treatmnt</td>
<td>0.033%</td>
<td>0.026%</td>
<td>0.065</td>
<td>-0.001</td>
<td>0.034</td>
<td>0.059</td>
<td>-0.839**</td>
<td>1.711**</td>
<td>0.258**</td>
<td>0.046</td>
<td>0.036</td>
<td>0.005</td>
<td>-0.025</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Listwise N=772

Note: Shaded areas indicate results of interest related to this particular analysis.
APPENDIX C. PRELIMINARY LOGISTIC REGRESSIONS

The first preliminary logistic regression model used the following independent variables with the dependent variable, change of major:

- SIS variables: GENDER, HS_RANK, ACT_CMPST, FT_CRED_ATT, HP_MEM, LC_MEM
- Mapworks variables: DEG_COMM, SA_MATH, SA_PLAN, SE_HARD_CRSE, SA_ACAD_LIFE

Table C1. Change of Major with significant SIS variables and Mapworks variables

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_ATT</td>
<td>.152</td>
<td>.054</td>
<td>8.024</td>
<td>1</td>
<td>.005</td>
<td>1.164</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-.121</td>
<td>.039</td>
<td>9.843</td>
<td>1</td>
<td>.002</td>
<td>.886</td>
</tr>
<tr>
<td>SE_HARD_CRSE</td>
<td>-.439</td>
<td>.102</td>
<td>18.435</td>
<td>1</td>
<td>.000</td>
<td>.645</td>
</tr>
<tr>
<td>Constant</td>
<td>.431</td>
<td>1.286</td>
<td>.112</td>
<td>1</td>
<td>.738</td>
<td>1.539</td>
</tr>
</tbody>
</table>

Forward logistic regression with LR was conducted to determine which independent variables were indicators of change of major. The regression results indicated that the overall model fit of three indicators was moderate (-2LL = 449.064) but statistically better in distinguishing change of major ($\chi^2(3) = 35.534, p < .0001$) over the constant model, classifying 66% of the cases correctly as compared to the constant model which failed to differentiate change of major. The area under the ROC curve value was 0.700 indicating that on average, interpreted as a fair test of sensitivity against specificity with a 70.0% probability that student is classified correctly. Regression coefficients are presented above. The significance of the Wald statistic for first-term credits attempted indicates it increased the odds of changing majors. The significance of the Wald statistic for ACT composite score and self-efficacy regarding ability to succeed in hardest course decreased the odds of changing majors significantly. The odds ratio for each coefficient indicated a small effect on the outcome. The small effect size was reflected in the Nagelkerke pseudo-$R^2$ (.096).
The second preliminary logistic regression model used the following independent variables with the dependent variable, change of major:

- SIS variables: GENDER, HS_RANK, ACT_CMPST, FT_CRED_ATT, HP_MEM, LC_MEM
- Mapworks factors: ACAD_SE, ACAD_INT, SOC_INT

Table C2. Change of Major with SIS variables and Mapworks factors

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>ρ</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_CRED_ATT</td>
<td>.164</td>
<td>.053</td>
<td>9.404</td>
<td>1</td>
<td>.002</td>
<td>1.178</td>
</tr>
<tr>
<td>ACT_CMPST</td>
<td>-.121</td>
<td>.038</td>
<td>10.061</td>
<td>1</td>
<td>.002</td>
<td>.886</td>
</tr>
<tr>
<td>ACAD_SE</td>
<td>-.565</td>
<td>.125</td>
<td>12.476</td>
<td>1</td>
<td>.000</td>
<td>.568</td>
</tr>
<tr>
<td>Constant</td>
<td>1.033</td>
<td>1.319</td>
<td>.613</td>
<td>1</td>
<td>.434</td>
<td>2.808</td>
</tr>
</tbody>
</table>

Forward logistic regression with LR was conducted to determine which independent variables and factors were indicators of change of major. The regression results indicated that the overall model fit of three indicators was moderate (-2LL = 453.225) but statistically better in distinguishing change of major ($\chi^2(3) = 39.251, p < .0001$) over the constant model, classifying 67.1% of the cases correctly as compared to the constant model which failed to differentiate change of major. The area under the ROC curve value was 0.716 indicating that on average, interpreted as a fair test of sensitivity against specificity with a 71.6% probability that a student is classified correctly. Regression coefficients are presented above. The significance of the Wald statistic indicated that an increase in first-term credits attempted increased the odds of changing majors. The significance of the Wald statistic for ACT composite score and self-efficacy regarding the ability to succeed in hardest course decreased the odds of changing majors. The odds ratio for each coefficient indicated a small effect on the outcome. The small effect size was reflected in the Nagelkerke pseudo-$R^2$ (.104).
The third preliminary logistic regression model used the following independent variables with the dependent variable, selection of program of study:

- SIS variables: GENDER, HS_RANK, ACT_CMPST, FT_CRED_COMP, HP_MEM, LC_MEM
- Mapworks variables: DEG_COMM, SA_MATH, SA_PLAN, SE_HARD_CRSE, SA_ACAD_LIFE

Table C3. Selection of program of study with SIS variables and Mapworks variables

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>ρ</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEG_COMM</td>
<td>.260</td>
<td>.075</td>
<td>11.818</td>
<td>1</td>
<td>.001</td>
<td>1.296</td>
</tr>
<tr>
<td>SA_ACAD_LIFE</td>
<td>.167</td>
<td>.057</td>
<td>8.539</td>
<td>1</td>
<td>.003</td>
<td>1.182</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.212</td>
<td>.501</td>
<td>19.493</td>
<td>1</td>
<td>.000</td>
<td>.109</td>
</tr>
</tbody>
</table>

Forward logistic regression with LR was conducted to determine which independent variables were indicators of change of major. The regression results indicated that the overall model fit of two indicators was not strong (-2LL = 1031.108) but statistically better in distinguishing change of major ($\chi^2(2) = 28.846$, $p < .0001$) than the constant model, classifying 58.4% of the cases correctly as compared to the constant model which failed to differentiate students who selection of program of study. The area under the ROC curve value was 0.592 indicating that on average, interpreted as a weak test of sensitivity against specificity with a 59.2% probability that a student is classified correctly. Regression coefficients are presented above. The significance of the Wald statistic for commitment to completing degree and self-assessed satisfaction with academic life indicates they increased the odds of selection of a program of study. The odds ratio for each coefficient indicated a small effect on the outcome. The small effect size was reflected in the Nagelkerke pseudo-$R^2$ (.049).
The fourth preliminary logistic regression model used the following independent variables with the dependent variable, selection of program of study:

- SIS variables: GENDER, HS_RANK, ACT_CMPST, FT_CRED_COMP, HP_MEM, LC_MEM
- Mapworks factors: ACAD_INT, SOC_INT, ACAD_SE

Table C4. Selection of program of study with SIS variables and Mapworks factors

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>ρ</th>
<th>Odds Ratio (Exp[B])</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAD_INT</td>
<td>.335</td>
<td>.075</td>
<td>20.068</td>
<td>1</td>
<td>.001</td>
<td>1.399</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.557</td>
<td>.420</td>
<td>13.738</td>
<td>1</td>
<td>.000</td>
<td>.211</td>
</tr>
</tbody>
</table>

Forward logistic regression with LR was conducted to determine which independent variables were indicators of change of major. The regression results indicated that the overall model fit of the single indicator was not strong (-2LL = 1063.507) but statistically better in distinguishing change of major ($\chi^2(1) = 20.718, p < .0001$) than the constant model, classifying 58.5% of the cases correctly as compared to the constant model which failed to differentiate students who selection of program of study. The area under the ROC curve value was 0.596 indicating that on average, interpreted as a weak test of sensitivity against specificity with a 59.6% probability that a student is classified correctly. Regression coefficients are presented above. The significance of the Wald statistic academic integration indicates it increased the odds of the selection of a program of study. The odds ratio for each coefficient indicated a small effect on the outcome. The small effect size was reflected in the Nagelkerke pseudo-$R^2$ (.035).
APPENDIX D. INSTITUTIONAL REVIEW BOARD APPROVAL

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4350
FAX 515 294-4267

Date: 9/1/2015
To: Sylvester Upah
    Rm 10 Enrollment Services
CC: Dr. Soko Starebin
    N221A Lagomarcino
From: Office for Responsible Research

Project Title: Investigating the impact of Predictive Analytics on First-term Freshmen Student Outcomes using a Quasi-experimental Design

The Co-Chair of the ISU Institutional Review Board (IRB) has reviewed the project noted above and determined that the project:

☐ Does not meet the definition of research according to federal regulations.
☒ Is research that does not involve human subjects according to federal regulations.

Accordingly, this project does not need IRB approval and you may proceed at any time. We do, however, urge you to protect the rights of your participants in the same ways you would if IRB approval were required. For example, best practices include informing participants that involvement in the project is voluntary and maintaining confidentiality as appropriate.

If you modify the project, we recommend communicating with the IRB staff to ensure that the modifications do not change this determination such that IRB approval is required.
REFERENCES


Dietz-Uhler, B., & Hurn, J. E. Using learning analytics to predict (and improve) student success: a faculty perspective. *Journal of Interactive Online Learning, 12*(1), 17-26


Schildkamp, K., & Poortman, C. L. Factors influencing the functioning of data teams. *Teachers College Record, 117*(5), 1-31.


UCLA Institute for Digital Research and Education. *How can I perform the likelihood ratio, Wald, and Lagrange multiplier (score) test in Stata?* Retrieved from http://www.ats.ucla.edu/stat/sas/notes2/


Xu, Y. J. (2016). The experience and persistence of college students in STEM majors. *Journal of College Student Retention: Research, Theory & Practice, 0*(0), 1-20.
