2017

Who Stays and Who Leaves? Predicting College Student Persistence Using Comprehensive Retention Models

Adam C. Elder
Louisiana State University and Agricultural and Mechanical College

WHO STAYS AND WHO LEAVES? PREDICTING COLLEGE STUDENT PERSISTENCE USING COMPREHENSIVE RETENTION MODELS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The School of Education

by
Adam C. Elder
B.S., Auburn University, 2010
M.A., Louisiana State University, 2014
May 2017

ACKNOWLEDGMENTS

My journey to this point has been full of mentors, colleagues, and cheerleaders who made it possible for me to reach this incredible milestone. Each of you inspired me, motivated me, and supported me in one way or another, and the joy of finishing this dissertation is shared with each of you.

Thank you to the extraordinary faculty at Louisiana State University for improving my intellect and challenging me to grow both as a scholar and as a person. I especially would like to express my appreciation for my dissertation committee members, Dr. Eugene Kennedy, Dr. Brian Marx, and Dr. Lauren Coats. Thank you for your time, your expertise, and your invaluable assistance throughout this process.

My deepest gratitude goes to my major professor and supervisor throughout the entirety of my journey as a graduate student, Dr. Keena Arbuthnot. You led me to discover a passion I did not know I had, you saw something in me that I did not see in myself, you consistently pushed me to always strive for the highest of heights, and you gave me the confidence to take on this degree and this dissertation. I would not have gotten to this point if it were not for you, and for that I will always be exceedingly grateful.

I also need to thank several friends and colleagues who made this study possible. Thank you to Christina Coovert and Justine Radunzel for helping me to fully formulate the topic of this study, providing me with access to data, and sharing your meaningful insight. Thank you to my dear friend Lauren Thackerson for graciously taking the time to review and provide feedback on various sections of the dissertation. A special note of thanks goes to Lupe Lamadrid for not only selflessly assisting me in any capacity as I worked on this project, but also doing it with an eagerness that was truly inspiring. It is an honor to have you as a friend and colleague.

Finally, I thank my family, friends, and classmates who have been steadfast supporters every step of this journey. I could not possibly express the appreciation I have for each of you and the love and encouragement you have shown me over the years. I specifically want to thank my incredible parents, Bruce and Dawn Elder. Your self-sacrifice afforded me the opportunity to fully pursue my dreams and made it possible for me to achieve my goals. Your reassurance and willingness to listen got me through the tough times; your excitement and pride made the good times even sweeter. These are the greatest gifts you have given to me. These are the gifts that made it possible for me to persevere. For these reasons and so many others, I dedicate this dissertation to you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this study was to use a comprehensive framework to examine academic, psychosocial, noncognitive, and other background factors that are related to retention at a large, public four-year institution in the southeastern United States. Specifically, the study examined what factors are most important in predicting first-to-second year retention both before the student enrolls at the university and after completion of their first semester of coursework. Data were drawn from institutional records, a survey instrument designed to measure psychosocial constructs, the ACT student record, and the National Center for Education Statistics. The sample for the study consisted of 12,342 students. Hierarchical generalized linear models and ensemble tree-based methods were utilized to identify important predictors of retention, ascertain the nature of the significant relationships, and to build models for predicting retention outcomes. An initial model was built for prediction before students enrolled followed by a second model with first semester performance variables added. Predictive validity was assessed by splitting the sample into a training and test set.

Findings from the study showed that nontraditional factors were significant predictors of retention along with traditional predictors such as high school GPA. The results showed that the influence of financial factors and high school characteristics were among the most significant predictors of retention. Moreover, the results showed that multiple psychosocial factors are influential variables in retention outcomes. This study demonstrated that considering a variety of factors when forecasting postsecondary retention outcomes is vital for more accurate predictions. The models in this study showed that pre-college predictive models have the potential to be nearly as effective as models incorporating college performance and activity. The results of this study have important implications for higher education policymakers, college administrators, and

high schools. Several of the relationships revealed have significant policy implications related to budget concerns, university programming, and college preparatory initiatives at the high school level. The study also provides a useful model for identifying students at risk of not being retained that could be adapted for implementation at other institutions and points the importance of a holistic understanding of the total student.

CHAPTER ONE
INTRODUCTION

The percentage of the United States population with postsecondary degrees has increased dramatically in the past twenty years. Forty six percent of 25 to 29 year olds have at least an associate's degree and 36 percent have at least a bachelor's degree, an increase of 13 and 11 percentage points, respectively (Kena et al., 2016). While the percentage of the population with postsecondary degrees has remained relatively unchanged in recent years, there has been an alarming downward trend in degree attainment for those who pursue a postsecondary education. The overall national six-year completion rate has dropped at an accelerated rate in recent years, with only 52.9 percent of students finishing a postsecondary degree in 2015 compared to 56.1 percent only two years earlier (Shapiro et al., 2015; Shapiro, Dundar, Yuan, Harrell, & Wakhungu, 2013). These percentages, and the rate at which they have declined in recent years, are concerning and problematic on many levels.

The potential effects of the declining rate of college graduates are dire. The demand for postsecondary education and certifications beyond a high school diploma is on the rise in the United States. Carnevale, Smith, and Strohl (2013) estimate that current economic growth trends indicate 65% of the 165 million jobs in the year 2020 will require additional education beyond a high school diploma. This is double the percentage from the early 1970s. They noted that if the current educational output were maintained, then several million jobs would go unfilled due to a shortage of qualified workers. Coupling this information with the lowering degree attainment rates, the United States is potentially facing a job vacancy problem that can seriously hinder economic growth because jobs would have to be outsourced to college graduates from other countries. At the student level, individuals who aspire to postsecondary degrees but do not attain them will face economic and other sociopolitical disadvantages compared to their peers who

complete college. Millennials with at least a bachelor's degree earn a median annual salary $17,500 higher than those with only a high school diploma and are less likely to be unemployed or live in poverty (Pew Research Center, 2014). Beyond the effects seen at the individual and global levels, students who do not complete a postsecondary degree also negatively impact colleges and universities.

Low graduation rates adversely affect postsecondary institutions in a multitude of ways. The first and most public consequence is lowered perception of institutional quality. A variety of college ranking systems exist, but the two most well-known are the annual U.S. News and World Report's Best Colleges Ranking and Forbes' Top Colleges Ranking. In both of these ranking systems, a college's graduation rate accounts for a portion of the final score that is used to order the schools. Specifically, six-year graduation rates account for 25% of the score in the U.S. News and World Report ranking while four-year graduation rates account for 7.5% of the score in the Forbes ranking (Howard, 2016; Morse, Brooks, & Mason, 2016). Therefore, institutions with low graduation rates are negatively impacted in the rankings which consequently has a deleterious effect on public perception of the college's quality. A less public, but highly consequential, effect of student attrition is the lost revenue absorbed by institutions. Since colleges earn money through paid student tuition and/or public funding formulas that are reliant on enrollment counts, students who fail to complete a degree represent lost revenue averaging over 20 million dollars annually (Raisman, 2013). This consequently has grave implications for institutions' operating budgets. It should be noted that this financial burden does not fall solely on colleges and universities because federal and state tax dollars fund higher education through direct appropriations and financial aid. First-year attrition alone costs federal and state governments nearly 1.5 billion dollars per year (Schneider, 2010). This has led to many states

cutting funding for higher education, which transfers the weight of funding colleges and universities onto students. These alarming statistics and their upward trend have pushed continued enrollment and degree completion to the forefront of higher education policy discussions.

College Student Retention

Student retention and graduation is one of the primary issues facing local and national higher education policy agendas, and extensive research has been conducted to better understand these two phenomena. The most proximal and straightforward means of understanding and improving graduation rates is to better understand and improve first-to-second year retention rates because the majority of students who leave college do so between their first and second year (Sullivan, 2010). For this reason, retention is one of the most discussed and researched topics in higher education (Berger & Lyon, 2012). The national average rate of retention is 80 percent for four-year institutions and 61 percent for two-year institutions (Kena et al., 2016). Those students who are not retained either transfer to another institution or drop out of college altogether; and students leave their initial college or university for a variety of reasons including academic failure, negative experiences with the institution, lack of institution-student fit, and external influences (Yorke & Longden, 2004). To combat these issues, postsecondary institutions spend extensive resources addressing enrollment issues in an attempt to retain and ultimately graduate students. Researchers have explored factors that are predictive of student retention in great detail in the last several decades in order to inform policy and practice at the institutional level.

Studies in this area traditionally have focused on demographic and academic characteristics related to student attrition. A multitude of studies have found a positive

relationship between academic success and standardized test scores, high school performance and preparation, and socioeconomic factors (J. Allen, Robbins, Casillas, & Oh, 2008; Godfrey, Matos-Elefonte, Ewing, & Patel, 2014; Radunzel & Noble, 2012; Robbins et al., 2004). Moreover, studies focusing specifically on retention also found these factors to be predictive of student attrition. For this reason, these metrics typically bear the most consideration in admissions decisions at colleges and universities around the country (Clinedinst, Koranteng, & Nicola, 2015). These metrics alone, however, do not fully address the variation in what leads to students being retained or not. Studies have shown that traditional metrics can account for as little as 10 percent of the variation in retention outcomes (Robbins et al., 2004). This has led to calls from both researchers and practitioners to examine additional factors that are predictive of success and consequently retention in college (Burrus et al., 2013; Conley, 2007; Mattern et al., 2014). These new factors, often called noncognitive factors, incorporate a plethora of constructs that examine aspects of students that cannot be understood from cognitive measures or demographic characteristics alone. The list of noncognitive factors that are studied throughout the retention literature is seemingly endless due to the broad definition used for this domain. Some of the most commonly studied factors that have been found to be related to college success are psychosocial factors such as educational goals and aspirations (D. Allen, 1999; Robbins et al., 2004), academic commitment and motivation (Robbins et al., 2004; Robbins, Allen, Casillas, Peterson, & Le, 2006), and self-efficacy (Gore, 2006; Komarraju & Nadler, 2013; Robbins et al., 2006). Additionally, other noncognitive measures such as institutional fit and major-interest fit have also been found to be significantly related to retention (J. Allen & Robbins, 2010; Mattern, Shaw, & Kobrin, 2010).

Even with the broadening definition of college readiness and the growing body of research in this area, colleges and universities across the country are still struggling to combat student attrition. There are two primary complications for institutions attempting to retain students. The first is that colleges are oftentimes unable to identify students who are likely to transfer or dropout. Typically, colleges are unable to identify these students until after the first semester, if at all. This constraint occurs due to incomplete and/or inefficient data management systems because colleges frequently store only high school and testing data that is used in their admissions process. This results in institutions losing information about students' pre-college traits and interests outside of academic performance. Additionally, it is not typical for institutions to collect comprehensive data related to noncognitive factors. The second complication facing institutions in their retention efforts is the dearth of research that studies a comprehensive approach towards retention by examining academic, psychosocial, and background characteristics conjunctively. The available literature in this area infrequently utilizes recent and/or comprehensive data sources. Furthermore, studies are oftentimes unable to use rigorous analysis methods due to data limitations. There are also several studies that explore specific noncognitive factors in depth, but a conglomerate study incorporating these factors with other comprehensive measures does not exist. Moreover, the majority of studies in this area typically examine large, multi-institutional samples. Tinto (1993) suggested that predictive accuracy and the relative strength of variables in retention models could vary across institutions. This hypothesis about differences between individual institutions and student populations has been supported by research; it is therefore inappropriate to interpret extant models as a one-size-fits-all understanding of factors related to retention (Astin & Oseguera, 2005; Bean, 2005;

Pascarella & Chapman, 1983). There is a need for rigorous research that is specific to individual institutions so that contextual issues are not overlooked in the findings.

Research Purpose and Questions

The purpose of this study was to use a comprehensive framework to examine academic, psychosocial, noncognitive, and other background factors that are related to retention at a large, public four-year institution. The research used rich data sources and rigorous analysis methods that are not common in the retention literature. The study examined retention from an institutional research perspective as it was focused on a single institution, but the results also have the potential to be viewed from individual and policy perspectives (Bean, 2005). The specific research questions that were answered were:

1. What pre-college academic, psychosocial, noncognitive, and background factors are predictive of first year retention at a large, public four-year institution in the Southeast?
2. How does the relationship between first year retention and pre-college academic, psychosocial, noncognitive, and background factors change when first semester college performance and activity is included in the model?
3. Are the relationships between first year retention and pre-college academic, psychosocial, noncognitive, and background factors moderated by demographic group membership?
4. Which factors are most important in predicting whether a student is retained?
5. What modeling approaches are the most accurate in predicting whether a student is retained?

Significance of the Study

This significance of this study primarily lies in the methodological approach used. This study used a large, thorough dataset that drew from multiple sources to examine factors related to

retention at a specific institution. Similar types of studies have been conducted before (e.g,. Singell & Waddell, 2010), but there are not any studies that use a composition of academic, psychosocial, and rich student background characteristics with a sizeable sample and rigorous analysis methods at a single institution. Furthermore, studies of this nature do not examine a rich set of pre-college factors. This study utilized variables from all of these domains to gain a more comprehensive understanding of the pre-college and post-matriculation factors that are related to first-to-second year retention.

The results of this study will be of primary interest to university administrators and student affairs practitioners at the university under study and potentially could be generalizable to similar institutions, though it would be most appropriate to replicate the study at other colleges and universities. The findings identified a variety of factors related to student retention and provided a model for identifying students at risk of dropout or transfer immediately upon enrollment. The study detailed an approach that is practical for institutions and can be a useful aid that can inform their retention efforts. Similarly, this study identified a variety of traditional and nontraditional variables related to retention that will be of particular interest to researchers and policymakers. The significant relationships found in this study have important implications for educational policy and illuminate opportunities for future research. Stakeholders at the secondary level will also find the results of interest due to the role high school factors played in the retention models.

Summary

Student retention is a major area of concern for higher education stakeholders due to the widespread effects it has on students and institutions. Colleges and universities have implemented programs and initiatives targeted at improving student retention in efforts to

address this issue; and researchers have extensively studied a variety of factors that are related to students staying enrolled at their institution. Furthermore, researchers and theorists call for the use of comprehensive measures beyond academic factors. Research has stressed the importance of individual institutional environments when examining factors related to retention, but minimal research has been conducted at the institutional level using rigorous data sources and analysis methods. The present study addresses this deficiency.

CHAPTER TWO
REVIEW OF THE LITERATURE

Retention of students at colleges and universities became a major point of concern in the 1960s and 1970s after several decades of burgeoning college enrollment. Despite this rapid growth, however, there was a dearth of empirically informed and theoretically grounded knowledge regarding undergraduate student persistence (Berger & Lyon, 2012). This newfound concern for postsecondary retention led to the development of several theories (e.g., Astin, 1977; Tinto, 1975) related to student attrition which were tested and refined through empirical research. As the field gained prominence, the body of research on retention became more extensive, existing theories were modified, and new theories emerged. Meanwhile, the face of higher education and higher education policy changed rapidly. This increased the need for meaningful research in this area. Retention is now a major higher education issue facing researchers, policymakers, and practitioners. The constant evolution of the higher education landscape demands the research on retention draws on sound theoretical perspectives, refine and expound upon the vast literature base on the topic, and employs methods that are reflective of the constantly changing populace and policy issues related to postsecondary education. This chapter highlights the theoretical perspective for this study and describes an overview of the literature on retention.

Theoretical Perspective

Tinto's (1975, 1993) Student Integration Model was a seminal theory regarding student departure from institutions. The foundational concept in this model was the idea of academic and social integration at the institution which, he argued, some degree of which "must exist as a condition for continued persistence" (Tinto, 1993, p. 120). Bean and Eaton (2000) proposed a model for retention that expanded on Tinto's model by shifting from a sociological to a

psychological perspective of student retention. This model, called the Psychological Model of College Student Retention, drew from attitude-behavior theory, coping behavioral theory, self-efficacy theory, and attribution theory. This model focuses on explaining persistence behavior choices and examining the motivations that drive these decisions. The model proposes that student interactions with the institutional environment are first shaped by past behaviors, beliefs, and attributes. The institutional environment then affects these interactions once the student enters college. Students' reactions to academic, social, and bureaucratic interactions as well as other interactions that are external to the institution then inform new psychological assessments about their self-efficacy and attributes. These linkages are constantly being evaluated and reevaluated, leading to the student's assessment of their academic and social integration at the university. These feelings of integration then inform the student's attitudes towards fitting in at the institution, which consequently inform the student's intention to persist. This model suggests that certain attributes are essential to academic and social integration including feelings of effectiveness in their social and academic environments, feelings of control of their own outcomes, and coping skills for dealing with academic and social difficulties that motivate them to face these challenges (Bean & Eaton, 2002).

The distinct difference between the Student Integration Model and the Psychological Model of Student Retention is that and Bean and Eaton's (2000) model is centered around student attitudes. Bean (2005) acknowledged that any person or experience on a college campus can impact these attitudes, that all campus entities are responsible for student persistence, and that "an institution needs to change what it is or what it does in order for retention rates to change" (p. 237). With this in mind, he outlined nine themes that affect retention: student background, money and finance, grades and academic performance, social factors, bureaucratic

factors, the external environment, psychological and attitudinal factors, institutional fit and commitment, and intentions. These themes address a variety areas and departments across a college campus and the overarching concept of college readiness, which is defined as the "set of skills, behaviors, attitudes, and knowledge, both cognitive and noncognitive, possessed by an individual student that shape their likelihood of attaining a college degree" (Nagaoka et al., 2013, p. 50). The theoretical perspective for this study is grounded in this framework and will explore factors and constructs that address each of these themes in order to identify specific areas where an institution can most strategically allocate their resources as they seek to improve student retention.

Overview of the Literature

The literature on factors related to college outcomes is extensive. This section highlights and summarizes three overarching areas that have been researched: academic factors, psychosocial factors, and background characteristics. Each of these areas and the various constructs within each attend to different components of Bean and Eaton's (2000) model of student retention. This overview will focus primarily on retention, but other postsecondary outcomes such as academic success are also addressed.

Academic Factors

High school grades and standardized test college admissions scores (i.e., ACT and SAT scores) are two measures of academic ability that have been studied extensively in relation to postsecondary outcomes. These metrics, along with high school coursework, are the leading measures used to make college admissions decisions (Clinedinst et al., 2015). High school GPA and ACT/SAT scores provide information about a student's previous academic performance and aptitude for success at the postsecondary level, respectively. High school coursework provides

information about the degree of rigor of a student's high school curriculum. Research has consistently supported the predictive relationship these variables have with academic success, retention, and degree completion in college. Moreover, high school academic factors consistently account for approximately one quarter of the variance in postsecondary outcomes (Kim, 2015; Robbins et al., 2004). These pre-college characteristics influence a student's collegiate academic interactions which is a guiding element of persistence in Bean and Eaton's (2000) framework.

High school grades. Extant research has consistently and convincingly shown that high school grades are one of the strongest predictors of academic success in college (Atkinson & Geiser, 2009). It is included as either a variable of interest or a control in virtually every study on postsecondary outcomes for this reason, and the findings are consistent across outcomes. Studies have shown high school GPA is strongly related to not only first year GPA in college, but also cumulative GPA, first-to-second year retention, and graduation (e.g., Radunzel & Noble, 2012; Robbins et al., 2004).

Kim (2015) examined the relationship between high school GPA, among other factors, and college performance and first-to-second year retention. Other predictors included standardized test scores, gender, ethnicity, and financial aid status. The study was conducted at a single public university in the Midwest over the course of three years from 2006 to 2009. The sample was comprised of 7,045 first-year students. One of the research questions for this study was focused on the difference between regular admission and special admission students, so 15% of the sample was comprised of students who had been specially admitted. Blockwise multiple and logistic regression analyses were used to assess the relationship between the predictors and criterion variables. Results showed that high school GPA was a highly significant predictor of

both first year GPA and retention for regular admission students, but it was not a significant predictor of retention for special admission students.

Geiser and Santelices (2007) conducted a study that examined the relationship of high school GPA along with other predictors with both short- and long-term postsecondary outcomes. The study was conducted at the University of California system with 79,785 freshmen between the fall of 1996 and the fall of 1999. For this study, unweighted high school GPA was used because previous research has demonstrated that weighted high school GPA lacks the predictive power of unweighted GPA (Geiser & Santelices, 2004). Multilevel models were used to examine the relationship while controlling for demographic and background characteristics. Results showed that high school GPA was the strongest predictor of academic performance in the first year of college, and that its predictive weight became stronger for cumulative GPA in subsequent years.

Despite the conclusive research on the strength of high school grades as a predictor of postsecondary outcomes, there are limitations with this measure which has a negative effect on their perceived reliability. First, high school GPA is subject to grade inflation or deflation—a situation in which a student's grades are not reflective of his or her true academic achievement or ability. This has long been a critique of this metric, and recent research has shown that this continues to be manifested across high schools (Zhang & Sanchez, 2013). Second, high school GPA is not standardized across high schools and students. GPAs are subject to varying grading scales and standards. Furthermore, students do not necessarily take the same courses throughout their high school curriculum, so a student's GPA can easily be influenced by the courses a student took. These limitations consequently threaten the validity of the interpretation of high

school GPA when comparing students. As a counterbalance to these issues, standardized test scores were developed in order to aid admissions personnel in their decision-making.

Standardized test scores. Standardized test scores, namely the ACT and SAT, are frequently used by admissions officers to make enrollment decisions at universities because they provide a common, standardized metric that allows for meaningful comparison of students from various backgrounds (Clinedinst et al., 2015). These assessments are meant to measure a student's aptitude for academic success in college. The two assessments have different foci, as the SAT is more focused on critical thinking and reasoning skills while the ACT is focused on curricular objectives (Zwick, 2007). Despite these different orientations, the composite scores for these two assessments are highly correlated (r = .92), as are their subtests (Dorans, 1999). This strong relationship has resulted in the two tests being used interchangeably throughout the literature on standardized assessment and its relationship with various postsecondary outcomes. Research has consistently shown that these assessments are positively related to postsecondary outcomes (ACT, 2014; Mattern & Patterson, 2009; Radunzel & Noble, 2012). Research has also shown that, along with high school GPA, standardized test scores regularly account for the greatest amount of variance in college success (e.g., Schmitt et al., 2009).

Like high school GPA, standardized test scores face limitations as a measure of academic ability. They are subject to confounding influences such as differences in assessed constructs across tests and varied curriculums across schools, districts, and states (Atkinson & Geiser, 2009). Another criticism of standardized assessment is that test scores are potentially influenced by environmental factors for students from low socioeconomic backgrounds, women, and minorities (Geiser & Santelices, 2007). This limitation subsequently has a negative influence on the predictive validity of standardized tests for these populations. Frequently cited possible

explanations for these trends are the availability of test preparation materials and other enhanced learning opportunities to higher income students, stereotype threat, and differences in cognitive processing (Arbuthnot, 2011; Briggs, 2009; Gallagher & Kaufman, 2005).

High school coursework. Students who undertake a more intense curriculum in high school are more prepared for the rigors of college coursework than students who do not. A rigorous high school curriculum consists of classes that are college preparatory in nature such as honors, advanced placement, dual enrollment, and international baccalaureate courses. Research has shown that students who take more difficult classes are more likely to be successful in postsecondary settings (Long, Conger, & Iatarola, 2012; Wyatt, Wiley, Camara, & Proestler, 2011). For example, An (2015) found a positive effect of participation in dual enrollment on first year college GPA, even after controlling for academic and background characteristics. This relationship was supported by Godfrey et al. (2014), who also found that students that perform well in advanced placement (AP) courses and on the AP test are more successful in college. In fact, Adelman (2006) found that high school coursework intensity (which he defined as highest math course taken) had a stronger relationship than GPA or standardized test scores with degree attainment. There are limitations to using the rigor of high school curriculum in predicting students' postsecondary achievement, however. The quality of these advanced courses can vary widely from school to school, so it can be difficult to understand the true degree to which a student took a rigorous curriculum in high school. It is also a difficult construct to measure consistently because the types of classes offered and taken are not consistent across students and schools.

Discrepancy between measures. It is important to recognize that each of the measures mentioned above face limitations in their validity for predicting academic success in isolation.

In order to minimize validity threats posed by these limitations, the measures are used jointly to gain a more reliable understanding of a student's academic capabilities (Clinedinst et al., 2015). However, these metrics can sometimes send mixed signals about a student if there is a discrepancy in these measures. For example, there are instances in which a student performs exceedingly well on a standardized test but has a poor high school GPA. Researchers have explored this discrepancy and its relationship with academic success in college.

Mattern, Shaw, and Kobrin (2011) investigated the prediction error of first year success for high school GPA and standardized test scores by controlling for the discrepancy between these two measures. The study utilized a sample of 150,377 students from 110 different institutions nationwide and multiple regression models with a variety of academic predictors. First year GPA was regressed on a series of various predictors including high school GPA and the three SAT sub-scores for critical reading, math, and writing. The composite SAT score and high school GPA were standardized and subtracted from one another to find the discrepancy. The correlation between the residuals and the discrepancy score was then computed in order to determine the nature of prediction error based on the traditional admissions metrics. The results showed that high school GPA or SAT scores alone can either over- or under-predict first year performance, especially for students with high or low values for either measure. The authors recommended individual institutions conduct validity studies on these measures in order to determine their relative importance in more specific contexts.

Lin and Sanchez (2017) explored the relationship between discrepant performance and first-to-second year retention utilizing a sample of nearly two million students at over two thousand colleges. High school GPA and ACT composite score were standardized and compared to create different discrepancy groups that reflected both direction (e.g., greater high

school GPA than ACT score) and magnitude (i.e., small or large discrepancy). Hierarchical logistic regression models controlled for demographic characteristics including gender, ethnicity, and family income. Interactions between discrepancy group and demographics were included in the model to examine differences in the effect of discrepant achievement between demographic groups. The results showed that students with ACT discrepancy (i.e., higher ACT score than high school GPA) were less likely to persist than students with GPA discrepancy. The magnitude of discrepancy was not found to have a strong relationship with persistence, but students with higher scores and GPAs were more likely to persist regardless of discrepancy group.

First year performance. The predictive ability of pre-college characteristics in understanding retention is well documented. Despite these significant relationships, however, many students who are expected to be successful in college are not. One possible explanation for this is a student's academic performance in their first year. A multitude of studies have shown that first year GPA is an highly significant predictor of retention (e.g., D. Allen, 1999; Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008). The effects of first year performance have been shown to diminish the effects of pre-college academic performance and aptitude, sometimes completely.

For example, I. Johnson (2008) conducted a study that examined the effects of both individual student and aggregate high school characteristics on enrollment, persistence, and graduation at a four-year university. For the retention model, a sample of 7,559 students from the 2001 to 2005 cohorts was used. Multilevel logistic regression models were run to analyze the relationships of interest. Results showed that first semester GPA was a highly significant predictor of retention. So much so, in fact, that once it was entered into the model, it made high

school GPA no longer significant and it changed the direction of the effect of SAT scores to negative. This finding and the comparable findings from similar studies suggest that first semester GPA is a critical variable to consider in studies on retention.

Shaw and Mattern (2013) conducted a study that examined the effects of the differential between a student's actual first year college GPA and their predicted GPA on retention to the second year of college and beyond. The study utilized a sample of 120,698 students from 78 different institutions around the country. Students' predicted first year GPA was calculated based on SAT scores and self-reported high school GPA within each institution. Hierarchical generalized linear models were run with first year GPA, a cubic discrepancy term, and controls for demographic characteristics at the individual level and controls for institution type and selectivity at the institution level. Results showed that students with higher GPAs were more likely to persist. An especially interesting finding was that students who performed as expected were the most likely to be retained, while students who either underperformed or over performed were more likely to not return to the institution.

These studies highlight the importance of taking a student's first year academic performance in college into consideration when discussing postsecondary outcomes. In addition to considering the student's GPA, discrepancy between expected and actual performance is also a meaningful factor to consider. In total, academic measures significantly aid researchers and practitioners in understanding postsecondary success. High school grades and curriculum and standardized test scores together are useful in understanding how a student may perform in college. This is especially true when these two measures are not congruent. Furthermore, the significance of a student's actual performance once they begin college cannot be ignored in discussions on retention. Understanding how actual performance and expected performance

interplay is additionally meaningful for future research to consider because this uses academic measures to explore a nonacademic construct (Shaw & Mattern, 2013). Nonacademic constructs, also called noncognitive or psychosocial factors, are meaningful variables to include and explore in retention studies.

Psychosocial Factors

Academic factors play a significant role in understanding postsecondary outcomes, but a large proportion of the variance remains unexplained by these variables alone. In an attempt to further explain the variation in educational performance measures, research has investigated psychosocial factors extensively in educational contexts, including the role these factors play in postsecondary outcomes. The definitions and constructs of psychosocial factors fluctuate widely across studies, so Robbins et al. (2004) conducted a rigorous meta-analysis of 109 studies in an attempt to clarify these constructs into common constructs with common meanings. In total, they produced nine different constructs that addressed some of Bean's (2005) nine themes that affect retention and persistence.

To synthesize the research on these constructs in the literature, Robbins and colleagues (2004) employed hierarchical linear models to identify the relationship between the traditional predictors of high school GPA, standardized test scores, and socioeconomic status as well as a host of psychosocial factors on first year GPA and retention. The inclusion of the traditional predictors allowed for a detailed understanding of the incremental validity of the psychosocial factors on these outcomes. For the retention models, factors included as predictors were academic goals, institutional commitment, social support, social involvement, academic self-efficacy, and academic-related skills. Results showed that, after controlling for socioeconomic status and academic background, the psychosocial factors contributed incrementally to the

prediction of retention. In particular, academic-related skills and institutional commitment were found to have a strong relationship with retention while the association with the other factors was moderate. The study showed that psychosocial factors explained over 13% of the variance in retention status above and beyond what is explained by traditional predictors, making a strong argument for the inclusion of these variables in retention studies.

Despite the evidence supporting the predictive value of psychosocial factors, a comprehensive instrument measuring the constructs outlined in Robbins et al. (2004) did not exist. Robbins et al. (2006) developed the Student Readiness Inventory (SRI) as an amelioration of this deficiency. This instrument measured ten different psychosocial scales and was used to determine which scales were the strongest predictors of college outcomes and their incremental validity over academic and background characteristics. A sample of 14,464 students from 48 different institutions completed the instrument, and hierarchical models were used to assess the predictive validity of the psychosocial factors above and beyond traditional academic, demographic, and high school characteristics. For the purposes of the present study, an overview of six of the SRI scales will be provided along with additional research related to each construct.

Academic discipline. Academic discipline, also referred to as academic engagement, describes the degree to which students value academics, identify themselves as conscientious, and the amount of effort they devote to academics (Astin, 1984; Robbins et al., 2006). Robbins et al. (2006) found that academic discipline had a significant, positive relationship with retention. Specifically, a one standard deviation increase on the academic discipline scale as measured by the SRI was associated with a 35% increase in the odds of being retained at four-year institutions and increased the accuracy of prediction when included in the model along with other

psychosocial factors. Other research which also explored the relationship between academic discipline and retention corroborates these findings.

J. Allen, Robbins, Casillas, and Oh (2008) conducted a study to examine the effects of motivation and social connectedness on college retention. The study specifically examined three different factors from the SRI, including academic discipline. The sample was drawn from a national dataset and consisted of 6,872 students from 23 different four-year institutions in the early 2000s. After including controls for pre-college academic performance, demographic characteristics, and institutional selectivity and enrollment size, a path model using hierarchical modeling was constructed that related the predictors to academic performance and retention. This study examined first year cumulative GPA and third year enrollment status (defined as retained, dropped, or transferred institutions) instead of first-to-second year retention. The results showed that academic discipline was a significant predictor of both first-year GPA and third-year enrollment status after controlling for the variables listed above. Specifically, it was positively related to GPA and negatively related to being retained versus dropping out. The researchers posited that the counterintuitive relationship between academic discipline and retention occurred because the effect was suppressed by first-year academic performance (i.e., after controlling for GPA, students with higher self-ratings were more likely to drop out). This hypothesis was supported by a positive indirect effect through first year performance.

Komarraju, Ramsey, and Rinella (2013) explored the relationship between ACT scores, high school GPA, academic discipline (as measured by the SRI), and college GPA. Specifically, they sought to determine the predictive power of these variables on college GPA with extra focus on academic discipline. Hierarchical regression results with a sample of 375 students showed that all three predictors were positively related to college GPA. The findings showed that

academic discipline accounted for an additional 2% of the variance in college GPA above and beyond ACT score and high school GPA, both of which accounted for 24% of the variance. In addition to exploring this relationship, the study also tested for a mediation effect. Results showed that academic discipline partially mediated the relationship between high school and college GPA, but it did not mediate the relationship between ACT score and college GPA.

Svanum and Bigatti (2009) examined the relationship between academic engagement and degree attainment, time to degree, and cumulative GPA in a single upper-level psychology course at a large, urban, commuter state university. The sample consisted of 225 students in the 2000-2001 school year. Using Pearson correlations and multiple regressions to examine the relationships, the study found that academic engagement was positively related to degree attainment and cumulative GPA. There was also a significant negative relationship between academic engagement and time to degree, indicating students who were more engaged in academics completed their degrees faster than less engaged students.

Kuh et al. (2008) conducted a study to determine the impact of engagement on retention. Their sample consisted of 6,193 first year students at 18 different colleges and universities between 2000 and 2003. Each of the students in the sample completed the National Survey of Student Engagement (NSSE), which measured a variety of factors related to engagement. This study used time spent studying, time spent in co-curricular activities, and engagement in effective educational practices as predictors in a logistic regression model that controlled for student background variables, previous academic performance, and high school engagement. The findings showed that the engagement scales, along with first year GPA and unmet financial need, explained an additional 25% of the variance in retention beyond demographics and prior achievement. Furthermore, two of the NSSE scales, hours spent in co-curricular activities and

engagement in effective educational practices, were positively related to student retention. The effect of academic engagement as measured and defined by this particular instrument, however, was fairly small.

Academic self-confidence. Academic self-confidence, also referred to as academic self-efficacy, describes the degree to which students believe they can perform well in academic settings (Bandura, 1997; Robbins et al., 2006). The concept of academic self-efficacy, and self-efficacy more broadly, is a fundamental component of Bean and Eaton's (2000) model of retention because they posit, as previously discussed, that "the factors affecting retention are ultimately individual and that the individual psychological processes form the foundation for retention decisions" (Bean & Eaton, 2002, p. 73). The Robbins et al. (2006) study found that academic self-confidence had a positive relationship with retention and college GPA. However, when academic self-confidence was included in a hierarchical regression model with other SRI factors and student background characteristics, it was not statistically significant. Other researchers have examined this construct and found a similar relationship, but other studies found results that were statistically significant.

Vuong, Brown-Welty, and Tracz (2010) conducted a study that examined the relationship between self-efficacy and academic success. Their study also examined a host of demographic characteristics. The study's sample consisted of 1,291 students from five diverse institutions in the California State University system who completed the College Self-Efficacy Inventory (CSEI). This instrument measured a variety of constructs related to self-efficacy including an index related to academic self-efficacy. Regression models were run on students' self-perceived likelihood of persistence (not whether or not they actually persisted) using the self-efficacy

measures. The results showed that greater academic self-efficacy was significantly related to a higher student self-perception of probable persistence.

Gore (2006) conducted a study that examined multiple measures of academic self-efficacy and their relationship with college outcomes, namely GPA and retention. The first analysis consisted of 629 first-year college students at a large, public university in the Midwest in the early 2000s who took the CSEI and items from the SRI related to academic self-confidence at various points in the school year. Hierarchical linear regression models were used to evaluate the degree to which these two measures and ACT composite score were able to predict college GPA. The results showed that CSEI scores were predictive of college GPA above and beyond ACT scores. However, there was no relationship for CSEI scores from the first administration at the beginning of the school year. This trend was also true when the SRI measure was examined. Hierarchical logistic regression models showed a similar trend, with end-of-semester CSEI scores adding to the prediction over ACT scores alone. The study also replicated the analysis with a different sample of 7,956 students from 25 different four-year universities from 2003 and 2004. In this analysis, only the academic self-confidence measure from the SRI was used. Results showed that academic self-confidence significantly added to the prediction of retention over ACT scores alone.

Komarraju and Nadler (2013) examined the relationship between self-efficacy as well as other psychosocial factors on college GPA. The researchers administered a survey to a sample of 257 undergraduate students selected from a psychology course. Correlations between the survey subscales and GPA revealed self-efficacy was positively related to academic performance. A hierarchical regression was also conducted to examine the incremental validity of the various measures. The results showed that self-efficacy was the strongest predictor of

GPA among all the survey scales. Mediation analysis revealed that this relationship was partly mediated by effort regulation.

Zajacova, Lynch, and Espenshade (2005) explored the effects of self-efficacy on academic outcomes at a City University of New York campus. The sample was comprised of 107 first-time freshmen who enrolled at the university in the spring of 1998. The majority of the sample was predominately female, Hispanic and White, and a couple years older than traditional college students. The researchers created a modified survey instrument that was designed to measure stress and self-efficacy, and factor analyses revealed two distinct indices related to academic self-efficacy that were combined with other self-efficacy indices to create a general self-efficacy factor. Structural equation models revealed that the general self-efficacy factor was positively related to GPA but was not related to retention. This finding is contradictory to other studies on academic self-confidence, suggesting that sample and analytical methods have an effect on the observed relationship of self-efficacy and postsecondary outcomes.

Commitment to college. Commitment to college, also referred to as educational or degree commitment, describes the degree to which students value a college degree and are committed to earning one (Robbins et al., 2006). It is an important construct to consider because students who are committed to completing a degree are more likely to be engaged in college both academically and socially (Tinto, 1993). The Robbins et al. (2006) study found that commitment to college had a positive relationship with retention and college GPA; a one standard deviation increase in commitment was related to a 31% increase in the odds a student would be retained. However, when commitment to college was included in a hierarchical regression model with other SRI factors and student background characteristics, it was not statistically significant. In a rigorous meta-analysis on precursors of college GPA, Richardson, Abraham, and Bond (2012)

found there was a positive relationship between degree commitment and postsecondary success. A somewhat contradictory result was found in the study that was conducted by J. Allen et al. (2008) which was described earlier. In that study, college commitment was found to not have a significant relationship with first year GPA but a significant, positive relationship was found with third-year retention after controlling for demographic, academic, and institutional factors. These findings are built on the work of other researchers who investigated the relationship between college commitment and postsecondary outcomes and have since been expounded upon further.

Cabrera, Nora, and Castañeda (1993) used structural equation modeling to test a combination of Tinto's (1993) retention model and an early conceptualization of Bean and Eaton's (2000) because, they argued, there was "considerable overlap between the two theoretical frameworks" (p. 124). One component of this merged model was the hypothesized relationship between institutional and goal commitment and a student's intention to persist. The longitudinal study was conducted at a large, urban institution in the South with 2,459 first time freshmen. The results showed that both institutional commitment and goal commitment, or commitment to college, had a positive direct effect on a student's intention to persist and consequently their persistence behavior.

Nieuwenhuis, Hooimeijer, and Meeus (2015) studied the relationship between educational commitment and educational attainment in the Netherlands. The sample consisted of 915 students with data collected over nine years from 2001 to 2010. The study used survival analysis to determine the association between degree of educational commitment (as measured by the Utrecht-Management of Identity Commitments Scale) and time to educational credentialing while controlling for an assortment of demographic variables. The findings

showed that students with higher levels of educational commitment earned educational qualifications faster than students with lower levels of commitment.

Emotional control. Emotional control, also referred to as resiliency, describes how students respond to strong feelings such as stress and anxiety, as well as how students manage emotions (Luthar & Cicchetti, 2000; Robbins et al., 2006). The Robbins et al. (2006) study found that emotional control was significantly related to first-to-second year retention and college GPA. Their model included a quadratic term for emotional control due to a curvilinear relationship between this predictor and GPA. The findings showed that emotional control had a significant inverted quadratic relationship with GPA, indicating that extremely high and low scores for emotional control were negatively related to academic performance in college. For retention outcomes, emotional control was related to persistence at two-year institutions; however, no significant effect was found at four year institutions. Richardson and colleagues' (2012) meta-analysis of psychosocial factors showed that anxiety, depression, and stress were negatively related to postsecondary success. Other researchers who have explored the relationship between retention and variables associated with emotional control further substantiate these findings.

Munt and Merydith (2012) utilized a sample of 216 undergraduate students at private, four-year technical college in New York to study personal characteristics that were related to academic achievement in college. A questionnaire that measured a variety of personality traits including anxiety and tough-mindedness was administered to the students. Multivariate analysis of variance and multiple regression were run to explore differences between academically successful and unsuccessful students on these measures. Results showed that tough-mindedness

and emotional stability were significantly related to retention. Students who scored lower on these metrics were less likely to be retained than their counterparts.

Daniels et al. (2009) studied the relationship between emotions and achievement. Specifically, they examined the associations between anxiety, feelings of helplessness, and performance in college. The sample for the study was comprised of 669 undergraduate students from 1997 and 2003 in Canada. A survey was administered to each student at two different points in their freshman year to measure the psychosocial variables of interest. Results of a structural equation modeling procedure indicated that anxiety and feelings of helplessness both were negatively related to academic performance in college.

Social activity. Social activity, also referred to as campus engagement, describes the degree to which students feel connected to and are involved with their college or university (Robbins et al., 2006). The Robbins et al. (2006) study found that social activity was significantly related to both first-to-second year retention and college GPA. Similar to emotional control, their model included a quadratic term for social activity due to a curvilinear relationship between this predictor and GPA. The findings showed that social activity had a significant inverted quadratic relationship with GPA, indicating that extremely high and low scores for social activity were negatively related to academic performance in college. For retention, social activity was negatively related to persistence after controlling for institutional, demographic, and academic background variables. Specifically, a one standard deviation increase in social activity was associated with an 11% decrease in the odds a student would be retained. The relationship found in this study is supported by other research on social activity. Studies have found that students who are more willing to engage in activities that will enhance their social integration into the university are more likely to be successful in college—especially for minority groups

and students from underprivileged backgrounds (e.g., Kuh et al., 2008; Pascarella, Pierson, Wolniak, & Terenzini, 2004).

Social connection. Social connection, also referred to as social comfort, describes the degree to which students are comfortable in meeting and interacting with others (Robbins et al., 2006). Robbins and colleagues (2006) found that social connection was positively related to academic performance and persistence. However, when student background and institutional characteristics were controlled for, this relationship was only significant for predicting retention at four-year universities. Specifically, a one standard deviation increase in social connection was associated with a 13% increase in the odds a student was retained to the second year. In the study conducted by J. Allen et al. (2008) described earlier, the relationship between social connection and first-year performance as well as third-year retention was also examined. Controls for academic achievement, gender, ethnicity, socioeconomic status, and institutional characteristics were included in the hierarchical path analysis model. The results of the study showed that social connection had a small, negative relationship with first-year GPA but a small, positive relationship with continued enrollment in the third year versus dropping out. The researchers conjectured the other predictors suppressed the effect of social connectedness on first-year GPA but that it did have a small direct effect on long-term enrollment. Other studies have shown that social connection is positively associated with social integration which, according to theoretical models, is a condition for persistence (e.g., Braxton et al., 2014).

Demographic and Other Factors

In addition to the intensive interest in academic and psychosocial factors and their relationships with college outcomes, a variety of other factors have also been explored

throughout the literature. A host of demographic characteristics have been studied as well as other constructs that address various components of the theories related to retention.

Demographic characteristics. Some of the most common demographic variables that have been included in studies on retention are gender, ethnicity, socioeconomic status, and first-generation status. Each of these variables have demonstrated associations with retention behaviors. Research demonstrates that these variables oftentimes moderate relationships between other variables, making them an important factor that should be included in educational studies.

Gender. Examination of general trends throughout the research shows that females tend to outperform males on college outcome measures including academic performance and retention (e.g., J. Allen et al., 2008; DeBerard, Spielmans, & Julka, 2004; Kuh et al., 2008; Mattern & Patterson, 2010). However, some studies have found no significant gender effect once other controls are entered into the model or that gender mediates and/or moderates relationships between variables, including interactions with other demographic characteristics such as ethnicity (e.g., Bridgeman, Burton, & Pollack, 2008; St. John, Hu, Simmons, & Musoba, 2001). This suggests that gender is an important characteristic to consider not only as a direct effect, but also as a possible moderator in studies on retention.

Ethnicity. Similar to gender, ethnicity is generally a significant factor in studies on college outcomes. The literature shows that, in general, White and Asian students are more likely than students of color to succeed in college and be retained (e.g., Arbona & Nora, 2007; Kao & Thompson, 2003; Porchea, Allen, Robbins, & Phelps, 2010; Reason, 2003). The research has also shown that ethnicity interacts with other demographic characteristics and its effect is sometimes attenuated when a variety of controls are included in the model (e.g., D. Allen, 1999;

Attewell, Heil, & Reisel, 2010). Accordingly, it is important to account for ethnicity in higher education research, as well as its interaction with other variables, in order to fully explore how it is related to both other explanatory variables and outcomes of interest.

Socioeconomic status. Socioeconomic status (SES) is a consistently important factor in educational research. Many studies found a positive relationship between SES and college success (e.g., J. Allen et al., 2008; Robbins et al., 2004). Additionally, SES has been shown to covary with other demographic characteristics and moderate associations with higher education outcomes (e.g., Paulsen & St. John, 2002). This suggests that SES is an important variable to include and that its effect is not straightforward—it should be examined in combination with other variables. SES is measured in a variety of ways throughout the literature, typically using family income and/or parental education level (Liu et al., 2004). The definitions of first generation students vary, but the most common definition is students whose parents did not obtain a bachelor's degree. The research has consistently shown that first-generation students are less likely to be successful in college, even after controlling for factors that are related to first-generation status (e.g., Choy, 2001; Pascarella, Pierson, Wolniak, & Terenzini, 2004). Bearing this in mind, first-generation status is an important variable to account for in higher education research. It especially should be examined alongside demographic and psychosocial factors that previous research has shown to be related to first-generation students.

As demonstrated throughout the literature, demographic factors cannot be ignored in educational research. Not only should they be included as main effects, but they should also be examined as potential moderators of other variables of interest in models on postsecondary outcomes including retention. Inclusion of demographic characteristics in research on retention will provide a more realistic and comprehensive understanding.

Institutional fit. Recent studies have examined factors beyond the academic, psychosocial, and demographic factors discussed above which address a variety of themes within Bean's (2005) framework. There are a plethora of variables within this realm including high school and background characteristics. An area of specific interest for this study is institutional fit. The notion of institutional fit directly aligned with Tinto's (1993) concept of integration. If a student feels they fit in academically and socially at their institution, they are more likely to persist. Research has shown that many activities, interactions, and events influence the degree to which students feel integrated (Burrus et al., 2013). The following studies address a different perspective of institutional fit, namely academic and interest fit.

Mattern, Shaw, and Kobrin (2010) examined the relationship between retention and college fit. Specifically, they explored the degree to which the academic fit between a student and their institution is related to academic success and first-to-second year retention. The study also explored the moderation of this relationship by gender, ethnicity, and language spoken. The sample consisted of a national sample of 143,624 college freshmen in the 2006 cohort at 106 different institutions across the United States. Predictor variables used in the study included official SAT scores and self-reported demographic and background information. Criterion variables were first-year college GPA and retention status reported by the students' institutions. Additionally, academic fit was assessed by determining the magnitude of difference between a student's SAT score and the mean SAT score of their institution. A logistic regression was performed to assess the relationship between retention and student SAT score, institutional average SAT score, academic fit, gender, ethnicity, and income. Interactions between academic fit and the demographic characteristics were also included in the model. The results showed student SAT score and institutional average SAT score were significant predictors while

academic fit was not; however, the effect sizes for these variables were small and consequently did not support a relationship between retention and academic fit overall or for demographic subgroups. Nonetheless, the results did show that students with higher test scores are more likely to be retained, but the strength of this relationship was marginal.

Mattern, Wyatt, and Shaw (2013) conducted a study examining the relationship between the distance between a student's home and institution and whether or not they transferred from their initial institution. A logistic regression was used to analyze this relationship with controls for SAT score, ethnicity, gender, and first-generation status utilizing a sample of 503,887 students. Student records regarding institution attended were obtained from the National Student Clearinghouse, and distance from home was calculated using the student's self-reported home zip code and the university zip code. Due to the positive skew of distance from home, a logarithmic transformation was applied to the variable. The results of the analysis showed there was a significant positive relationship between distance from home and the likelihood a student would transfer. The effect size, however, was small. The study also found that males, first-generation students, Asian and Hispanic students, and students with higher SAT scores were less likely to transfer than their counterparts. While distance from home is not a direct measure of institutional fit, the findings suggest that it is related to how well a student acclimates to their university.

Summary

The research has shown that academic, psychosocial, demographic, and other background characteristics are significantly related to first-to-second year retention. Generally, students who perform at higher levels in high school are more likely to be successful in college. Research has also shown, however, that this relationship is affected by interrelationships between metrics used

to measure academic characteristics as well as actual performance once students matriculate to college. Psychosocial factors meaningfully add to the prediction of retention over academic factors by explaining variance that cannot be accounted for by academic factors alone. The relationship between retention and academic or psychosocial factors oftentimes is moderated by demographic traits, and these characteristics are important to consider when examining retention. Finally, a core tenet of theoretical frameworks for persistence behaviors is the concept of integration. It is therefore important to also consider both academic and social fit between students and their institutions when exploring retention.

As demonstrated, the understanding of retention has expanded over the past several decades. Early studies on factors related to retention serve as an important foundation on which to build future research, but it is important that further inquiry incorporates the issues that are prevalent in modern retention conversations. The demographic profile of students who attend college is more diverse than ever before. Consequently, it is imperative that studies on retention incorporate a plethora of variables that are able to account for the diversity of this population. Studies should include traditional demographic characteristics such as race, gender, age, and socioeconomic status and examine the interactions between these variables to advance the understanding of retention (Pascarella & Terenzini, 1998; Reason, 2003). Previous research has shown the nature of the relationship relationships between high school GPA and standardized test scores with retention have changed over time, but it remains clearly and consistently significant. For this reason, all studies studying retention need to also include these measures (Reason, 2003).

Many studies are constrained by limitations related to data. For example, several studies rely on student self-reported GPA and/or standardized test scores. While student self-reports of

these measures are reasonably reliable for high achieving students, they are less accurate for students with low grades and performance (Kuncel, Credé, & Thomas, 2005). This suggests that studies should not rely on student self-reported data whenever possible. Another data limitation evident throughout the literature was the inability to control for a rich set of variables, which is critical for a meaningful understanding of the complex issue of retention (Pascarella & Terenzini, 1998). Furthermore, many studies are limited in their sample size and selection due to data limitations. This consequently affects the power of the statistical models and the generalizability of the findings. Research that is not hindered by these limitations are frequently conducted at the national level in multi-institutional studies. It has been noted by multiple researchers and theorists that the general does not necessarily apply to the specific, and that these types of rigorous studies need to be conducted at individual institutions to best understand retention in specific contexts. This study contributes to the body of knowledge by addressing each of these limitations using a large dataset with an abundant set of variables at a large, public research university in the Southeast.

CHAPTER THREE
METHODOLOGY

In this chapter, the research methodology for the study is outlined. The design of the research is discussed, followed by information about the data sources that were utilized including the reliability and validity of the measures. The sampling techniques will be described as well as the analytic and statistical procedures that were used to answer the research questions.

Research Design

This study utilized a nonexperimental, correlational research design to answer the research questions. This specific design was chosen because the research questions were interested in the association between variables and the demands of an experimental design, namely control of the independent variables, could not be utilized (Creswell, 2015; R. B. Johnson & Christensen, 2014). More specifically, the research questions sought to understand the academic, psychosocial, demographic, and other variables that were related to and predictive of the criterion variable, first-to-second year retention. For this reason, the specific type of correlational design that was used in this study was a prediction research design (Creswell, 2015).

Data and Sample

In order to have the most comprehensive information possible, four different data sources were utilized for this study:

- **Institutional record.** The institutional record of the university in the study provided postsecondary records of academic performance and enrollment patterns of the students as well as traditional high school measures used to make admissions decisions. The primary benefit of the institutional record was the accurate measures it provided regarding high school and college academic performance because all variables were

recorded from high school transcripts, standardized test records, and the university's enrollment tracking system. Additionally, this database contained information reported from the Free Application for Federal Student Aid (FAFSA) which provides financial information about students including annual household income and financial aid.

- **ACT student record.** The student record provided by ACT provided a detailed portrait of students' pre-college interests and performance. The ACT student record is comprised of three different components. The first was student performance on the assessment itself, including composite and subtest scores. The second was information from the Interest Inventory and Student Profile Section which provides student background information, academic and extracurricular interests, and educational and career plans. This information is self-reported by students at the time they register to take the ACT. The third component is the high school course/grade information section which is also student self-reported at the time of test registration. This section includes information regarding the courses a student has taken or plans to take as well as the grades earned in completed courses. This study primarily utilized information from the Student Profile Section.
- **Student Strengths Inventory.** The Student Strengths Inventory (SSI) is an instrument developed by Campus Labs (2012a) that measures a variety of psychosocial factors which have been shown to be related to collegiate success. It was administered to incoming students at their summer orientation session before the semester they enrolled at the institution in this study. The instrument consists of 48 items measured on a six point Likert-type scale with options for strongly disagree, disagree, somewhat disagree, somewhat agree, agree, and strongly disagree. The items were grouped into six

constructs consisting of eight items each by conducting an exploratory and confirmatory factor analysis (Campus Labs, 2012a). All six factors had reliabilities that were .8 or higher, indicating good reliability (Nardi, 2006). The validity of the SSI was assessed by the developers in three different ways in order to assure confidence in the credibility of the measures (Campus Labs, 2012a). First, items were shared with experts in the field to provide evidence for face validity. Second, predictive validity was assessed by regressing retention and academic success on the six constructs plus standardized test scores. Finally, discriminant validity was assessed by assessing the dissimilarity of the constructs with constructs from other instruments that they theoretically should be dissimilar from (e.g., dangerous drinking behaviors).

- **National Center for Education Statistics**. Variables regarding high school characteristics were drawn from the National Center for Education Statistics (NCES) Common Core of Data (CCD) and Private School Universe Survey (PSS) databases. Public school characteristics were drawn from the CCD for the 2013-2014 school year and private school characteristics were drawn from the PSS for the 2011-2012 school year, as these were the most recent years for which data was available. Information for public schools was complete due to the nature of data reporting for public schools to the NCES, but private school data was incomplete or missing from the PSS for schools who did not complete or only partially completed the survey (Burns, Wang, & Henning, 2011). Private schools with missing data were contacted via phone call and/or email to find this information.

Data Collection

The data collection procedures for this study occurred in three steps. First, a data request was made through the registrar's office at the university for students who fit the criteria for the sampling frame. All variables of interest were requested along with identifying information for each student for matching purposes. Second, the data from the SSI was shared from the university's Division of Student Affairs. These two datasets were then merged using student codes from the registrar data. Unmatched observations were cross-referenced, and it was determined that these students had not completed the SSI so these cases were removed from the dataset. Third, a data-sharing request with ACT was filed and subsequently approved. A file with student identifying information was sent to ACT who then appended the ACT student record and returned the file through a secure connection with all identifying information deleted from the dataset. Finally, variables from the NCES datasets were merged into the final dataset by matching on high school name, city, and state. Unmatched observations were due to naming differences between the different databases, so data for unmatched high schools were matched by hand. The original files for each dataset were password protected and saved on a secure server. Approval for this study was obtained from the Institutional Review Board (see Appendix A).

Variables and Factors

There were a variety of variables that were considered in the study. These variables fell under seven major categories: academic, psychosocial, noncognitive, financial, demographic, and high school characteristics as well as first semester performance variables. This section describes each of these variables in terms of the construct they measure and the type of variable (i.e., continuous or categorical). A comprehensive summary chart of the variables used in the

study is presented in Appendix B along with the data source(s) the variable was drawn from and a description of the possible values for each.

Academic variables. Four variables were academic variables taken from the institutional and ACT student records. High school GPA was a continuous variable that could range from 0.00 to 4.00. This variable represented a composite measure of a student's unweighted academic performance in their high school coursework. ACT composite score was a continuous variable that could range from one to 36 that measured a student's academic readiness for college. The third academic variable was an indicator variable for whether a student self-described their high school curriculum as college preparatory, meaning it consisted of honors or other advanced coursework. Finally, the last academic variable was the number of hours a student registered to take in their first semester of college.

Psychosocial variables. Six variables were the psychosocial variables taken from the SSI. Table 3.1 presents the six constructs, conceptual definitions for each, sample items constituting each factor, and the reliability of the scale as measured by Cronbach's alpha.

Noncognitive variables. Seven variables were noncognitive factors that measured various aspects of students' backgrounds and expectations. Three variables were constructed indicator variables from the ACT Student Profile Section (SPS). The first of these variables indicated whether a student participated in extracurricular activities in high school. This variable was constructed using items from the SPS that asked a student to indicate whether or not they participated in any of 13 different extracurricular activities. If a student said yes to any of the 13 activities, they were coded as having participated in extracurricular activities in high school. If they said no to all 13 activities, they were coded as not having participated in any extracurricular activities. The second noncognitive variable indicated whether or not students had out-of-class

Table 3.1
Summary of Student Strengths Inventory

Factor	Definition	Sample Item	Reliability
Academic Engagement	The value an individual places on academics and attentiveness to school work	I turn in my homework on time.	.80
Academic Self-Efficacy	An individual's confidence in his or her ability to achieve academically and succeed in college	I will excel in my chosen major.	.86
Campus Engagement	An individual's desire to be involved in campus activities and their attachment to the university	Being active in extra-curricular activities in college is important to me.	.88
Educational Commitment	An individual's dedication to college and the value placed upon a college degree	I see value in completing a college education.	.89
Resiliency	An individual's approach to challenging situations and stressful events	I manage stress well.	.81
Social Comfort	An individual's comfort in social situations and ability to communicate with others	I am comfortable in groups.	.83

Note. Information retrieved from Campus Labs (2012a, 2012b).

accomplishments in high school such as awards or recognitions. This was similarly constructed from a list of 21 different accomplishments from the SPS. The third noncognitive variable was a variable indicating whether or not a student had intentions of being involved in extracurricular activities in college. It was constructed from a list of 14 different items on the SPS using the same method as the previously described noncognitive variables.

Three additional noncognitive variables were drawn from the SSI. The first was a categorical variable indicating whether a student planned to have a job in college. The options for this item were yes, no, and unsure. The second was an ordinal variable representing an item

that asked students to indicate how many hours per week they anticipated studying. The options for this item were less than 10 hours, 11 to 20 hours, 21 to 30 hours, and more than 30 hours. The third was a variable that represented the highest level of education a student expected to attain. The options for the original item were associate's/technical degree, bachelor's degree, master's degree, and doctorate/professional degree. If a student skipped this item on the SSI, their response on the same item from the SPS of the ACT was used. The original categories were collapsed and the variable was recoded to a dichotomous variable with levels for bachelor's degree or less and graduate or professional degree.

The final noncognitive variable was institutional fit. This was a constructed variable from four different items on the SPS of the ACT. These items asked students to indicate their preference for various characteristics influencing their college choice. The first item asked students to indicate the type of institution they preferred (e.g., public two-year, private four-year), the second item asked students to indicate the gender composition they preferred (e.g., coeducational or all female), the third item asked students to indicate up to two states in which they preferred to attend college, and the fourth item asked students to indicate the size of the student body they preferred from one of five different ranges. If a student indicated a preference that matched the characteristic of the institution in the study, they were classified as having a match for institutional fit on that characteristic. The overall institutional fit variable was a continuous variable that represented the number of institutional characteristics that the student's preference matched; thus, the scores could range from 0 (completely mismatched) to 4 (completely matched).

Demographic characteristics. Three different demographic variables were included as possible predictors. The first two were gender (coded as male or female) and ethnicity (coded as

African American, Asian, Hispanic, White, or Other). The third demographic variable was first generation status. This was constructed by taking the highest parental education level between either parent listed on the ACT SPS, institutional record via admission application, and/or the demographic portion of the SSI. If a student indicated that neither of their parents had attained at least a bachelor's degree then they were coded as a first-generation student; conversely, if one of their parents earned a bachelor's degree or higher, they were coded as a non-first generation student.

Financial characteristics. Three variables were related to financial characteristics. One was a variable indicating the total dollar amount for all earned scholarships. This included scholarships awarded by the institution, state funded scholarships, and external scholarship awards. The second variable was the total amount financial aid received in the form of federal loans and/or grants. The final financial variable was the student's family income as reported via tax forms on the FAFSA. This variable was highly right skewed, so a log base ten transformation was applied.

High school characteristics. Five different variables drawn from the NCES databases were related to high school characteristics and considered as possible predictors. The first was the type of high school (i.e., public or private). Two variables were related to the demographic composition of the school. These were the percentage of minority students and the school's total enrollment. The fourth variable was an indicator for whether the high school was located in the same state as the institution in the study or in another state, thereby identifying whether the student was an in-state or out-of-state resident. Finally, the distance of the high school from the institution was included. This was calculated by finding the distance between the midpoint of the high school's zip code and the university. The distribution of distances was highly right

skewed, so a log base ten transformation was applied. Since there were students with a distance of zero and logarithmic transformations can only be applied to positive numbers, a constant of 0.01 was added before applying the transformation (Mattern et al., 2013).

First semester performance and activity. Three different variables related to first semester academic performance and activity were considered in the modeling. Two of these were related to academic performance in the first fall semester. The first variable was the magnitude of the difference between high school GPA and first semester collegiate GPA. This was called the GPA differential. The second variable was the number of credit hours a student earned in their first semester of college. The final college performance and activity variable was participation in a fraternity or sorority, also referred to as participating in Greek life. This variable was dichotomous, indicating whether or not a student was in a fraternity/sorority.

Data Screening

Before conducting any analyses, the dataset was screened for accuracy. The data were checked for duplicate observations to ensure that each student was only represented once in the final dataset. The dataset was also thoroughly examined to ensure the accuracy of the values for each variable. This was accomplished by analyzing descriptive statistics for each variable, checking for values that were out of range and/or unreasonable, and ensuring the means and standard deviations for each variable were plausible. The overall distribution for each continuous variable was checked, and a log base ten transformation was applied to any variables that were highly skewed. Furthermore, the variance inflation factors were examined to assess potential issues with multicollinearity. All were lower than 2.5, so there were not any issues with collinear predictors.

Sample

The population for this study was all first time college freshmen at a large, predominantly White research institution in the southeastern region of the United States. The institution is the state's flagship institution, and it accordingly his the highest admissions standards of all the public colleges and universities in the state. Moreover, the institution is located in a state that has a state-funded scholarship program, has experienced higher education budget cuts for several years, and has a large private school culture largely due to the perception of public school quality. Students selected to be part of the sample had to have been enrolled at the institution for at least one full semester so that the outcome, first-to-second year retention, could be measured. Since second year retention typically is measured from first fall semester to second fall semester, students who enrolled in the spring semester were excluded. Furthermore, students had to enroll at the institution no earlier than the first year the SSI was administered in 2013. This limited the sampling frame to first time freshmen who enrolled at this particular institution in either the summer or fall semesters between the 2013-2014 and 2015-2016 school years and remained enrolled for the duration of the first fall semester. In total, the sampling frame consisted of 16,590 students. Students within this sampling frame were selected for inclusion in the final sample for this study if they had taken the both the ACT and the SSI as well as filled out the FAFSA. Any students who did not meet these criteria were removed. Students who did not fully complete the SSI caused validity concerns for the six factor scores so they were removed from the sample. Furthermore, students that were home schooled, attended a virtual school, attended a foreign high school, or attended a high school for which characteristics could not be ascertained were also removed from the sample. This brought the final sample size to 12,342

students. Table 3.2 provides a comparison of the sample and population on key demographic information.

Table 3.2
Demographic Comparison of the Sample and Population

		Sample		Population	
Characteristic	Category	n	%	n	%
Gender	Male	5134	42%	7437	44.8%
	Female	7208	58%	9153	55.2%
Ethnicity	White	9098	74%	12108	73.0%
	African American	1605	13%	2149	13.0%
	Hispanic	712	6%	752	4.5%
	Asian	561	5%	1064	6.4%
	Other	366	3%	517	3.1%
Residency	In-State	10908	88%	13415	80.9%
	Out-of-State	1434	12%	3175	19.1%
Total		12342		16590	

Missing Data

To preserve the integrity of the dataset and the generalizability of the findings, the data were examined for prevalence and pattern of missing values. The pattern of missing data was assessed to determine if the pattern of missing data was arbitrary (Tabachnick & Fidell, 2013). In total, 7,152 cases (57.9%) were missing data for at least one variable. Five percent of the sample was missing data for either five or six variables; all remaining cases were missing data for four or less variables. Table 3.3 presents the number and proportion of missing values for the variables that had missing data. The pattern of missingness suggested the data were at least missing at random, so multiple imputation was used to impute values for all missing data. Since the variables in this study were a combination of continuous and categorical variables, the fully conditional specification method for multiple imputation was utilized. This allowed for more

Table 3.3
Summary of Variables with Missing Data

Variable	n	Percentage
Had Out-of-Class Accomplishments in High School	4406	35.7%
Intend to Participate in Extracurricular Activities in College	3031	24.6%
College Preparatory Curriculum	2745	22.2%
Participated in Extracurricular Activities in High School	1568	12.7%
Institutional Fit	827	6.7%
Expected Educational Attainment	76	0.6%
Hours Anticipated Studying per Week	30	0.2%
Plan to Have a Job in College	30	0.2%

flexibility than traditional multivariate normal imputation because the conditional distribution for each variable was considered and, consequently, the procedure yielded more accurate imputations (Berglund, 2015; Kropko, Goodrich, Gelman, & Hill, 2014). All independent variables of interest were included as predictors in the imputation model. The discriminant function method was used to impute values for two multi-level categorical variables with missing data. The logistic function method was used to impute values for all dichotomous variables that had missing data. The regression method was used to impute values for the continuous variable institutional fit. In total, five different imputed datasets were created and used in subsequent analyses. Results were pooled from the analysis of all five datasets to account for the uncertainty of imputed values, thereby generating valid statistical inferences.

Splitting the Sample

Since predictive validity of the models was of interest in the study, the sample was randomly split into a training set and a test set with approximately half of the overall sample in each subset. The demographic composition of the training set differed from the population on key measureable characteristics. For example, in-state students were overrepresented in the training set subsample compared to the institutional population. Consequently, weights were

applied to the training set to account for the overrepresentation of certain student demographic groups in the sample. These weights were obtained by dividing the proportion of students in the population by the proportion of students in the training set using a combination of three variables: gender, ethnicity, and residency status. Table 3.4 presents the number and proportion of students in the training set compared to the overall population of the institution for these three variables.

Table 3.4
Descriptive Statistics for Training Set Subsample Compared to the Population

		Training Set			Population	
Characteristic	Category	n	%	Weighted %	n	%
Gender	Male	2638	42.4%	44.7%	7437	44.8%
	Female	3584	57.6%	55.3%	9153	55.2%
Ethnicity	White	4594	73.8%	73.0%	12108	73.0%
	Black	795	12.8%	13.0%	2149	13.0%
	Asian	279	4.5%	4.6%	752	4.5%
	Hispanic	373	6.0%	6.4%	1064	6.4%
	Other	181	2.9%	3.1%	517	3.1%
Residency	In State	5489	88.2%	80.9%	13415	80.9%
	Out of State	733	11.8%	19.2%	3175	19.1%
Total		6222			16590	

Note. Some percentages do not add to 100 due to rounding.

Hierarchical Generalized Linear Models

The first three research questions were interested in the relationships between retention and a host of predictors. The criterion variable, first year retention status, was dichotomous so a logistic regression model would have been appropriate for answering these questions; however, students were naturally clustered within high schools. Thus, the chosen analysis method for answering these questions was hierarchical generalized linear models (HGLM). This method

was chosen over single-level logistic regression because ignoring the nested structure would have inflated the Type I error rate due to the failure to meet the assumption of independence (Hox, 2010; Tabachnick & Fidell, 2013). All models were estimated using Laplace's method to approximate the maximum likelihood. This approach was taken because Laplace estimates are relatively less biased compared to pseudo-likelihood estimates and more computationally effective than a quadrature approach (Hox, 2010; Raudenbush, Yang, & Yosef, 2000). Since the outcome was dichotomous, a binary distribution and logit link function were applied. The first level of the model consisted of the individual students, and the second level of the model consisted of the high school each student attended.

Main Effects Models

After all preliminary steps were completed, the HGLM was conducted by building a series of models. The first model constructed was the unconditional model with random intercepts. This was conducted for two reasons. First, it allowed for a test of whether retention rates varied across high schools. Second, it allowed for the calculation of the intraclass correlation (ρ) to determine the degree to which the errors were correlated in order to better understand the grouping structure of the data (Ene, Leighton, Blue, & Bell, 2015; Hox, 2010). The equation for used calculating the intraclass correlation was

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2},$$

where $\sigma_{u_0}^2$ represents the variance of the level-two errors and σ_e^2 represents the variance of the level-one errors (Hox, 2010). An assumption of HGLMs with dichotomous outcomes, however, is that there is no error at level-one. Snijders and Bosker (1999) (as cited in O'Connell, Goldstein, Rogers, & Peng, 2008) proposed a modification that assumes the outcome is a dichotomization of an unknown latent continuous variable with a level-one residual that follows

the logistic distribution with a mean of zero and variance of 3.29. Applying this modification, the equation that was used to calculate the intraclass correlation was

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + 3.29}.$$

The initial model with random intercept was followed by models with predictor variables added at the individual level and predictors added at the high school level. All variables of interest were run in the HGLM as fixed effects for each imputed dataset separately. Variables that were significant in any of the runs were retained, while variables that were not significant in any of the five runs were removed. The HGLM model was rerun on all five imputed datasets with only the significant predictors, and the results were combined across the imputed datasets to draw valid statistical inferences. In order to have a parsimonious model, insignificant variables in the combined model were removed. The HGLM was run again with the remaining predictors to arrive at the final, pre-college fixed effects model.

The parameter estimates and odds ratios were examined to ascertain the nature of the relationship between each significant predictor and retention status. Predicted probabilities were calculated for each variable as a means of understanding effect size. Predicted probabilities are calculated using the equation

$$p = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}},$$

where η_{ij} represents the linear predictor from the regression equation for the *i*th observation in the *j*th high school. The linear predictor was calculated by holding all other variables in the model constant at their mean. For categorical predictors, predicted probabilities were compared across the different levels of the variable. For continuous predictors, the predicted probabilities

for all possible values were calculated. These probabilities were then graphed for a visual representation of the effect size.

The final model that was developed from the first research question served as the baseline model for answering the second research question. Variables affiliated with first semester performance and activity were added to the pre-college main effects model to determine the degree to which they affected the significance of the pre-college variables. After examining the relationship and changes in significance, all insignificant variables were removed from the model to create the final post-first semester main effects model. Odds ratios and predicted probabilities were then examined to glean more information about changes in relationships and effect sizes.

Interaction Models

After the main effects for the pre-college and post-first semester models were determined, interaction terms between main effects and demographic variables were entered into each model and tested for significance. Specifically, interactions between all the main effects in each model and gender, ethnicity, and first generation status were examined. Interaction terms that were not significant were removed from the models. Any lower order effects that were part of a significant interaction effect were retained in the model, even if they were not significant. The resulting models were considered to be the final pre-college and post-first semester interaction models. Similar to the main effects models, odds ratios were computed to better understand effect sizes and the nature of the moderation. Interaction effects were examined through plots of the odds ratios to fully understand the nature of the moderated relationships.

Classification Trees

The fourth research question was interested in identifying the most important variables in predicting student retention. For this reason, classification tree based models were constructed in

addition to the four HGLMs. As noted by Hastie, Tibshirani, and Friedman (2009), a major limitation to classification trees is their instability. This occurs since estimation of a single decision tree is highly variable due to the hierarchical nature of splitting the tree and the propagated effects of errors down the tree. In order to allow for a reduction in the variance that occurs from a single decision tree, two different ensemble methods were utilized instead of a constructing one classification tree. Specifically, random forest and gradient tree-boosting were conducted. Random forests are an enhanced ensemble tree-based method in which multiple decision trees are constructed through bootstrapped training samples and a random sample of predictors considered at each split (James, Witten, Hastie, & Tibshirani, 2013). This prevents any single predictor from dominating all the trees, thereby decorrelating all the trees in the random forest. Boosting is another ensemble tree method that combines a large number of trees that are built sequentially using the residuals from the prior tree instead of the outcome variable itself, thereby yielding lower test error rates (Hastie et al., 2009; James et al., 2013).

The training set that was used to construct the HGLMs was used to build the two ensemble tree models. All original pre-college variables were considered as possible predictors. The random forest model was built using 1,000 trees with five randomly selected variables considered at each split to reduce test error and out-of-bag error (James et al., 2013). For the gradient tree boosting method, 1,000 possible trees were fit using a maximum depth of two variable interactions and a learning rate of 0.05. Half the observations were randomly selected to propose the next tree in order to sufficiently introduce randomness into the model. Finally, 10 cross-validation folds were performed to determine the optimal number of trees. For each model, the relative importance of the variables in the model was examined to understand the most influential factors. Partial dependence plots were developed for variables that warranted

further investigation. These plots describe the effects of each plotted variable on retention after accounting for all other variables in the model. This process was then repeated to develop the post-first semester tree-based models using the original variables plus earned credit hours, GPA differential, and participation in Greek life. Similar to the pre-college model, partial dependence plots were created for the most important variables to ascertain the nature of the relationship between these variables and retention.

Predictive Validity

The fifth research question was interested in assessing the performance of the models described above and their validity in predicting student retention. In total, the predictive validity of eight different models was examined. Four of the models were the pre-college main effects HGLM, interaction HGLM, random forest, and gradient tree boosting. The remaining four models were the post-first semester main effects HGLM, interaction HGLM, random forest, and gradient tree boosting. The equations for the HGLMs were used to calculate the predicted probabilities of retention for each observation in the test set. Students with probabilities of at least .5 were predicted to be retained while those with probabilities less than .5 were predicted to not be retained. The random forest algorithms were applied to the test set to predict whether students would be retained or not for both the pre-college and post-first semester models. Similarly, the algorithms produced by the optimal number of trees in the gradient tree boosting was used to calculate predicted probabilities of retention for each observation in the test set. Students with probabilities of at least .5 were predicted to be retained while those with probabilities less than .5 were predicted to not be retained. These predictions were then compared to the observed retention outcome in a confusion matrix. For each model, the accuracy of prediction, sensitivity, specificity, positive predictive value, and negative predictive

value were calculated. Sensitivity is the probability of correctly identifying a student who is retained, while specificity is the probability of correctly identifying a student who was not retained. The positive predictive value is the probability a student who was predicted as being retained was actually retained, and the negative predictive value is the probability a student who was predicted to not be retained was not. All five measures were compared across the models to assess predictive validity and model strength.

Summary

This study utilized a nonexperimental, correlational research design at a large, public predominantly White research institution in the Southeast. The sample consisted of first-year college freshmen at the institution in the 2013 through 2015 cohorts. Data was collected for each student from the institutional record, the ACT student record, NCES data, and the Student Strengths Inventory which measured six psychosocial constructs related to postsecondary success. Pre-college and post-first semester variables were used to construct HGLMs and ensemble tree-based models on a test subset of the sample. Relationships and variable importance were examined for each model. Finally, the various models were used to predict whether each student in the test subset would be retained. These predictions were compared to the actual outcomes to assess and compare model performance.

CHAPTER FOUR
RESULTS AND FINDINGS

In this chapter, the results of the study are presented. The chapter begins with a description of the sample using descriptive statistics. The findings related to the research questions for the study are then detailed. This first includes the results of the main effects and interaction effects models of both the pre-college and post-first semester HGLMs. Next, the results from the ensemble classification tree-based methods are presented. The chapter concludes by comparing the performance of the various models and assessing their predictive validity.

Summary of Sample

The sample for this study consisted of 12,342 first-time college freshmen at a large, public four-year institution in the southeastern region of the United States. All students in the sample enrolled in either the summer or fall semester between 2013 and 2015, completed the FAFSA, took both the ACT and the SSI, and attended either a public or private high school. Additional details regarding the sample selection are detailed in Chapter 3. The sample was randomly split into a training set and a test so that approximately half of the sample was in each set. The same training set was used to build each statistical model and the same test set was used to assess the predictive accuracy. The training set consisted of 6,222 students, and the test set was comprised of the remaining 6,120 students. The training set differed from the population on the key demographics of gender, ethnicity, and residency status, so sample weights were applied for the analyses. Table 4.1 presents the weighted and unweighted descriptive statistics for continuous variables before they were transformed (distance from home and family income) and rescaled (scholarships and financial aid). Table 4.2 presents weighted and unweighted descriptive statistics about the training set sample for categorical variables.

Table 4.1
Unweighted and Weighted Descriptive Statistics for the Training Set

	Unweighted		Weighted	
Variable	Mean	SD	Mean	SD
Academic				
High School GPA	3.47	0.37	3.46	0.37
ACT Score	25.7	3.5	25.7	3.6
Registered Credit Hours	14.9	1.3	14.9	1.3
Psychosocial				
Academic Self-Efficacy	71	23	71	23
Academic Engagement	63	26	62	26
Educational Commitment	61	27	60	27
Resiliency	56	29	56	28
Campus Engagement	67	24	68	24
Social Comfort	55	29	55	29
Noncognitive				
Institutional Fit	3.10	0.78	3.07	0.80
Financial				
Scholarships	7652	4192	7753	4999
Financial Aid	6261	8083	6922	9001
Family Income	149043	184682	149682	181823
High School				
Distance	123	247	164	310
Enrollment	1190	606	1237	660
Percent of Minority Students	27.1	22.7	27.7	22.8
Post-First Semester				
Earned Credit Hours	13.2	3.2	13.2	3.2
GPA Differential	0.64	0.65	0.65	0.66

Research Question One

The first research question sought to find which pre-college academic, psychosocial, noncognitive, and background factors were predictive of first-to-second year retention at a large, public four-year institution in the Southeast. To answer this question, a series of HGLMs were

Table 4.2
Unweighted and Weighted Frequencies for the Training Set

Variable	Categories	Unweighted Percentage	Weighted Percentage
Academic			
College Preparatory Curriculum	Yes	98%	98%
	No	2%	2%
Noncognitive			
Participated in Extracurricular Activities	Yes	98%	98%
	No	2%	2%
Had Out-of-Class Accomplishments in High School	Yes	99%	99%
	No	1%	1%
Plan to Participate in Extracurricular Activities in College	Yes	97%	97%
	No	3%	3%
Plan to Work in College	Yes	63%	62%
	No	19%	19%
	Unsure	18%	19%
Hours Anticipated Studying Per Week	Less than 10	5%	5%
	10 to 19	60%	60%
	20 to 29	29%	29%
	30 or more	6%	6%
Expected Educational Attainment	Bachelor's	21%	22%
	Graduate	79%	78%
Demographic			
Gender	Male	42%	45%
	Female	58%	55%
Ethnicity	White	74%	73%
	African American	13%	13%
	Asian	4%	5%
	Hispanic	6%	6%
	Other	3%	3%
First Generation	Yes	29%	28%
	No	71%	72%

(Table 4.2 continued)

Variable	Categories	Unweighted Percentage	Weighted Percentage
High School			
Status	Public	59%	61%
	Private	41%	39%
Residency	In-State	88%	81%
	Out-of-State	12%	19%
College			
Greek Life	Yes	28%	29%
	No	72%	71%
Retained	Yes	86%	86%
	No	14%	14%

run. The first model was the null unconditional model with no predictors and random intercepts for high schools. The second set of models were full models with all variables of interest included as possible predictors run on each individual imputed dataset. The final model was a single, combined model that consisted of all significant predictors from the individual imputations.

Null Model

The estimated intercept in the null, intercept only model was 1.89. This represented the log odds of a student from a typical high school being retained at the institution. Transforming the log odds to a predicted probability gives

$$p = \frac{e^{1.89}}{1 + e^{1.89}} = .8688.$$

This suggests that the likelihood a student from a typical high school would be retained to their second year was 86.9%. The estimated covariance of the intercept was 1.40. The intraclass correlation calculated using the modified formula for HGLMs with dichotomous outcomes was

$$\rho = \frac{1.40}{1.40 + 3.29} = .2985,$$

suggesting that 29.9% of the variability in retention rates is accounted for by high school differences. This variability was statistically significant, $Z(871) = 5.09$, $p < .01$, suggesting that the probability of being retained varied considerably across high schools.

Variable Reduction Models

An HGLM with all 28 variables of interest was run on each imputed dataset. The results of these models are presented in Table 4.3. Twelve of the variables were significant in all five individual models. These variables were included in the final pre-college HGLM with combined results across the five imputed datasets.

Pre-College Main Effects Model

Twelve total pre-college variables were significant predictors of first-to-second year retention. There were two significant predictors from each domain:

- Academic factors: High school GPA and registered credit hours
- Psychosocial factors: Academic self-efficacy and campus engagement
- Noncognitive factors: Plan to work in college and institutional fit
- Demographic factors: Gender and first generation status
- Financial factors: Scholarships and financial aid
- High school factors: Public/private status and enrollment size

The results for this model are presented in Table 4.4 and are discussed in further detail for each variable.

High school GPA and the number of registered hours in a student's course load were both positively related to retention. Odds ratios indicated that a one point increase in high school GPA was associated with more than a fourfold increase in the odds a student would be retained, holding all other variables constant. It should be noted that a full point change in high school

Table 4.3
HGLM Results for Each Imputed Dataset

Variable	1	2	3	4	5
High School GPA	.00*	.00*	.00*	.00*	.00*
ACT Score	.14	.17	.13	.19	.12
Registered Credit Hours	.00*	.00*	.00*	.00*	.00*
College Preparatory Curriculum	.10	.24	.77	.24	.27
Academic Self-Efficacy	.01*	.01*	.01*	.01*	.01*
Academic Engagement	.99	.90	1.00	.98	.99
Educational Commitment	.11	.08	.09	.08	.09
Resiliency	.50	.51	.51	.59	.53
Campus Engagement	.00*	.00*	.00*	.00*	.00*
Social Comfort	.39	.37	.40	.41	.40
Extracurriculars in High School	.47	.04*	.94	.85	.62
Accomplishments in High School	.51	.08	.57	.68	.93
Extracurriculars in College	.96	.42	.76	.38	.62
Plan to Work in College	.00*	.00*	.00*	.00*	.00*
Hours Anticipated Studying	.19	.19	.19	.16	.18
Expected Educational Attainment	.95	.93	.97	.92	.79
Institutional Fit	.00*	.00*	.00*	.00*	.00*
Gender	.02*	.01*	.02*	.01*	.02*
Ethnicity	.25	.24	.30	.29	.27
Scholarships	.00*	.00*	.00*	.00*	.00*
Financial Aid	.01*	.01*	.01*	.01*	.01*
Family Income	.90	.88	.89	.89	.94
First Generation	.00*	.00*	.00*	.00*	.00*
Residency Status	.61	.63	.58	.91	.65
High School Status	.00*	.00*	.00*	.00*	.00*
Distance from Home	.07	.07	.07	.19	.06
Enrollment	.00*	.00*	.00*	.00*	.00*
Percent of Minority Students	.58	.62	.57	.75	.60

Note. Significant *p*-values are designated with a *. Significance values less than .01 are rounded to .00.

GPA is dramatic, and that more realistic variations in high school GPA would result in smaller increases in the odds of retention. Each additional credit hour a student registered to take in their first semester increased the odds a student would be retained by 23%, holding all other variables constant. Since most courses are three credit hours, registering for an additional three hour

Table 4.4
Fixed Effects Estimates for the Pre-College Main Effects Model

Predictor	B	SE	OR	95% CI	*p*
Intercept	-8.50	0.725			< .01
High School GPA	1.45	0.130	4.25	[3.29, 5.49]	< .01
Registered Credit Hours	0.21	0.033	1.23	[1.15, 1.31]	< .01
Academic Self-Efficacy	-0.01	0.002	0.99	[0.99, 1.00]	.01
Campus Engagement	0.01	0.002	1.01	[1.01, 1.01]	< .01
Plan to Work					
Yes	-0.38	0.123	0.69	[0.54, 0.87]	< .01
Unsure	0.01	0.152	1.01	[0.75, 1.36]	.94
No					
Institutional Fit	0.28	0.059	1.32	[1.17, 1.48]	< .01
Scholarships	0.10	0.013	1.10	[1.08, 1.13]	< .01
Financial Aid	0.01	0.006	1.01	[1.00, 1.03]	.01
Gender	0.21	0.097	1.24	[1.02, 1.49]	.03
First Generation	-0.33	0.095	0.72	[0.60, 0.87]	< .01
High School Status	1.11	0.198	3.02	[2.05, 4.45]	< .01
High School Enrollment	0.04	0.011	1.04	[1.01, 1.06]	< .01

Note. The reference group is italicized. Females were the reference group for gender and public schools were the reference group for high school status.

course increased the odds of retention by a factor of 1.85, or 85%. The change in predicted probabilities for each variable (holding all other variables constant at their means) provided a practical estimate of effect size. These probabilities are presented in Figure 4.1. At the low end of the scale for high school GPA, a 0.1 point increase was associated with approximately a 3% increase in the probability of retention. This effect became smaller at the high end of the distribution. Registering for an additional credit hour was associated with approximately a 1% increase in the probability of being retained. As previously mentioned, courses typically are in three hour increments, so the estimated effect on retention probability of registering for an additional course was a 3% increase.

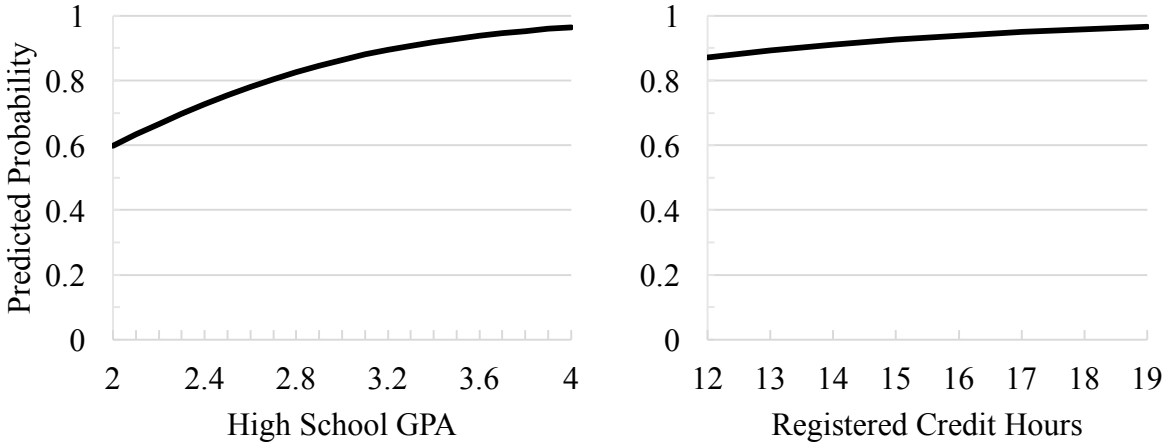

Figure 4.1. Predicted Probabilities of High School GPA and Registered Credit Hours in the Pre-College Main Effects HGLM

Academic self-efficacy was negatively related to retention, while campus engagement was positively related. Odds ratios indicated that a one point increase on the academic-self-efficacy scale was associated with a half percent decrease in the odds a student would be retained, holding all other variables in the model constant. A one point increase on the campus engagement scale was associated with a one percent increase in the odds of retention, holding all other variables constant. Since both measures were scales constructed from an inventory of items and a one point increase does not have substantive meaning, examining these in the context of standard deviations provides more meaningful interpretation. For a one standard deviation increase on the academic self-efficacy scale (s = 22.79), the odds of being retained decreased by 11%. Similarly, a one standard deviation increase on the campus engagement scale (s = 23.70) was associated with a 28% increase in the odds of retention. The predicted probabilities presented in Figure 4.2 showed that these effect sizes were relatively small. A one standard deviation change from the mean for the academic self-efficacy scale was associated with a less than a 1% change in either direction. Similarly, a one standard deviation change from the mean

for the campus engagement scale was associated with about a 1.5% change in the probability in either direction.

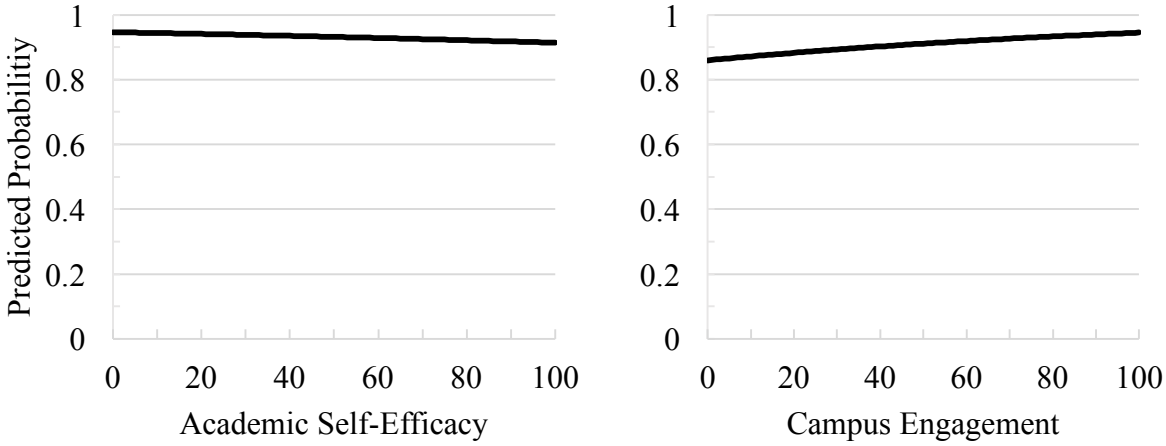

Figure 4.2. Predicted Probabilities for Academic Self-Efficacy and Campus Engagement in the Pre-College Main Effects HGLM

Planning to have a job while enrolled in college and the institutional fit score were both significant predictors of retention. Students who were unsure of their job plans were not significantly different than those who did not plan have a job (p = .94). However, the results suggest students who did plan to have a job were less likely to be retained than students who did not plan to work while in college. Odds ratios indicated that planning to have a job was associated with a 31% decrease in the odds a student would be retained, holding all other variables constant in the model. Institutional fit was positively related to retention. A one point increase in institutional fit score (i.e., a student's preference matched one additional characteristic of the institution) was associated with a 32% increase in the odds a student would be retained, holding all other variables constant. The predicted probability (holding all other variables constant at their mean) for students who did not plan to work and students who were unsure about their plans to work in college was .94. Students who planned to work had a predicted probability of .91—three percentage points lower than the other two groups. The

results for institutional fit were more pronounced and are visually represented in Figure 4.3. Students whose preferences did not match the institution in any way had retention probabilities that were 10% lower than students who fully matched. Each additional preference match was associated with approximately a 2% increase in the probability of retention.

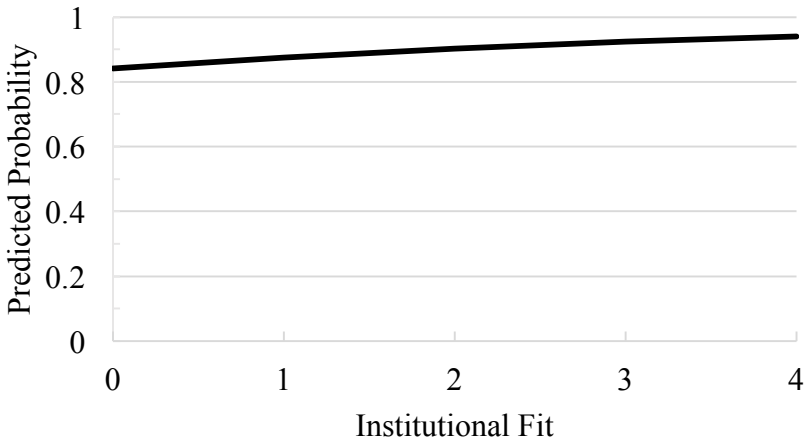

Figure 4.3. Predicted Probabilities for Institutional Fit in the Pre-College Main Effects HGLM

Gender and first generation status were both significantly related to retention. Male students were 24% more likely to be retained than female students, holding all other variables constant in the model. The odds ratio for first generation status suggests that students whose parents did not obtain a bachelor's degree were 28% less likely to be retained than non-first generation students. The predicted probabilities for male and female students suggested that males had a little over one percentage point higher retention probabilities than female students when all other variables were held constant at their means. Male students had a predicted retention probability of .93 while females had a predicted probability of .92. The predicted retention probability for first generation students was two percentage points lower than non-first generation students, as first generation students had predicted probabilities of .91 and non-first generation students had predicted probabilities of .93.

The amount of financial assistance in the form of scholarships and/or financial aid was positively related to retention. Odds ratios suggested that a one thousand dollar increase in scholarships was associated with a 10% increase in the odds of retention while a one thousand dollar increase in financial aid was associated with a 1% increase in these odds, holding all other variables constant. The predicted probabilities for changes in scholarships and financial aid evaluated holding all other variables constant at their means are shown in Figure 4.4. The results showed that each additional thousand dollar increase in scholarship monies increased the retention probability by approximately one percentage point, but this effect was diminished at the high end of the scale. The effect of increased financial aid was much smaller, with each additional thousand dollars increasing the probability of retention by at most one tenth of a percentage point.

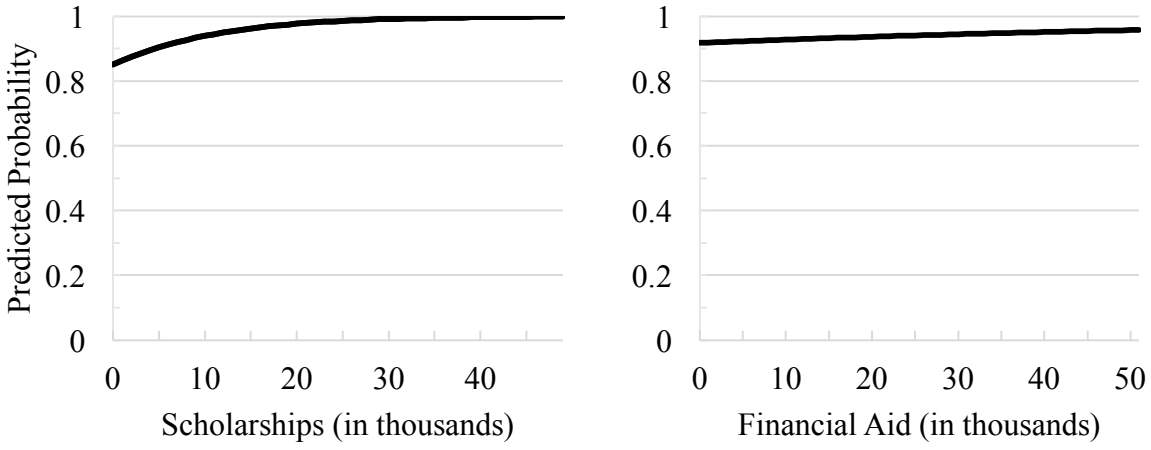

Figure 4.4. Predicted Probabilities for Scholarships and Financial Aid in the Pre-College Main Effects HGLM

Two high school characteristics were significantly related to retention. Students who attended a private high school were significantly more likely to be retained than students who attended a public high school. The odds ratio suggested that private school students were three times more likely to be retained than public school students on average, holding all other

variables constant. High school enrollment was also positively related to retention. The odds ratio suggests that a one hundred student increase in high school enrollment was associated with a 4% increase in the odds a student would be retained. Predicted probabilities evaluated at the mean for all other variables in the model showed that students who attended a private school had retention probabilities that were nine percentage points higher than public school students. Students from private schools had predicted probabilities of .95 while students from public schools had predicted probabilities of .86. High school enrollment also had a sizeable effect, with students from the smallest high schools having predicted retention probabilities that were seven to nine percentage points lower than students from the largest high schools (see Figure 4.5).

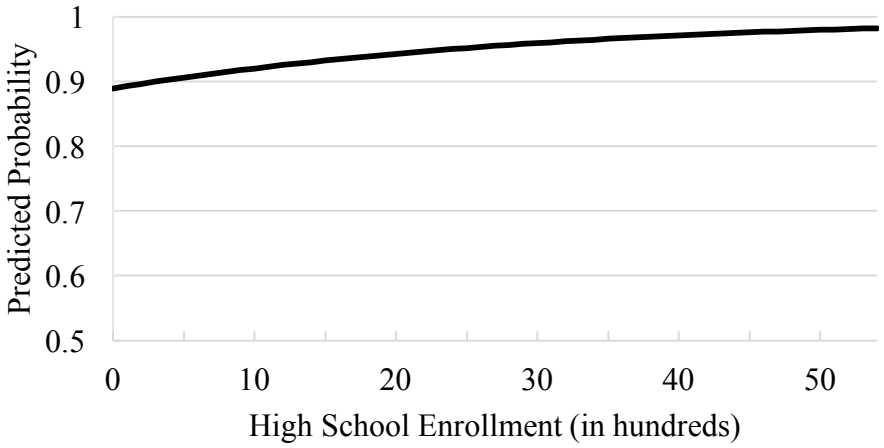

Figure 4.5. Predicted Probabilities for High School Enrollment in the Pre-College Main Effects HGLM

Research Question Two

The second research question sought to understand how the relationship between the pre-college variables and retention outcomes were affected once first semester college performance and activity was taken into consideration. To answer this question, three variables were added to the pre-college main effects model. Changes in parameter estimates and significance were

examined. Finally, variables that were not significant were removed from the model to create a final, parsimonious fixed effects model for post-first semester prediction of retention. The results of these two models in comparison to the pre-college main effects model are presented in Table 4.5. The results showed that the number of hours earned and participating in Greek life were both positively related to retention, while the differential between high school and college GPA was negatively related to retention. The inclusion of these three post-first semester variables made several factors that were significant in the pre-college model no longer significant. Specifically, academic self-efficacy, planning to have a job in college, financial aid, first generation status, and high school enrollment were no longer significant predictors of retention. These five variables were consequently removed from the final post-first semester main effects model, then odds ratios and predicted probabilities calculated by holding other variables at their means were examined. This allowed insight into the effect sizes and changes in relationships from the pre-college main effects model for the new and remaining predictors.

Table 4.5
Fixed Effects Estimates for the Post-First Semester Main Effects Model

Predictor	B	SE	OR	95% CI	*p*
Intercept	-4.57	0.71			< .01
High School GPA	0.88	0.14	2.42	[1.85, 3.16]	< .01
Registered Credit Hours	0.10	0.04	1.11	[1.03, 1.19]	< .01
Campus Engagement	0.01	0.00	1.01	[1.00, 1.01]	< .01
Institutional Fit	0.25	0.06	1.28	[1.14, 1.44]	< .01
Gender	0.30	0.10	1.34	[1.10, 1.63]	< .01
Scholarships	0.04	0.01	1.04	[1.01, 1.06]	< .01
High School Status	0.46	0.16	1.59	[1.16, 2.17]	< .01
Earned Credit Hours	0.09	0.02	1.09	[1.05, 1.13]	< .01
Greek Life	0.83	0.13	2.29	[1.78, 2.95]	< .01
GPA Differential	-1.08	0.09	0.34	[0.28, 0.41]	< .01

The odds ratio for high school GPA suggested that a one point increase in GPA was associated with a 142% increase in the odds of retention on average, holding all other variables constant at their mean. As noted before, a full point change in high school GPA is dramatic and more realistic variations in high school GPA would result in smaller increases in the odds of retention. Predicted probabilities showed that at the low end of the scale for high school GPA, a 0.1 point increase was associated with approximately a 1.5% increase in the probability of retention. This effect became smaller at the high end of the distribution. Odds ratios showed that for each additional credit hour registered for was associated with an 11% increase in the odds a student would be retained an average, holding all other variables constant. Therefore, registering for an additional three credit hour course was associated with a 35% increase in the odds of retention. Predicted probabilities presented in Figure 4.6 shows that registering for an

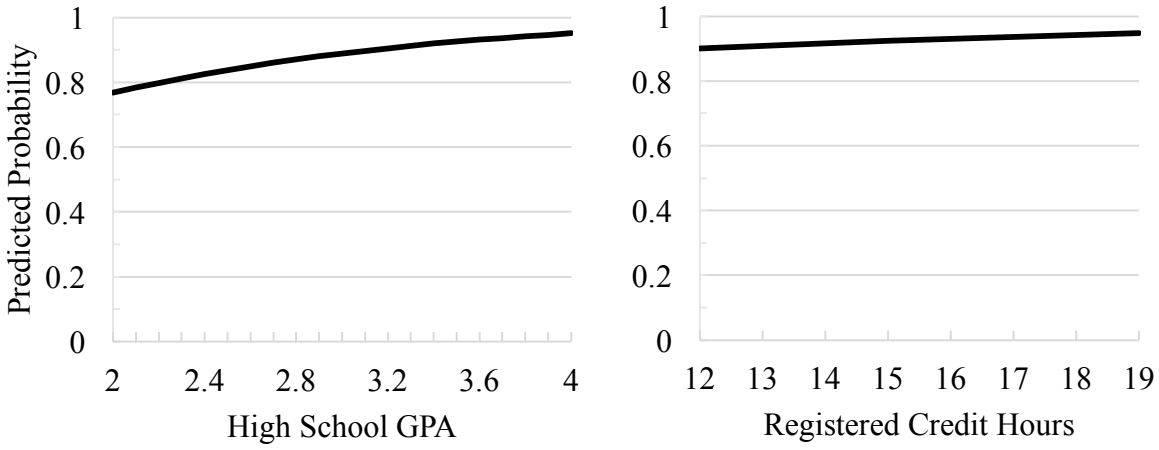

Figure 4.6. Predicted Probabilities for High School GPA and Registered Credit Hours in the Post-First Semester Main Effects HGLM

additional credit hour was associated with less than a 1% increase in the probability of being retained. As previously mentioned, courses typically are in three hour increments, so the estimated change in retention probability when registering for an additional course ranged from a

1.7% to 2.4% increase. Comparing these results to the pre-college main effects model, the relationships of both high school GPA and hours registered with retention are attenuated once first semester performance and activity were taken into account.

The only psychosocial variable that remained significant in the post-first semester main effects model was campus engagement. Odds ratios showed that a one point increase on the campus engagement scale was associated with a little over half a percent increase in the odds of retention on average, holding all other variables in the model constant. In terms of standard deviations, a one standard deviation increase was associated with 15% increase in these odds. Predicted probabilities showed that a one standard deviation change from the mean for the campus engagement scale had a minimal change in the predicted probability of retention of less than 1% (see Figure 4.7). Similar to the pre-college academic factors, the association between

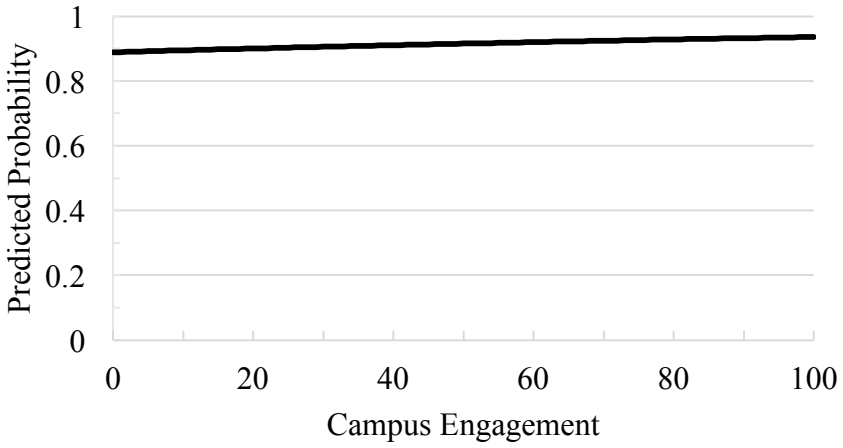

Figure 4.7. Predicted Probabilities for Campus Engagement in the Post-First Semester Main Effects HGLM

campus engagement and retention was diminished once first semester performance and participation were taken into account.

The institutional fit metric was significant in the post-first semester main effects model. Odds ratios showed that a one point increase in institutional fit was associated with a 23%

increase in the odds a student would be retained on average, holding all other variables constant. Predicted probabilities showed that this change was associated with approximately a 1.5% to 2.9% increase in the probability of retention. Scholarship money also remained significant in this model. Odds ratios indicated that a one thousand dollar increase in scholarships was associated with a 4% increase in the odds a student would be retained on average, holding all other variables constant. Predicted probabilities presented in Figure 4.8 showed that this change was related to a less than 1% increase in the probability of retention. The strength of these factors, like the other pre-college factors previously discussed, were lessened once post-first semester performance and activity were taken into account.

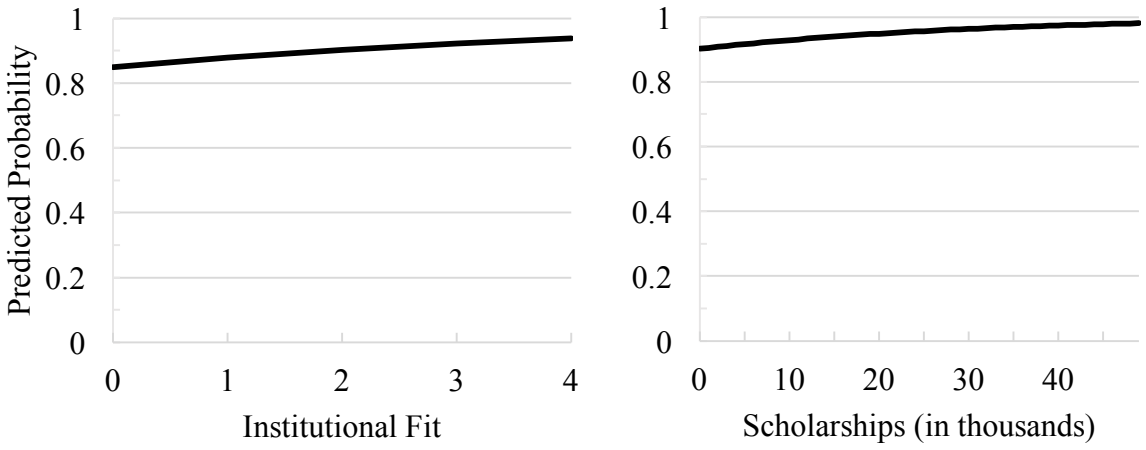

Figure 4.8. Predicted Probabilities for Institutional Fit and Scholarships in the Post-First Semester Main Effects HGLM

The final two factors from the pre-college main effects model that remained significant were gender and high school status. Odds ratios indicated that male students were 34% more likely to be retained than female students on average, holding all other variables constant. The predicted probabilities showed that this effect size was small because the probability of a male being retained was two percentage points higher than the probability of a female being retained (.93 compared to .91). Odds ratios for high school status indicated that private school students

were 59% more likely to be retained than public school students on average, holding all other variables constant. Predicted probabilities showed that this difference amounted to a three and a half point increase in the probability of retention for private school students (.93 compared to .90). The estimated difference between males and females was magnified when post-first semester factors were considered, but it was still small overall. Conversely, the difference between students from private schools compared to students from public schools was lessened.

Hours earned and participation in Greek life were both positively related to retention. Odds ratios indicated that, on average, each addition earned credit hour was associated with a 9% increase in the odds a student would be retained, holding all other variables constant. Considering this in terms of earning credit for a typical three-hour course, the odds retention increased by 29% when credit is earned for an additional course. The predicted probabilities presented in Figure 4.9 showed that students earning minimal credit had significantly lower probabilities of retention than students earning several credit hours. The results showed that earning credit for a three-hour course increased the probability of retention by nearly one and a half percentage points at the high end of the scale to nearly four percentage points at the low end of the scale. Odds ratios suggested that students who participated in Greek life were 2.3 times more likely to be retained on average than students who did not, holding all other variables constant. Predicted probabilities indicated that students in Greek life had a predicted retention probability of .96—five percentage points higher than non-Greek students who had a predicted probability of .91. The magnitude of the differential between high school and college GPA was negatively related to retention. Specifically, the odds ratio indicated that a one point difference was associated with a 66% decrease in the odds a student would be retained. In terms of predicted probabilities, students with a full point differential had predicted probabilities that were

seven percentage points lower than students with no differential. This decrease was more dramatic with the largest discrepancies, as students with the largest discrepancies had retention probabilities that were 68 percentage points lower than students with no differential.

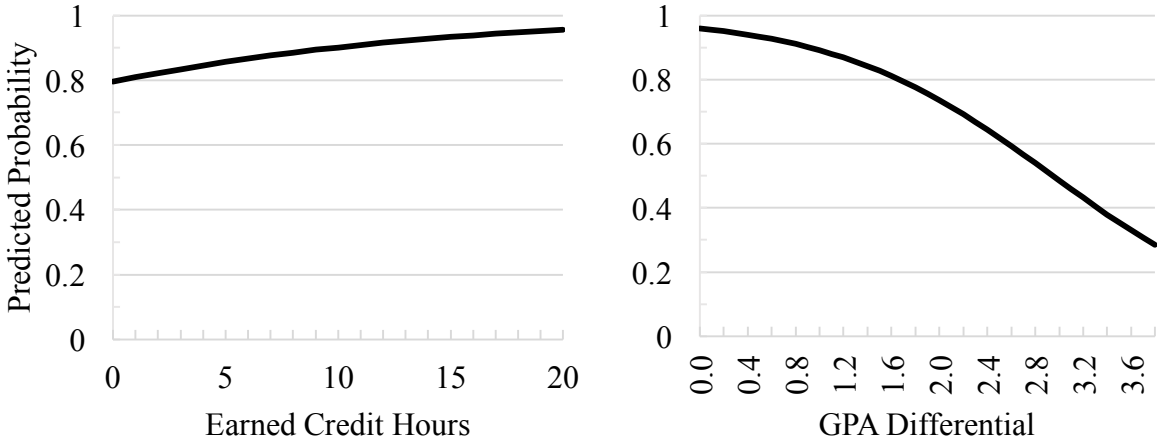

Figure 4.9. Predicted Probabilities for Earned Credit Hours and GPA Differential in the Post-First Semester Main Effects HGLM

Research Question Three

The third research question sought to determine how the relationships revealed in the pre-college and post-first semester main effects models were moderated by gender, ethnicity, and first generation status. To answer this question, all possible interactions between the main effects and the three demographic variables of interest were added to each model. Significant interactions were retained for a final model that allowed for examination of moderated relationships in both the pre-college and post-first semester models.

Pre-College Interaction Model

The results of the pre-college interaction model presented in Table 4.6 showed that five interaction terms were significant. When ethnicity was entered into the model as a main effect it was not significant, but it was a significant moderator for the relationships between gender and retention as well as scholarships and retention. The results showed that the odds ratios

comparing the odds of retention for males to the odds of retention for females for White and African American were not significantly different from one another. The other three ethnic groups, however, had significantly moderated relationships compared to White students. Figure 4.10 shows that, on average and holding all other variables constant, the odds of retention for males were higher than females for White and African American students, while females had higher odds of retention than males for the other three ethnic groups.

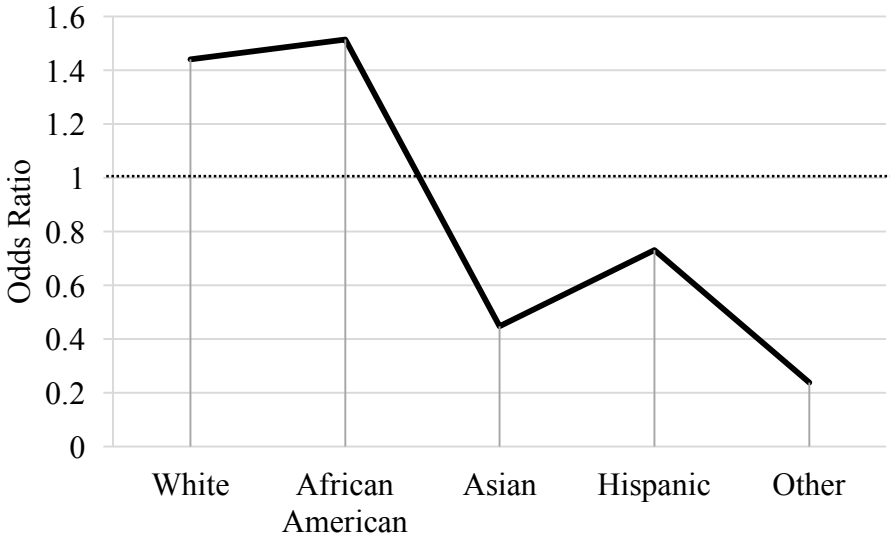

Figure 4.10. Odds Ratios for Ethnic Groups Comparing Males to Females for the Pre-College Interaction HGLM. Females were the reference group.

The interaction term between ethnicity and scholarships was significant, which indicated that the relationship between scholarships and retention was moderated by ethnicity. The interaction was only significant between White and Hispanic students. The relationship between scholarships and retention for all other ethnic groups was not significantly different than it was for White students. Figure 4.11 represents the odds ratios comparing the odds of retention for White students to the odds of retention for Hispanic students. The results showed that, on average and holding all other variables constant, an increase in scholarship dollars increased the odds of retention for White students at a greater rate than it did for Hispanic students.

Table 4.6
HGLM Results for the Pre-College Interaction Model

Variable	B	SE	OR	95% CI	*p*
Intercept	-8.45	0.744			< .01
High School GPA	1.40	0.132	4.04	[3.11, 5.23]	< .01
Registered Credit Hours	0.20	0.033	1.22	[1.15, 1.30]	< .01
Academic Self-Efficacy	-0.005	0.002	1.00	[0.99, 1.00]	.02
Campus Engagement	0.01	0.002	1.01	[1.01, 1.01]	<.01
Plan to Work					
Yes	-0.37	0.124	0.69	[0.54, 0.88]	< .01
Unsure	0.03	0.153	1.03	[0.76, 1.39]	.84
No					
Institutional Fit	0.27	0.059	1.31	[1.17, 1.47]	< .01
Financial Aid	0.01	0.006	1.01	[0.99, 1.02]	.34
First Generation	-0.42	0.308	0.65	[0.36, 1.20]	.17
High School Status	1.18	0.192	3.24	[2.23, 4.72]	< .01
Scholarships	0.09	0.017	1.10	[1.06, 1.14]	< .01
Gender	0.36	0.116	1.44	[1.15, 1.81]	< .01
High School Enrollment	0.05	0.012	1.05	[1.03, 1.07]	< .01
Ethnicity					
African American	-0.17	0.243	0.85	[0.53, 1.36]	.49
Asian	0.31	0.598	1.37	[0.42, 4.42]	.60
Hispanic	0.87	0.330	2.40	[1.25, 4.58]	.01
Other	0.86	0.445	2.37	[0.99, 5.66]	.05
White					
Gender*Ethnicity					
African American	0.05	0.250	1.05	[0.64, 1.72]	.84
Asian	-1.17	0.473	0.31	[0.12, 0.78]	.01
Hispanic	-0.68	0.332	0.51	[0.26, 0.97]	.04
Other	-1.81	0.476	0.16	[0.06, 0.42]	< .01
White					
Scholarships*Ethnicity					
African American	0.02	0.032	1.02	[0.96, 1.08]	.57
Asian	0.11	0.077	1.11	[0.96, 1.29]	.16
Hispanic	-0.11	0.032	0.90	[0.84, 0.96]	< .01
Other	-0.03	0.044	0.97	[0.89, 1.05]	.45
White					
Scholarships*First Generation	0.06	0.027	1.06	[1.00, 1.12]	.04
Financial Aid*First Generation	0.03	0.011	1.03	[1.01, 1.05]	.01
HS Enrollment*First Generation	-0.04	0.015	0.96	[0.93, 0.99]	.01

Note. The reference group is italicized.

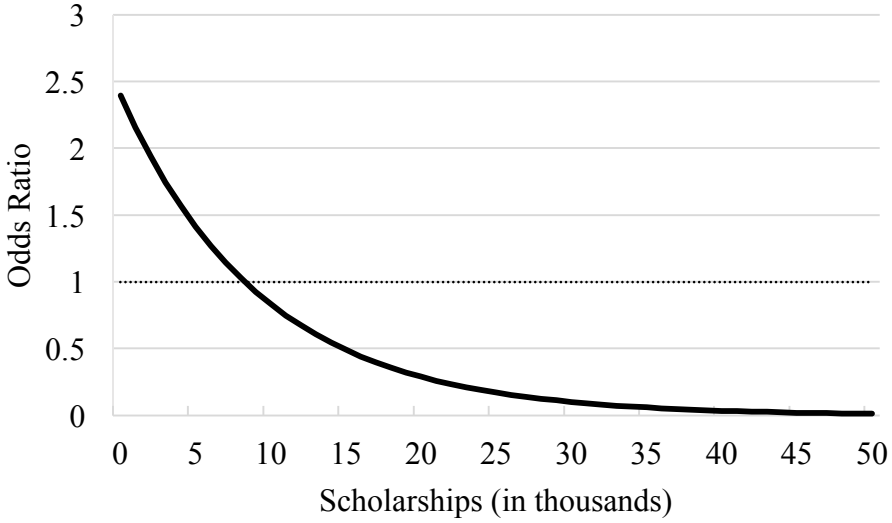

Figure 4.11. Odds Ratios for Scholarships Comparing Hispanic Students to White Students for the Pre-College Interaction HGLM. White students were the reference group.

First generation status was a significant moderator of the relationship between scholarships and retention as well as financial aid and retention. Figure 4.12 shows the change in the odds ratio between the odds of retention for first generation students and the odds of retention between non-first generation students as the amount of scholarships increased. The results showed that, on average and holding all other variables constant, the odds of retention for first generation students were higher with larger amounts of scholarship money, while non-first generation students had higher retention odds with scholarships of less than about $8,000. These results were similar with the moderated relationship between financial aid and retention. Figure 4.13 shows the change in the odds ratio between the odds of retention for first generation students and the odds of retention between non-first generation students as the amount of financial aid increased. The results showed that, on average and holding all other variables constant, the odds of retention for first generation students were higher with larger amounts of financial aid, while non-first generation students had higher retention odds with financial aid of less than about $15,000.

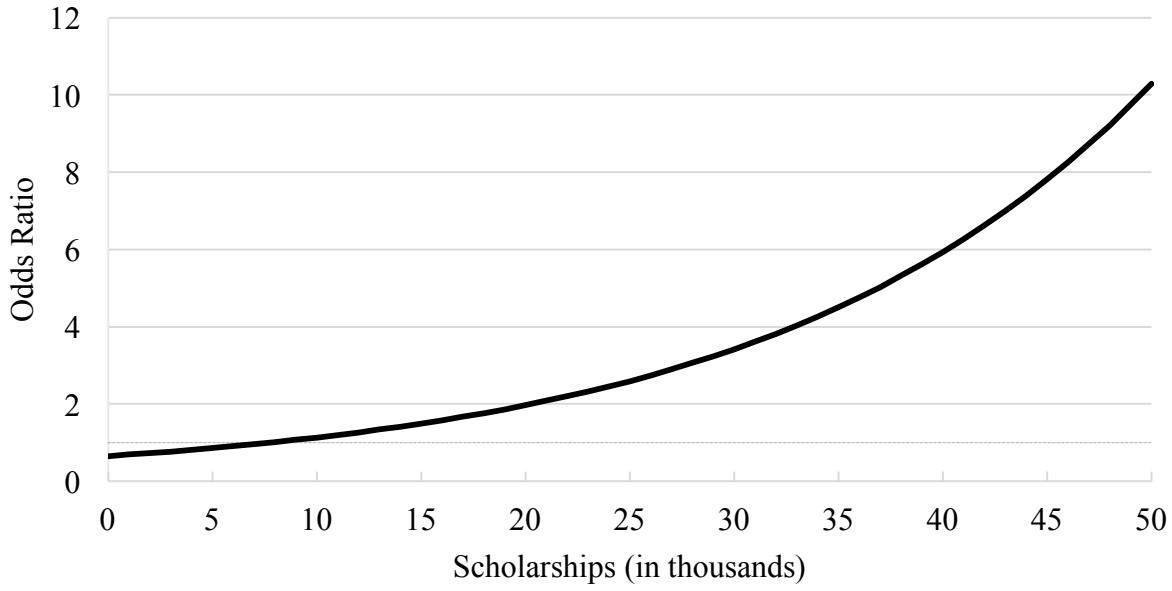

Figure 4.12. Odds Ratios for Scholarships Comparing First Generation Students to Non-First Generation Students for the Pre-College Interaction HGLM. Non-first generation students were the reference group.

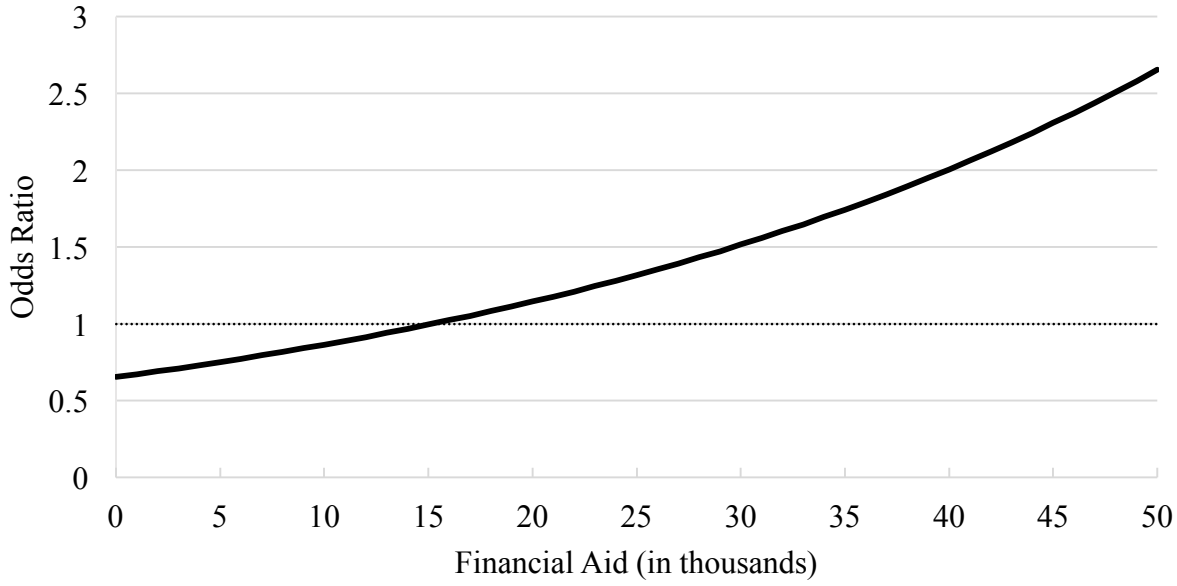

Figure 4.13. Odds Ratios for Financial Aid Comparing First Generation Students to Non-First Generation Students for the Pre-College Interaction HGLM. Non-first generation students were the reference group.

The relationship between high school enrollment and retention was also moderated by first generation status. The results shown in Figure 4.14 reflect the change in the odds ratio between the odds of retention for first generation students and the odds of retention between non-first generation students as high school enrollment increased. The results showed that, on average and holding all other variables constant, larger high school enrollments negatively magnified the difference in the odds of retention for first generation students compared to the odds of retention for non-first generation.

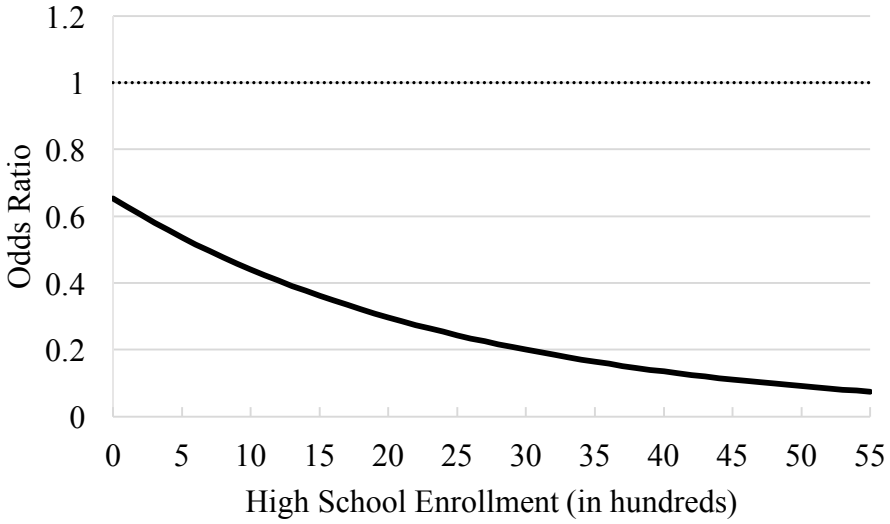

Figure 4.14. Odds Ratios for High School Enrollment Comparing First Generation Students to Non-First Generation Students for the Pre-College Interaction HGLM. Non-first generation students were the reference group.

Post-First Semester Interaction Model

The inclusion of the three post-first semester variables into the pre-college interaction model caused some of the variables resulted in a similar effect to the main effects model. Academic self-efficacy, planning to have a job in college, financial aid, first generation status, and high school enrollment were no longer significant predictors of retention. Additionally, the interactions between first generation status and scholarship money, financial aid, and high school

enrollment were no longer significant in the post-first semester model. These variables and interactions were consequently removed from the model. Interactions were also tested between the three post-enrollment variables and gender, ethnicity, and first generation status. None of these interactions were significant, so they were not included in the final post-first semester interaction model. This left two significant moderated relationships in the post-first semester interaction model. The results are shown in Table 4.7.

The first significant interaction was between gender and ethnicity. The results showed that the odds ratios comparing the odds of retention for males to the odds of retention for females for African American and Hispanic students were not significantly different from White students. The other two ethnic groups, however, had significantly moderated relationships compared to White students. Figure 4.15 shows that, on average and holding all other variables constant, the odds of retention for males were higher than females for White and African American students, while females had higher odds of retention than males for Asians and students from other ethnic groups. This differed slightly from the pre-college interaction model in that the interaction was not significant for Hispanic students. The second significant interaction was between ethnicity and scholarship money. The interaction was only significant between White and Hispanic students. The relationship between scholarships and retention for all other ethnic groups was not significantly different than it was for White students. Figure 4.16 represents the odds ratios comparing the odds of retention for White students to the odds of retention for Hispanic students. The results showed that, on average and holding all other variables constant, an increase in scholarship dollars increased the odds of retention for White students at a greater rate than it did for Hispanic students.

Table 4.7

HGLM Results for the Post-First Semester Interaction Model

Variable	B	SE	OR	95% CI	p
Intercept	-4.77	0.715			< .01
High School GPA	0.91	0.139	2.48	[1.89, 3.25]	< .01
Registered Credit Hours	0.09	0.036	1.1	[1.02, 1.18]	.01
Campus Engagement	0.01	0.002	1.01	[1.00, 1.01]	.01
Institutional Fit	0.25	0.060	1.29	[1.14, 1.45]	< .01
High School Status	0.54	0.155	1.71	[1.26, 2.32]	< .01
Scholarships	0.04	0.013	1.04	[1.01, 1.07]	< .01
Gender	0.47	0.120	1.60	[1.26, 2.02]	< .01
Ethnicity					
African American	0.25	0.234	1.28	[0.81, 2.02]	.29
Asian	0.47	0.581	1.60	[0.51, 5.00]	.42
Hispanic	0.79	0.329	2.20	[1.15, 4.19]	.02
Other	1.03	0.458	2.81	[1.15, 6.90]	.02
White					
Gender*Ethnicity					
African American	-0.02	0.259	0.98	[0.59, 1.63]	.95
Asian	-1.17	0.488	0.31	[0.12, 0.81]	.02
Hispanic	-0.56	0.347	0.57	[0.29, 1.13]	.11
Other	-1.84	0.518	0.16	[0.06, 0.44]	< .01
White					
Scholarships*Ethnicity					
African American	0.02	0.029	1.02	[0.96, 1.08]	.48
Asian	0.08	0.071	1.08	[0.94, 1.25]	.26
Hispanic	-0.08	0.029	0.92	[0.87, 0.98]	.01
Other	0.01	0.043	1.01	[0.93, 1.10]	.75
White					
Earned Credit Hours	0.09	0.019	1.09	[1.05, 1.13]	< .01
Greek	0.89	0.131	2.44	[1.88, 3.15]	< .01
GPA Differential	-1.09	0.093	0.34	[0.28, 0.40]	< .01

Note. The reference group is italicized. Public school students were the reference group for high school status and females were the reference group for gender.

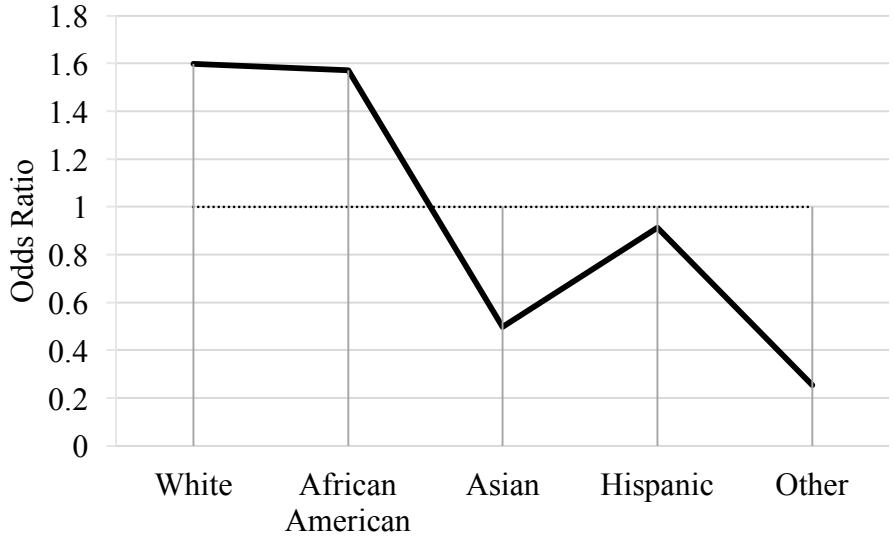

Figure 4.15. Odds Ratios for Ethnic Groups Comparing Males to Females for the Post-First Semester Interaction HGLM. Females were the reference group.

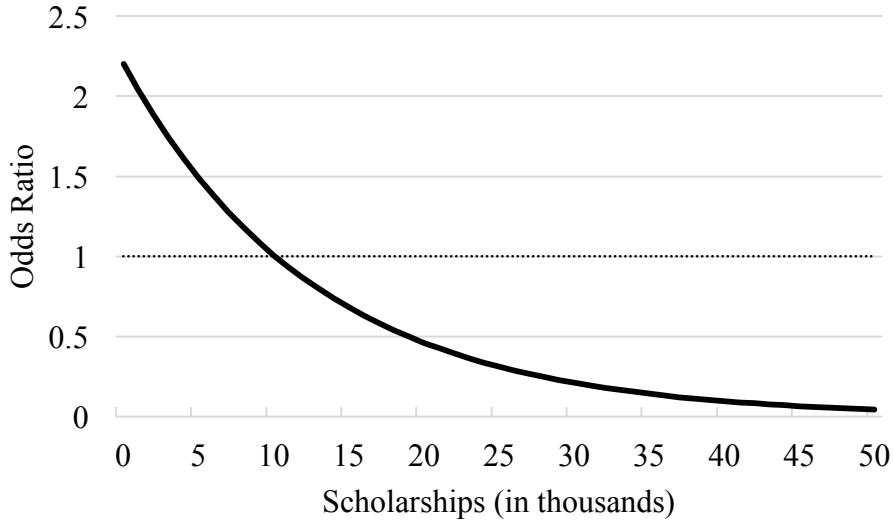

Figure 4.16. Odds Ratios for Scholarships Comparing Hispanic Students to White Students for the Post-First Semester Interaction HGLM. White students were the reference group.

Research Question Four

A series of tree-based models were built which allowed for analysis of variable importance and construction of predictive models. A random forest was constructed on the training set using the pre-college variables, followed by gradient tree boosting applied to the

same data for 378 trees because it was the best cross validation iteration out of 1,000 iterations. The results from the random forest showed that the five most important variables were scholarships, family income, high school GPA, distance from home, and percent of minority students at the student's high school. The results from the gradient tree boosting showed that the amount of scholarships had the greatest relative influence on retention, followed by the distance from home, high school GPA, and amount of financial aid. Together, these four variables contributed to half the boosted tree models. The variable importance graphs are presented in Figures 4.17 and 4.18 for the random forest and gradient tree boosting, respectively.

Scholarships, financial aid, high school GPA, and high school enrollment were significant predictors in the HGLM. However, distance from home, family income, and percentage of minority students were not. These seven variables together were the most important variables in both tree-based methods. The partial dependence plots in Figures 4.19, 4.20, and 4.21 describe the effects of each of these variables on retention after accounting for the other variables in the model. Distance from home and family income are both represented in logarithmic units; therefore the actual value represented in the plot should be reverse transformed from the logarithmic scale. For example, a value of zero represents one, a value of one represents ten, a value of two represents 100, and so on. The results showed an overall negative relationship between distance from home and retention. Students who attended a high school near the university were most likely to be retained, while students who attended high schools further away were less likely to be retained. For family income, it should be noted that the middle 80% of the training set fell between 4.3 and 5.5 logarithmic units (i.e., approximately between $20,000 and $315,000). The partial dependence plot therefore shows a positive relationship between family income and retention when concentrating on the range described,

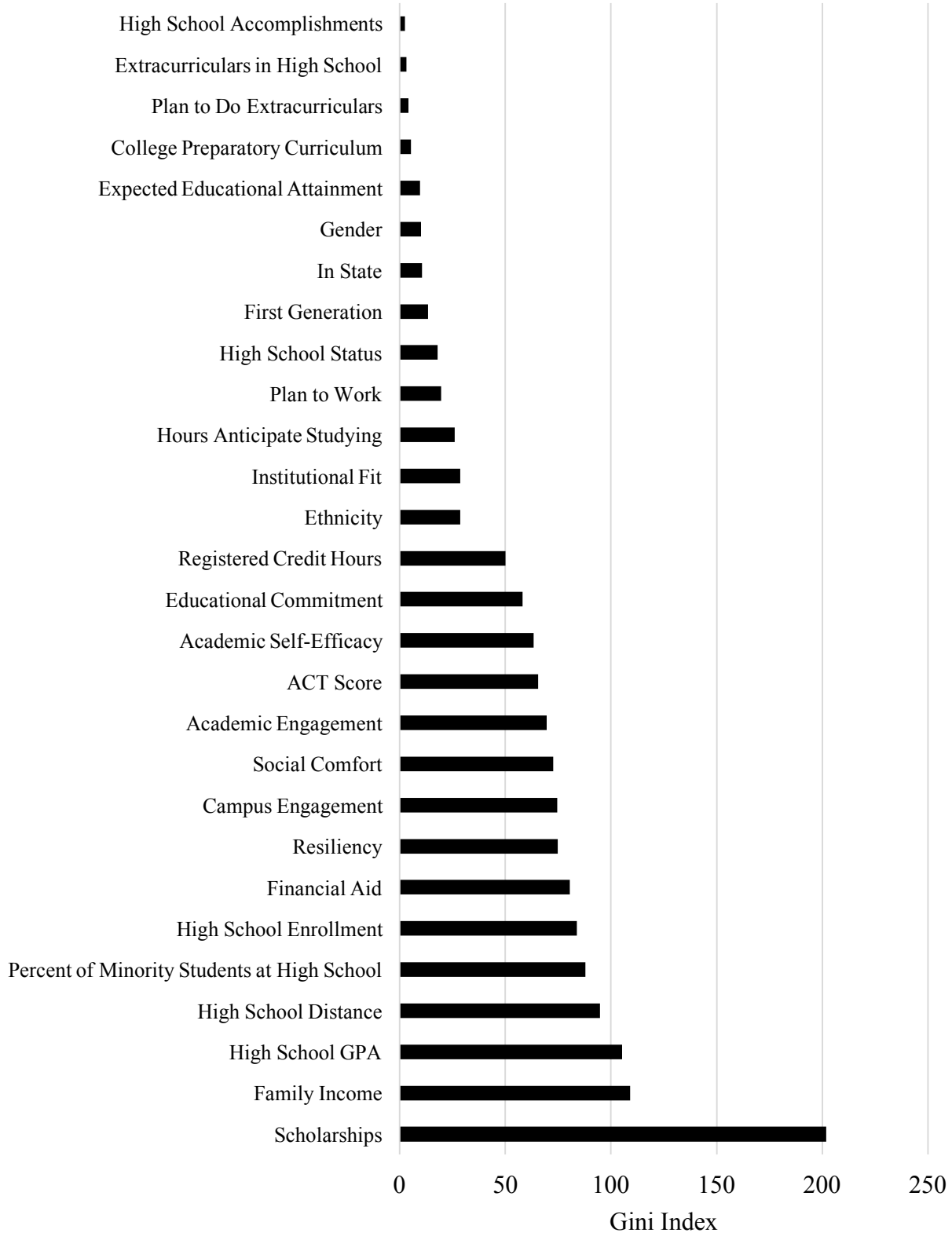

Figure 4.17. Relative Importance of Variables in the Pre-College Random Forest Model

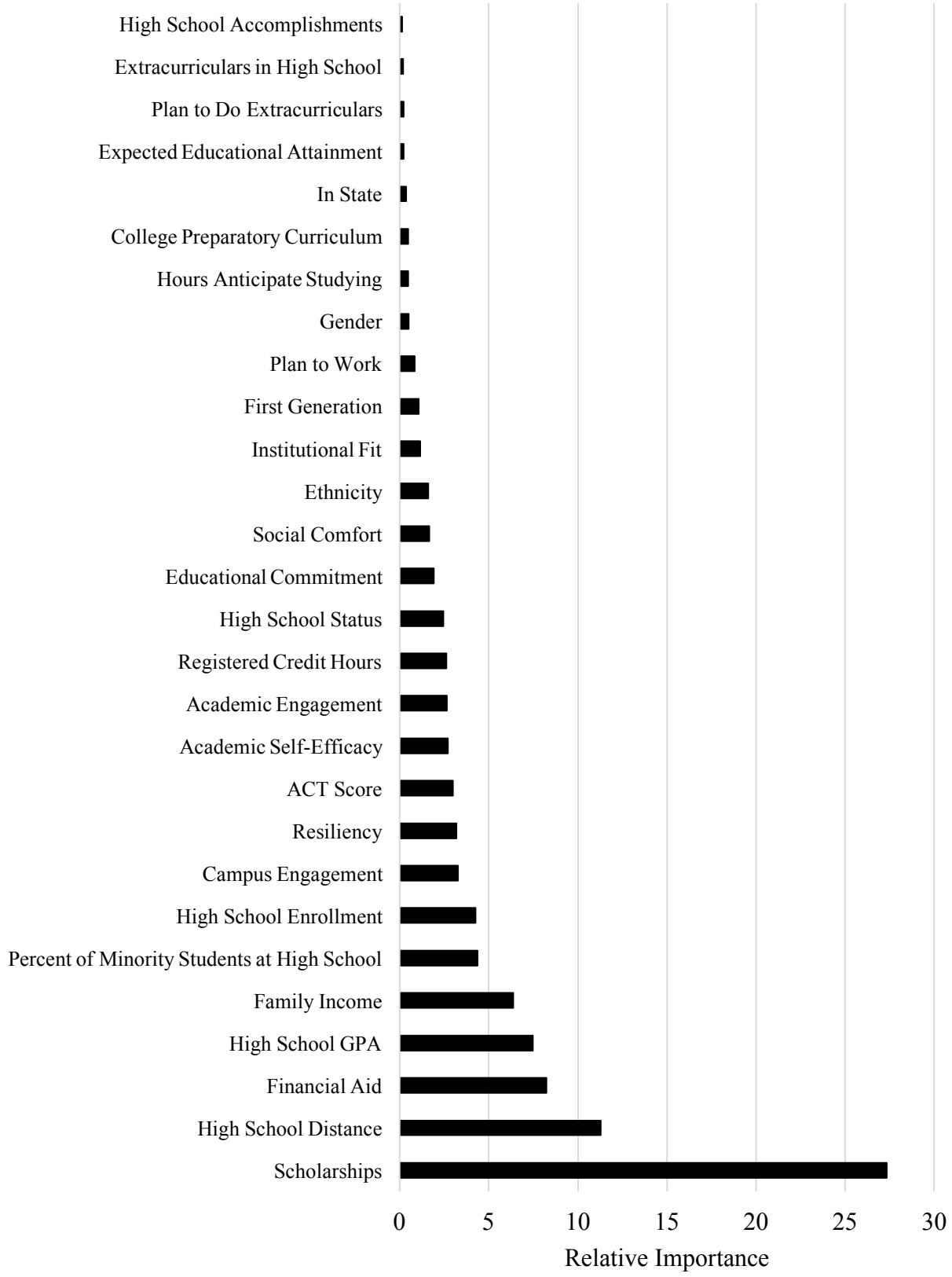

Figure 4.18. Relative Importance of Variables in the Pre-College Boosted Tree Model

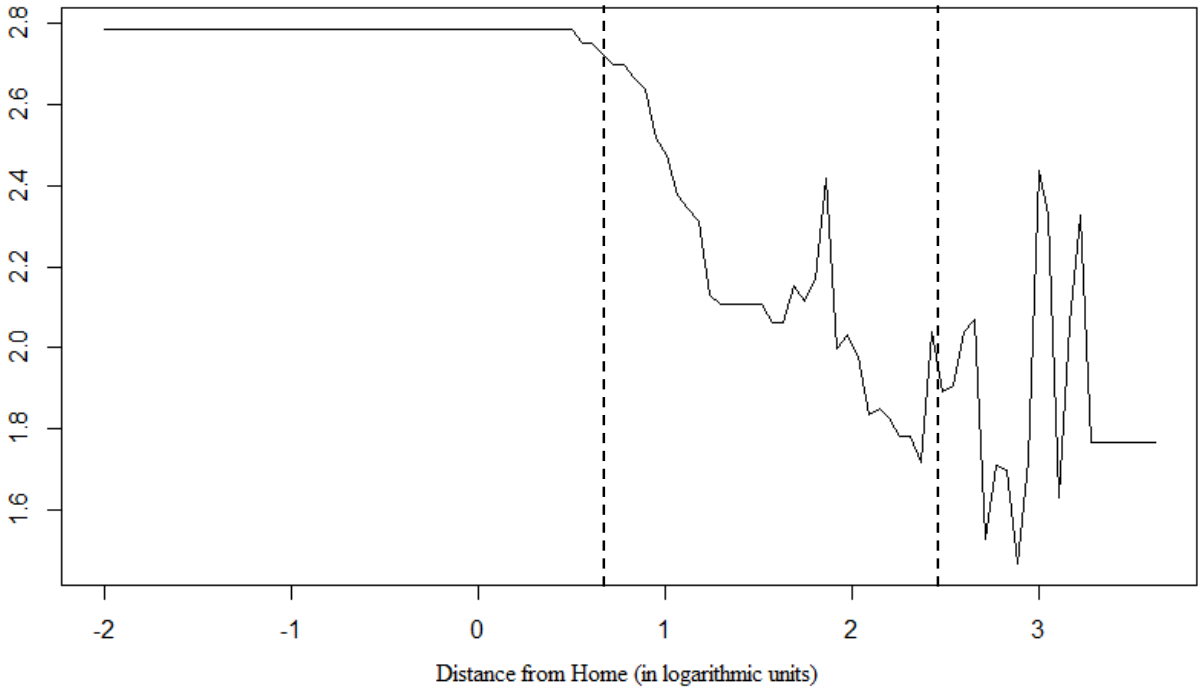

Figure 4.19. Pre-College Model Partial Dependence Plot for Distance from Home. Scale is in logarithmic units. The dashed lines represent the middle 80% of the sample.

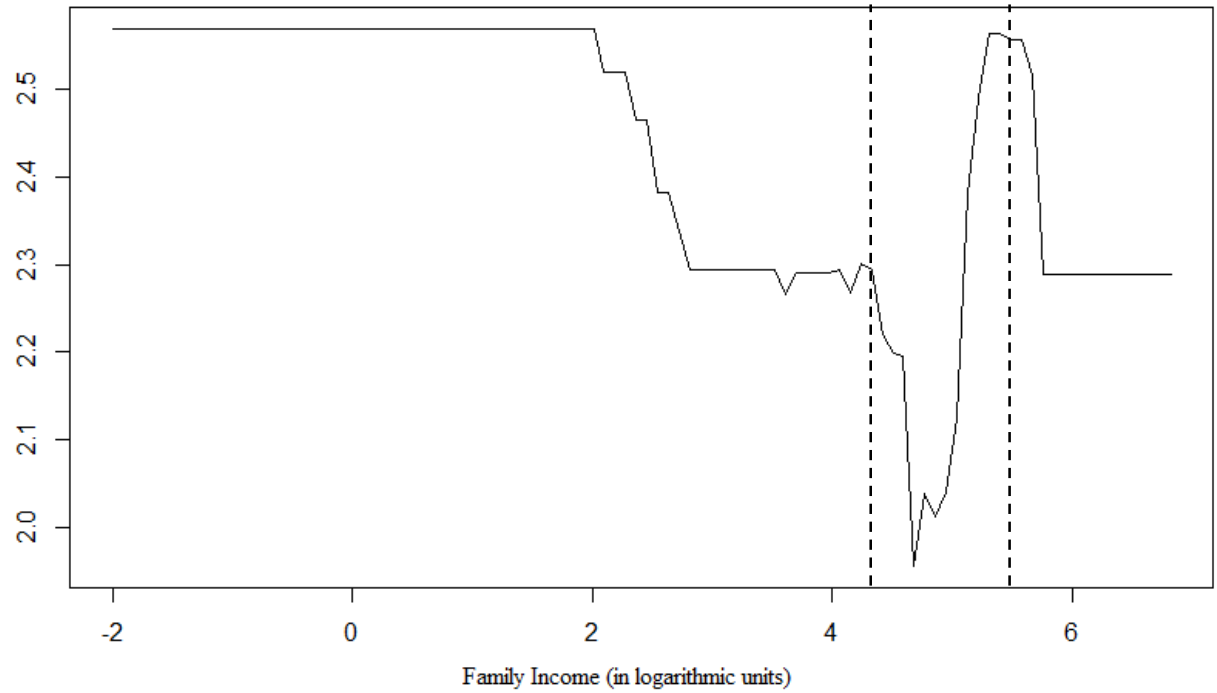

Figure 4.20. Pre-College Model Partial Dependence Plot for Family Income. Scale is in logarithmic units. The dashed lines represent the middle 80% of the sample.

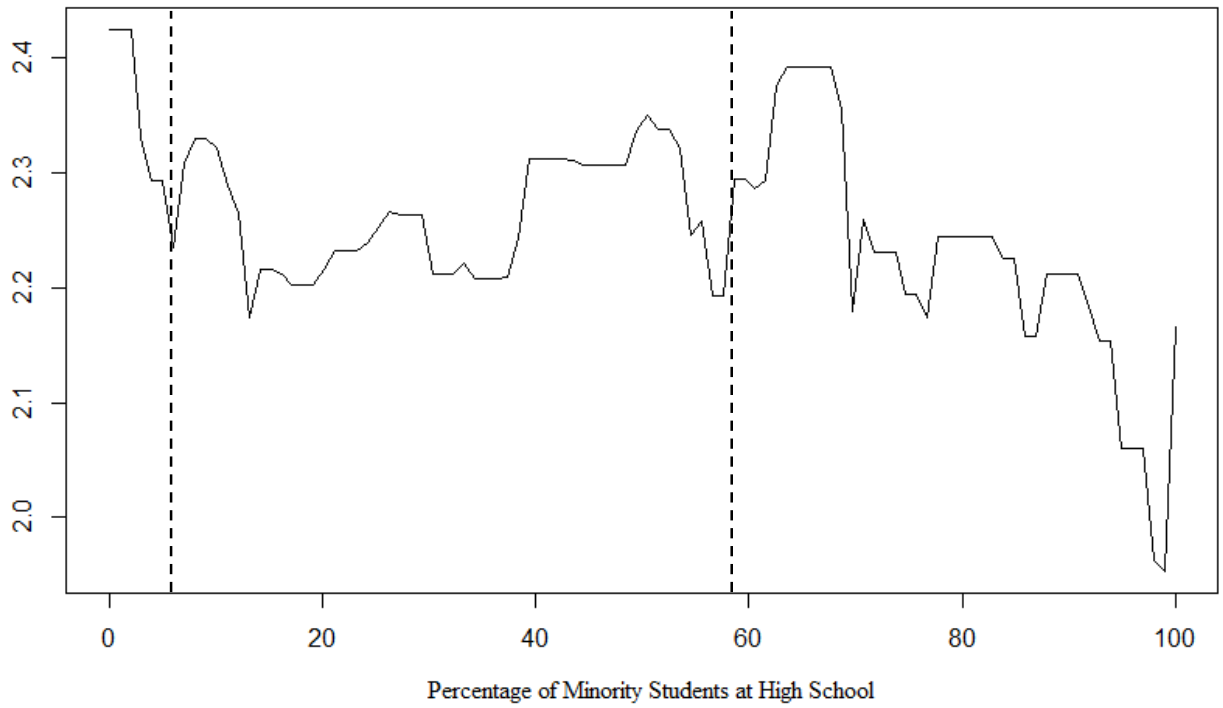

Figure 4.21. Pre-College Model Partial Dependence Plot for Percentage of Minority Students at the Student's High School. The dashed lines represent the middle 80% of the sample.

with students from families making less than $100,000 being the least likely to be retained and students from families exceeding incomes of $300,000 being the most likely to be retained. For the majority of the sample, there was not a clear relationship between percentage of minority students at the student's high school and retention. However, the partial dependence plot showed an overall negative trend for percentage of minority students at the student's high school—predominantly at the extremes. Students from high schools that had few to no minority students were the most likely to be retained, while students from high schools where the vast majority of the student body were minority students were least likely to be retained.

The same process was repeated with the post-first semester variables to create a post-first semester random forest model and a post-first semester boosted model. The gradient tree boosted model was constructed using 410 trees because it was the best cross validation iteration out of 1,000 iterations. The results of the random forest showed that the five most important

variables were GPA differential, earned credit hours, scholarships, family income, and high school GPA. The results from the gradient tree boosting showed that the GPA differential was overwhelmingly the most important variable, followed by earned credit hours and scholarships. These three alone were accounted for in over half of the boosted trees. The variable importance graphs are presented in Figures 4.22 and 4.23 for the random forest and gradient tree boosting, respectively. Each of the most important variables in the post-first semester model paralleled the relationships observed in the HGLM and pre-college tree-based models.

Research Question Five

Confusion matrices were constructed to assess the predictive validity of each model. The overall predictive accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for each model are presented in Table 4.8. The results for the pre-college models show that all four models performed relatively well, with predictive accuracies exceeding 85%. The interaction HGLM performed slightly better than the main effects only model on all measures except for a slightly lower sensitivity. However, the tree-based models outperformed both the main effects only and the interaction effects HGLMs. The overall predictive accuracy of these models was a little over one percent higher than the HGLMs. While the sensitivity was slightly lower than the HGLMs, the positive predictive value was slightly higher. The specificity and negative predictive values were greatly improved in the tree-based models which indicated that they could more correctly predict which students would not return for their second year. The results for the post-first semester models showed a similar trend to the pre-college models, though the differences between the HGLMs and the tree-based models was almost nonexistent. All models predicted retention well, but the random forest model had the highest negative predictive value while the boosted tree model had the highest specificity. Compared to the pre-

college models, the post-first semester models improved the accuracy of prediction and performed better across the various measures.

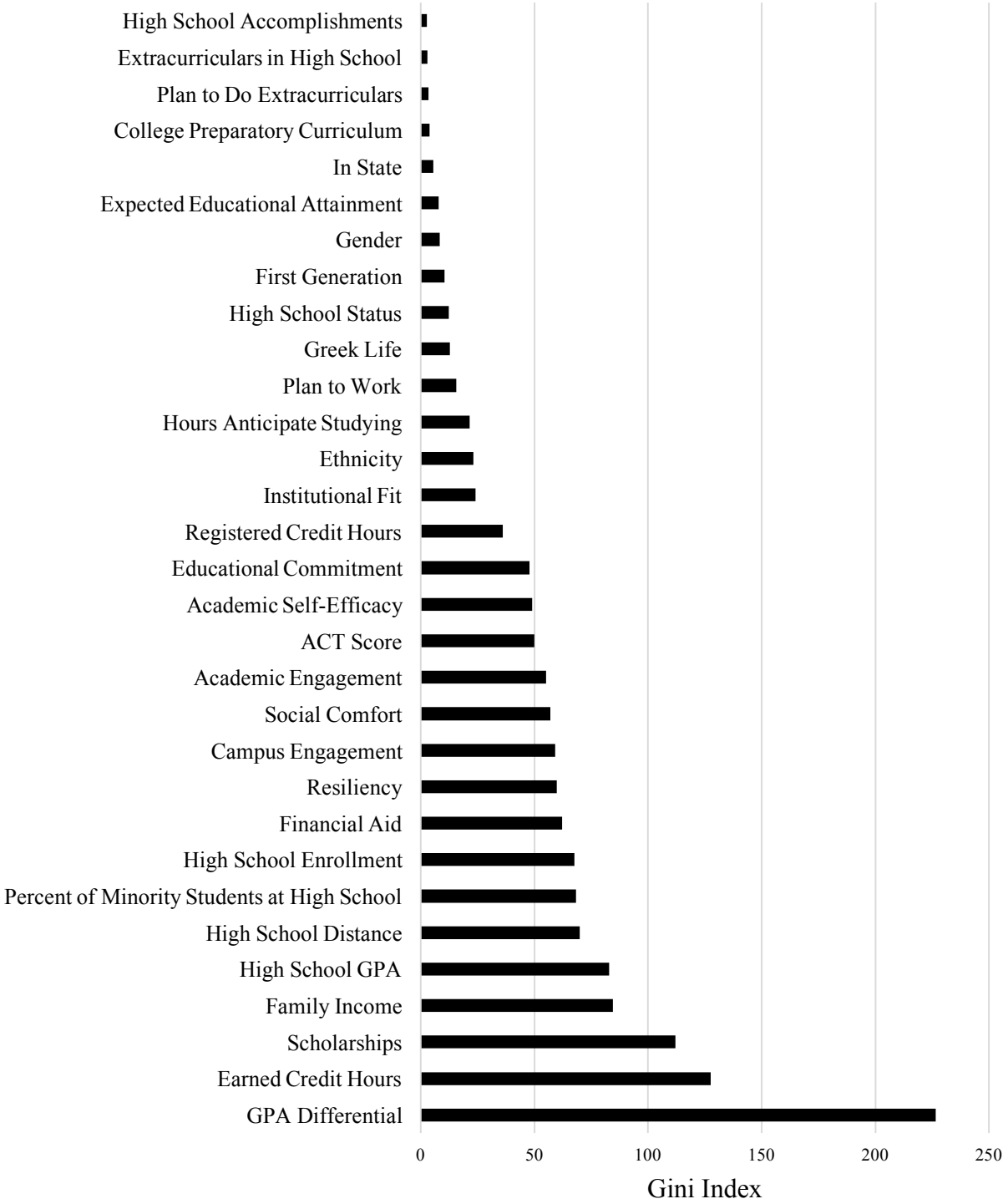

Figure 4.22. Relative Importance of Variables in the Post-First Semester Random Forest Model

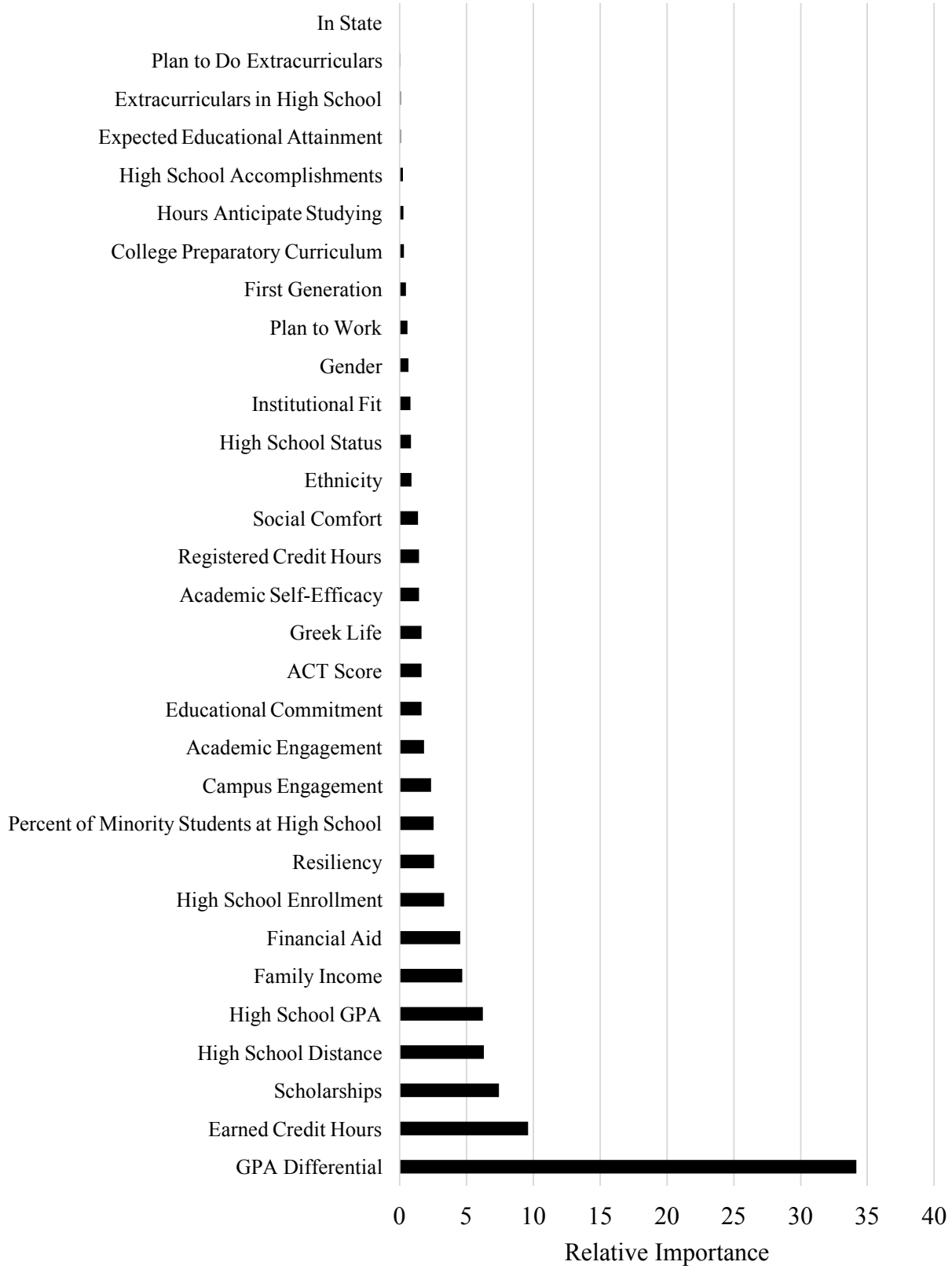

Figure 4.23. Relative Importance of Variables in the Post-First Semester Boosted Tree Model

Table 4.8
Summary of Statistics for Predictive Validity of the Retention Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV
Pre-College Models					
Fixed Effects	.86	.99	.06	.86	.52
Interaction	.86	.99	.09	.86	.53
Random Forest	.87	.98	.20	.88	.68
Boosting	.87	.98	.25	.88	.63
Post-First Semester Models					
Fixed Effects	.88	.98	.28	.89	.72
Interaction	.88	.98	.28	.89	.71
Random Forest	.88	.98	.27	.89	.74
Boosting	.88	.98	.29	.89	.71

Note. PPV = Positive Predicted Value, NPV = Negative Predicted Value

Summary

This study utilized a total sample of 12,342 first-time freshmen at a large, public four-year university in the southeastern region of the United States that was split into a training set consisting of 6,222 students and a test set of 6,120 students. HGLMs and ensemble classification tree-based methods were used to examine which pre-college and post-first semester academic, psychosocial, noncognitive, and background factors were related to first-to-second year retention. In the pre-college models, a variety of factors were found to be significant predictors of retention. The pre-college HGLM results showed that high school GPA, registered credit hours, academic self-efficacy, campus engagement, planning to work in college, institutional fit, gender, first generation status, scholarships, financial aid, high school status (i.e., public or private), and high school enrollment were significantly related to retention outcomes. Some of these relationships were moderated by ethnicity and/or first generation status. The post-first semester HGLM results showed that earned credit hours, GPA differential, and participation in Greek life were significant predictors of retention along with high school GPA, registered credit

hours, campus engagement, institutional fit, gender, scholarships, and high school status. Ethnicity was also a significant moderator of some of the factors in this model.

The ensemble classification tree-based models used were random forest and gradient tree boosting. The results from these models showed that the most important pre-college factors were scholarships, family income, distance from home, high school GPA, and percentage of minority students at the student's high school. These variables remained influential in the post-first semester models, but GPA differential and earned credit hours became the most important factors. Each of these models performed well, accurately predicting retention for nearly 90% of the test set. The tree-based methods performed the best of the pre-college models—especially when it came to correctly predicting the students that would not be retained. The post-first semester models all performed about equally well, and all performed slightly better than the pre-college models.

CHAPTER FIVE
DISCUSSION AND CONCLUSIONS

This chapter summarizes the study and highlights the findings related to each of the five research questions. The relevance of these findings to the literature are also discussed. Additionally, the chapter considers the implications of the study for policy and practice at the secondary and postsecondary levels. The chapter concludes with an overview of the limitations of the study and a discussion of recommendations for future research.

Summary of Study

There were four primary goals of this study: (1) identify the most significant pre-college and post-first semester predictors of first-to-second year retention at a large, public four-year university in the southeastern region of the United States, (2) understand the nature of these relationships, (3) determine if these relationships are moderated by demographic characteristics, and (4) ascertain the predictive validity of the various models developed using these predictors. The results of the study related to each of these goals, along with their relevance to extant literature on retention, are discussed below.

Significant Predictors of Retention and their Relationships

This study sought to identify the most relevant predictors of first-to-second year retention among a set of 28 different pre-college academic, psychosocial, noncognitive, financial, and background characteristics and three post-first semester academic and involvement characteristics. Any relationships moderated by gender, ethnicity, and/or first generation status were also of interest.

Academic characteristics. Among the academic characteristics, the results from the pre-college HGLM showed that high school GPA and registered credit hours were positively related to retention. This was supported by the tree-based models, which also showed high

school GPA to be one of the most important predictors of retention. The literature has consistently supported this finding in studies conducted nationally, statewide, and at various individual institutions across the country (e.g., Geiser & Santelices, 2007; Radunzel & Noble, 2012; Robbins et al., 2004). The results also showed that students who registered for more credit hours were more likely to be retained. This is likely because registering for more credit hours allows students to earn more credit hours, which was also positively related to retention. This is a reasonable result because students who earn more credit hours are considered to be making satisfactory academic progress and consequently are "on track" to complete their degree programs in a timely manner.

ACT score was not found to be a significant predictor in the HGLMs, but it was found to be moderately important in the tree-based methods. This is aligned with previous research at the national level that has found a weaker effect of standardized test scores compared to high school grades (Elder & Radunzel, 2017; Robbins et al., 2004). This finding suggests that the weight that is placed on standardized test scores for admissions and scholarship purposes may be unwarranted when considering a student's propensity to persist to their second year. This does not, however, suggest that standardized test scores are not meaningful predictors of other postsecondary outcomes of interest. The final academic factor, *taking a college preparatory curriculum in high school*, was not significant in either the HGLMs or the tree-based models. This contradicts research on the relationship between high school course rigor and retention (An, 2012; Long et al., 2012). There are two related reasons that potentially explain this finding. First, there was very little variability within the sample as 98% of the students in the training set classified themselves as having taken a college preparatory curriculum in high school. Second, the variable was a student self-reported dichotomous measure. Using a different measure such as

the number or types of college preparatory courses would allow for more variability and a more thorough understanding than the binary measure used in this study.

The trends for these academic characteristics were consistent in the post-first semester models. The major finding from these models was related to the *GPA differential* which was the difference between a student's high school GPA and their first semester college grades. In fact, GPA differential was the most influential variable in all the models. The magnitude of discrepant achievement in either direction significantly impacted the likelihood of retention. Students whose college performance was similar to their high school performance were most likely to be retained. As the differential between these two measures increased, however, the likelihood dropped dramatically. The literature on the importance of first semester academic performance with regards to postsecondary outcomes is extensive. This finding adds to the body of literature by examining this importance not in terms of actual performance, but by college performance relative to high school performance. This result shows that congruent achievement between high school and college is an important factor in student retention.

Psychosocial characteristics. Of the six psychosocial characteristics tested, only two were significant in the pre-college HGLMs. Campus engagement had a positive relationship with retention, while academic self-efficacy had a negative relationship. The positive relationship between campus engagement and retention is supported by prior research on this association (Braxton et al., 2014; Robbins et al., 2004, 2006). This is also supported by the social integration components of the primary theories on retention (Bean & Eaton, 2000; Tinto, 1993). This is a commonsensical finding when considered within this theoretical framework because students who suggest they are eager to participate in campus activities are more likely to integrate to college effectively than students who do not express this same keenness.

The models found a negative relationship between academic self-efficacy and retention, which contradicts the literature, though some studies found that academic self-efficacy was positively related to academic success but not related to retention (Brown et al., 2008; Robbins et al., 2004; Zajacova et al., 2005). The present study is the first to find a negative relationship between academic self-efficacy and retention. The effect found in this study was small, however, and it was not significant in the post-first semester HGLMs. Taken together with the finding related to differential achievement, a possible explanation for this finding is that students with a high degree of academic self-efficacy are more likely to be negatively affected if their college performance is not reflective of their achievement in high school. This could especially be true for students who were in the top tier of students from their high school and/or students who come from high schools where grade inflation is prevalent. These students could develop an inflated academic self-efficacy leading to overconfidence that they are adequately prepared for the realities of college level coursework (Elias & MacDonald, 2007). This conclusion is further supported by the types of items that constituted this scale used in the SSI. Nearly all the questions asked students about their confidence about earning at least a B in various course subjects in college. The other psychosocial factors were not significant predictors in the HGLMs, but some scales such as resiliency and academic engagement were among the important predictors in the tree-based models. This indicates that these other factors were possibly moderated by other variables.

Noncognitive characteristics. Two pre-college noncognitive factors were significant predictors of retention in the pre-college HGLM. The first was *planning to have a job in college*. Students who planned to have a job were less likely to be retained than students who did not plan to have a job or were unsure. It is possible that this variable is moderated by socioeconomic

status as measured by family income; however, this relationship was not examined in this study. It can be assumed that many students expecting to need a job while in school come from low to middle class backgrounds and are unable to fully fund their education and related expenses without some supplementary income. Additionally, some financial aid comes in the form of work study which could also be a contributing factor. The second significant noncognitive variable was *institutional fit* which was positively related to retention. Students who indicated preferences for college characteristics that did not match the characteristics of the institution they attended were less likely to be retained. This relationship is supported by Tinto's (1993) theory of integration. In the post-first semester HGLM, institutional fit remained important while planning to have a job while in college did not. This points to the importance of a student fitting in at their chosen institution, even after their first semester. It is possible that planning to have a job was no longer significant in the post-first semester model because students are more aware of the academic, financial, and time commitments after completing a semester of coursework; therefore, measuring student employment by pre-college intentions to work is no longer appropriate.

Participation in Greek life was also significantly related to retention. The results showed that students who were members of Greek letter organizations were more likely to be retained than students who were not. This is consistent with previous research on retention and membership in fraternities or sororities (DeBard & Sacks, 2011; Nelson, Halperin, Wasserman, Smith, & Graham, 2006). One possible explanation for this result is that Greek organizations provide an immediate sense of belonging at the university for first year students which is a critical component in theoretical models of retention (Branand, Mashek, Wray-Lake, & Coffey, 2015). A second explanation is that these organizations have support structures in place for their

members that potentially aid students be academically successful. Some of these structures include mandatory study hours and resource sharing from upperclassmen familiar with the classes and instructors.

Financial characteristics. The most significant factor in the pre-college models was the amount of *scholarships* a student was awarded. This variable was positively related to retention overall; however, this relationship was moderated by ethnicity. The results for both the pre-college and post-first semester models showed that the odds Hispanic students would be retained decreased compared to White students as the amount of scholarships increased. This is possibly explained by the fact that the distribution of scholarships was similar between White and Hispanic students up to about $10,000, and then the proportion of White students who received larger scholarship awards was greater than the proportion of Hispanic students. Consequently, this observed effect is conceivably a function of access/opportunity for large scholarships for Hispanic students compared to White students more so than it is of scholarships themselves. There were not significant differences in the relationship of scholarships and the odds of retention for other ethnic groups compared to White students. This relationship was also moderated by whether a student was a first generation student because additional scholarship dollars increased the odds of retention for first generation students more than it did for non-first generation students. Since first generation status can be used as a proxy for socioeconomic status (Liu et al., 2004), this finding is logical since larger scholarships decrease the financial burden associated with a college education. This was only true for the pre-college models, however, as first generation status was no longer a significant predictor after the student's first semester performance was taken into account which will be discussed further in the following section on background characteristics.

Financial aid was a significant predictor in the pre-college HGLMs, but it was moderated by first generation status. The results showed that larger amounts of financial aid were related to better odds of retention for first generation students, while this effect was considerably smaller for non-first generation students. The tree-based models showed that financial aid was one of the most important variables in both the pre-college and post-first semester models, but it was not as important as the amount of scholarships a student received. One possible explanation for this is that the financial burden posed by student loans are not realized in a student's first semester of college, so the impact of taking out loans is not as noticeable as the impact of having scholarships which can cover tuition, books, and housing. The relative importance of financial aid suggests that it possibly interacts with some of the other variables in the model; future research should investigate this further.

Interestingly, family income was not a significant predictor in any of the HGLMs. Previous research has found that students' socioeconomic status is significantly related to college outcomes (J. Allen et al., 2008; Lotkowski, Robbins, & Noeth, 2004; Robbins et al., 2004). This was, however, reflected in both the pre-college and post-first semester tree-based models. The results showed that for the majority of students, increased family income was related to increased odds of retention. The discrepancy between models can be attributed to the way in which family income was measured from the FAFSA. For example, a student could file their FAFSA as an independent with no parental contribution towards the captured annual income which allows for measurement error at the extremes. The constraints of HGLMs made it so these models were unable to account for this anomaly, but close examination of the partial dependence plots in the tree-based models captured a more realistic representation of socioeconomic status and retention and showed a strong positive relationship overall.

Background characteristics. *Gender* and *ethnicity* were significant predictors of retention that moderated one another in both models. These results were somewhat contrary to existing literature on retention. Most studies have found that females are more likely to be retained than male students or that there was not a significant difference between the two genders (J. Allen et al., 2008; DeBerard et al., 2004; Kuh et al., 2008; St. John et al., 2001). Similarly, studies have found that minority students oftentimes underperform White and Asian students (Arbona & Nora, 2007; Kao & Thompson, 2003; Porchea et al., 2010; Reason, 2003). The present study found that the relationship between these characteristics and retention were interdependent which is also consistent with some studies in the literature (D. Allen, 1999; Attewell et al., 2010). Specifically, White and African American males had higher odds of retention than their female counterparts, while Asian females and females from other, non-Hispanic ethnic groups had higher retention odds than their male counterparts. The pre-college HGLM showed this difference was also significant for Hispanic females. This can potentially be explained by the fact that other studies did not utilize the types of comprehensive, reliable data sources that were utilized in this study. Thus, the models in this study were able to account for a variety of factors beyond demographic characteristics that previous studies could not. This finding can also be reinforced by Tinto (1993) who suggested that female attrition is "more determined by social forces than academic ones" compared to males (p. 73). Therefore, the effect observed could be related to unmeasured social variables after accounting for the various other factors in the model. It is also worth noting that this finding could be unique to the context in which this study was conducted and could differ at other institutions and/or regions of the country.

Two high school factors were significantly related to retention in the HGLMs. The results showed that students from private schools had better odds of retention than students from public schools. It is possible that this is indicative of high school quality and consequently the quality of preparation for college. However, it is also highly likely that this is reflective of socioeconomic status and the quality of the student's high school resources. The second significant high school factor in the HGLMs was *enrollment size*, though this was only significant in the pre-college model and it was moderated by first generation status. Non-first generation students from larger high schools were more likely to be retained, while the increase in retention likelihood for first generation students from larger high schools was significantly smaller. This can be explained by the easier transition from a large high school to a large institution. If students are accustomed to being "just another face" at a large high school, then it is plausible that they would adjust better at a large university than students from small high schools where they knew much of the student body. This effect would be magnified for first generation students who also have additional adjustments to make that non-first generation students might not. Furthermore, first generation students are more likely to come from under-resourced high schools. This makes it possible for these students to not have the type of advising and support that are more readily available at large, higher-resourced schools, thereby minimizing the advantage provided by attending a large high school and its effect on retention. This importance of high school characteristics was also supported in the literature by I. Johnson (2008).

In addition to high school type and enrollment size, the tree-based models suggested that a variety of other high school characteristics were important predictors of retention—especially the distance of the high school from the university. The results showed that as this distance

increased, the likelihood a student would be retained decreased. This can be explained in multiple ways. First, students who attend college further from home incur additional living expenses that students who attend a college near their high school may not sustain. The cost of college could also be higher for students with large distances because, at a certain point, they would have to pay out-of-state tuition. These increased financial burdens could encourage students to leave the institution for more affordable options (Pryor, Hurtado, DeAngelo, Palucki Blake, & Tran, 2009). A second explanation is that students from high schools closer to the college are more familiar with the area, are nearer to support in the form of family and friends, and will have a higher percentage of their peers also enrolled at the institution. All of these factors would make the integration into the college community easier by minimizing the potential negative effects of homesickness (Fisher & Hood, 1987). The other important high school characteristic was *percentage of minority students*. The results showed that students from extremely low minority high schools had higher odds of retention while students from high minority high schools were least likely to be retained. This suggests that the effect of attending a predominately White institution had negative effects on student integration for students from predominantly minority high schools, likely because of the incongruence between the high school's ethnic composition with the university's.

First generation status was a significant moderator of multiple factors in the pre-college model, but it was not significant in the post-first semester model. First generation status moderated the relationships between scholarships, financial aid, and high school enrollment in the pre-college models. Each of these moderated relationships have been previously discussed. A potential reason that first generation status was no longer significant in the post-first semester models is that there usually are not many structured programs built to specifically assist first

generation students. Before one of these students begin college, they oftentimes have a steeper learning curve than their non-first generation peers. This distinction likely becomes blurred after the first semester of college which would reduce the observed effects of being a first generation student on retention.

Predictive Validity of Retention Models

The literature on retention has consistently suggested that a student's academic performance and social activity in their first year of college are the strongest predictors of first-to-second year retention (I. Y. Johnson, 2006, 2008; Kuh et al., 2008). The results of this study corroborate these findings. However, the results of this study also showed that it is possible to predict retention outcomes before a student begins classes with roughly the same accuracy as predictions after the student completes their first semester. All four post-first semester models correctly predicted retention for 88% of the students. The pre-college models were only one to two percent less accurate than the post-first semester models. This shows that, while college performance is important, high school factors and background characteristics are extremely meaningful predictors of retention and can predict it nearly as well as models incorporating college performance and activity.

While all the models had similar predictive accuracies, there were also important differences between them. Both the pre-college and post-first semester models had high sensitivity values which suggested that any of the models successfully predicted the students who would be retained. All the models were not as successful in predicting the students who would not be retained, however. Among the pre-college models, the results showed that the tree-based models were two to four times more successful in predicting the students who would not be retained than the HGLMs. This is important because ultimately the goal for universities is to

predict the students that will *not* be retained—not necessarily predicting the students that will. Comparing the specificities of the pre-college tree-based models with the specificities of the post-first semester models, the models in this study were able to correctly predict the students who would not be retained before they began college with nearly the same accuracy as the models after the first semester. Bearing this in mind, multiple predictive models should be used when predicting retention outcomes in order to have more accurate and robust predictions (Delen, 2010).

Implications for Policy and Practice

There are several findings from this study that have important implications for policy and practice not only in higher education, but secondary education as well. The first implication for practice is related to the potential power of using rich predictive models of student retention. The use of these predictive models has important implications for retention efforts at colleges and universities. Variable-rich predictive models could be used to identify students at risk of not persisting at the institution so that these students could be targeted for specific interventions tailored to their needs. For example, if a student was predicted not to be retained and their profile suggested this occurred due to academic concerns, advising and counseling staff could provide them with information about tailored support services and interventions that could benefit them such as academic support services. The ability to identify areas that are potential barriers to a student acclimating and becoming successful at their university is critically important as university administrators and student academic advisors seek to deter students from leaving college during or after their first year.

To have a robust prediction of student retention outcomes, the models used to identify students at risk of not being retained should draw from rich, comprehensive data sources.

Additionally, comparing results across multiple models such as the ones in this study could improve the accuracy of prediction. The present study showed that combining multiple data sources to create rigorous predictive models resulted in robust forecasts of retention outcomes. This also has implications for policy. It is incumbent upon secondary and postsecondary institutions as well as policymakers at the state and federal level to investigate ways to bridge secondary level data to the postsecondary institutions, thus creating a comprehensive pre-college and collegiate profile of each student. The postsecondary institutions can then utilize a comprehensive profile for each student which considers cognitive, psychosocial, noncognitive, and background measures. In this way, colleges and universities can more proactively anticipate barriers to student success and retention.

The second implication stems from the importance of the magnitude of high school/college GPA differential in the post-first semester models. The results of this study showed that students whose academic performance in college differs notably from their academic performance in high school were significantly less likely to be retained than students whose performance was consistent. Coupling this with the fact that 98% of the students in the sample said their high school curriculum was college preparatory in nature, there seems to be a problematic disconnect between college preparatory courses at the high school level and the rigors of actual college level coursework for some students. This has major implications for initiatives that encourage early college options such as dual enrollment and advanced placement courses. Policymakers should put policies and procedures in place that ensure these programs are implemented with a degree of fidelity that is reflective of college level courses. In the case of dual enrollment courses, for example, colleges and high schools should work closely together

to ensure that the content, delivery, and assessments are in alignment so that these classes are truly parallel to college level coursework.

This result also has important implications for practice. There is growing momentum for the idea that students graduating from high school should be college ready and therefore should not require remediation upon enrolling at a postsecondary institution (Venezia & Jaeger, 2013). The costs of remediation in the first year of college are high. Remedial coursework costs students approximately $1.5 billion per year and decreases their likelihood of college success and completion (Nguyen Barry & Dannenberg, 2016). Consequently, remediation should be identified and resolved prior to the student graduating from high school. Colleges and universities also need to have clear plans in place for addressing the needs of students who are unprepared when they matriculate. In order to be successful in this endeavor, colleges need to work closely with high schools so that both entities are informed of expectations and can more fully understand students' levels of academic preparation. An extension of this implication also applies to instruction in both secondary and postsecondary settings. Faculty members at colleges need to be well trained in pedagogy, and high school teachers need to ensure they are preparing students for the higher order thinking skills that tend to be emphasized in college. This is especially true in college preparatory courses where a rapid pace of delivery can be misconstrued as rigor.

A third implication is related to the significance of the institutional fit score and the distance of the university from students' homes. The results of this study suggest that as high school guidance counselors advise students on their college planning, they should focus conversations on what characteristics students are looking for in an institution. Understanding the institutional characteristics that match a student's preferences and helping the student to find

institutions fitting this profile will better position the student for success once they start college because they are more likely to integrate without difficulty. High school guidance counselors should also include the distance of the institutions in these conversations. Attending a college that is further from a student's home lowers their odds of retention on average, so having conversations about whether moving far from home is best for an individual student could help students to better gauge how distance could affect their ability to acclimate to postsecondary education. The significance of these measures also has important implications for colleges and universities. Student affairs practitioners should provide programs specifically designed to help students from long distances from the university assimilate effectively to the campus and community. Additionally, similar services and programs would be beneficial for students whose profile indicates they have poor institutional fit.

A fourth implication of this study is the importance of scholarships in college student retention. This study suggests that state funded scholarship programs such as the Taylor Opportunity Program for Students (TOPS) in Louisiana, Helping Outstanding Pupils Educationally (HOPE) in Georgia, and the Bright Futures Scholarship Program in Florida should be bolstered and expanded. These programs have been the victims of state budget cuts which have limited the money that students receive towards their college education by reducing the payout of these awards and/or increasing the minimum requirements for eligibility in order to reduce the number of scholarships awarded (McGlade & Travis, 2016; Raines, 2013; Sentell, 2017). The results of this study show that cutting these programs could have damaging effects on student persistence at colleges and universities. An additional unintended consequence of these reductions is that students will have to depend on other forms of financial aid to fund their education. This could have an adverse effect on all students, but it could especially be

detrimental to students from disadvantaged backgrounds such as first generation students. Consequently, the importance of both need- and merit-based scholarships must be an important factor in conversations around higher education policy and funding. At the institutional level, colleges would benefit from prioritizing scholarship awards to students in their budgets. This is especially true for students from disadvantaged and minority backgrounds.

Limitations of the Study

There were limitations worth noting in this study. The first is that the results are not necessarily generalizable to other colleges and universities—especially not to institutions that differ significantly in student population, geographic region, and/or mission from the university in this study. Therefore, the methods and findings in this research should be used as guide for replication by other institutions and not generalized to different contexts. A second limitation of this study was that it was unable to differentiate between students who transferred, students who stopped attending but later reenrolled (sometimes referred to as "stopouts"), and students who dropped out of college altogether. The factors related to each of these groups of students who are not retained could be different, and these differences were not be captured in this study. Therefore, this study should only be examined primarily from an institutional perspective and not interpreted from an individual student's perspective in the context of overall college persistence, regardless of institution. Third, the definitions and construction of some of the variables of interest limited the potential meaningfulness of these variables in the models. For example, the dichotomous variable indicating whether a student took a college preparatory curriculum in high school did not provide enough variability to be meaningful for this sample. This variable would have been better measured as a quantity of college preparatory classes taken, not a student self-reported indicator of overall high school curriculum type.

An additional limitation for the study was the instrument used to measure the psychosocial factors (i.e., the SSI). The instrument measured six different psychosocial factors that had high reliability. However, the results of the study were not consistent with the literature on some of these factors. Further investigation into the items revealed that there were potential validity issues. For example, the academic self-efficacy scale consisted of items that solely measured students' confidence in their ability to earn good grades. This reflects construct underrepresentation because the widely accepted definition of academic self-efficacy throughout the literature is more broadly defined. This consequently threatened the validity of the interpretation of this construct.

Future Research

There are multiple opportunities for future research based on the results of this study. One of the most promising areas for further investigation is exploring additional high school factors in relation to retention. High school characteristics were some of the most important variables in the predictive models in this study, and a sizeable percentage of the variance in retention outcomes was explained by high school differences. Collecting additional information about high school characteristics and including it in the predictive models from this study could greatly improve the accuracy of prediction and provide meaningful information that could be useful in informing various services that universities provide to students. Variables used in this study were focused on demographic characteristics of high schools. Future research should explore additional demographic characteristics as well as other types of traits—especially academic characteristics, given the importance of students' academic performance in the predictive models.

Future research should also thoroughly examine additional interactions that were not explored in this study. Several variables were noted as highly influential in the tree-based models that were not significant in the HGLMs, suggesting possible significant interaction effects that were not investigated in this study. Examples of variables that should be studied include socioeconomic status and other financial factors as well as additional student characteristics such as residency status and academic factors. Examining these relationships in the context of both HGLMs and ensemble tree-based models will provide a richer understanding of the interplay between the numerous variables related to college student retention, thereby better informing university initiatives and services designed to improve college student success. Moreover, the importance of the financial characteristics in the models points to the need for further investigation of these variables. Differentiating between the types of scholarships and financial aid (e.g., merit-based versus need-based, loans versus grants) could provide meaningful insight into the strong relationship financial characteristics had with retention. Similarly, the role of the psychosocial factors explored in this study as well as additional ones in the literature should be further explored—especially academic self-efficacy. This could be done in the form of interactions as well as measuring these constructs using different instruments such as the SRI from the Robbins et al. (2006) study and comparing the results to the scales from the SSI. The use of different instruments would allow for an assessment of the validity of the scales measured on the SSI.

Finally, this research should be replicated and expounded upon at other institutions. The findings from this study are highly context specific and should not necessarily be generalized to other colleges and universities. There are specific institutional and state policies and characteristics that could influence the variables that were important in the models as well as the

nature of the relationships of the variety of variables with retention. The methods utilized in this study should therefore be replicated at other institutions to determine the best models for predicting retention at different colleges. Additionally, these replicated studies could also improve on the methods in this study by incorporating additional factors like the ones described throughout this section.

Conclusion

This study used a comprehensive framework to examine academic, psychosocial, noncognitive, financial, and background characteristics that were related to retention at a large, public four-year institution in the Southeast. Data were collected from multiple sources to create predictive models that drew from a variety of pre-college factors. The models were built using a variety of statistical techniques including hierarchical generalized linear models and ensemble tree-based methods. The results showed significant relationships of various factors with retention, the nature of these relationships, moderated effects, and variable importance.

The results of this study demonstrated that considering a variety of factors when forecasting postsecondary retention outcomes is vital for more accurate, comprehensive predictions. The relationships revealed in this study can be used to help inform programs and services aimed at improving student success and retention. They provide important implications for a variety of issues in higher education policy—especially in regards to funding and programming. The study also showed that valid prediction models can be built using only high school level and background characteristics. This is an important tool that can aid administrators and practitioners in higher education to identify students at risk of not persisting to their second year at the institution early in their first semester of college and subsequently tailor interventions to individual student needs.

REFERENCES

ACT. (2014). *The ACT technical manual*. Iowa City, IA.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, D.C. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.495

Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence. *Research in Higher Education*, *40*(4), 461–485.

Allen, J., & Robbins, S. (2010). Effects of interest-major congruence, motivation, and academic performance on timely degree attainment. *Journal of Counseling Psychology*, *57*(1), 23–35. http://doi.org/10.1037/a0019085

Allen, J., Robbins, S. B., Casillas, A., & Oh, I.-S. (2008). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, *49*(7), 647–664. http://doi.org/10.1007/s11162-008-9098-3

An, B. P. (2012). The Influence of Dual Enrollment on Academic Performance and College Readiness: Differences by Socioeconomic Status. *Research in Higher Education*, *54*(4), 407–432. http://doi.org/10.1007/s11162-012-9278-z

An, B. P. (2015). The Role of Academic Motivation and Engagement on the Relationship Between Dual Enrollment and Academic Performance. *The Journal of Higher Education*, *86*(1), 98–126.

Arbona, C., & Nora, A. (2007). The influence of academic and environmental factors on hispanic college degree attainment. *The Review of Higher Education*, *30*(3), 247–269. http://doi.org/10.1353/rhe.2007.0001

Arbuthnot, K. (2011). *Filling in the blanks: Understanding standardized testing and the black-white achievement gap*. Charlotte, NC: Information Age Publishing, Inc.

Astin, A. W. (1977). *Four critical years: Effects of college on beliefs, attitudes, and knowledge*. San Francisco, CA: Jossey-Bass Publishers.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Development*, *25*, 297–308.

Astin, A. W., & Oseguera, L. (2005). Pre-college and institutional influences on degree attainment. In A. Seidman (Ed.), *College Student Retention: Formula for Student Success* (pp. 245–276). Westport, CT: Praeger.

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665–676.

Attewell, P., Heil, S., & Reisel, L. (2010). Competing explanations of undergraduate noncompletion. *American Educational Research Journal*, *48*(3), 536–559. http://doi.org/10.3102/0002831210392018

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Worth.

Bean, J. P. (2005). Nine themes of college student retention. In A. Seidman (Ed.), *College Student Retention: Formula for Student Success* (pp. 215–244). Westport, CT: Praeger.

Bean, J. P., & Eaton, S. B. (2000). A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the student departure puzzle*. Nashville, TN: Vanderbilt University Press.

Bean, J. P., & Eaton, S. B. (2002). The psychology underlying successful retention practices. *Journal of College Student Retention*, *3*(1), 73–89.

Berger, J. B., & Lyon, S. C. (2012). Past to present: A historical look at retention. In A. Seidman (Ed.), *College Student Retention: Formula for Student Success* (2nd ed., pp. 7–34). Lanham, MD: Rowman & Littlefield.

Berglund, P. A. (2015). *Multiple imputation using the fully conditional specification method: A comparison of SAS, Stata, IVEware, and R*. Cary, NC.

Branand, B., Mashek, D., Wray-Lake, L., & Coffey, J. K. (2015). Inclusion of college community in the self: A longitudinal study of the role of self-expansion in students' satisfaction. *Journal of College Student Development*, *56*(8), 829–844. http://doi.org/10.1353/csd.2015.0080

Braxton, J. M., Doyle, W. R., Hartley, H. V., Hirschy, A. S., Jones, W. A., & McLendon, M. K. (2014). *Rethinking college student retention*. San Francisco, CA: Jossey-Bass Publishers.

Bridgeman, B., Burton, N., & Pollack, J. (2008). Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches. *Journal of College Admission*, (199), 19–25. http://doi.org/Article

Briggs, D. C. (2009). *Preparation for college admission exams*. Arlington, VA. Retrieved from http://www.solutionsbayarea.com/TestPrepDiscussionPaper.pdf

Brown, S. D., Tramayne, S., Hoxha, D., Telander, K., Fan, X., & Lent, R. W. (2008). Social cognitive predictors of college students' academic performance and persistence: A meta-analytic path analysis. *Journal of Vocational Behavior*, *72*(3), 298–308. http://doi.org/10.1016/j.jvb.2007.09.003

Burns, S., Wang, X., & Henning, A. (2011). *NCES handbook of survey methods* (2nd ed.). Washington, D.C.: National Center for Education Statistics.

Burrus, J., Elliott, Di., Brenneman, M., Markle, R., Carney, L., Moore, G., … Roberts, R. D. (2013). *Putting and keeping students on track: Toward a comprehensive model of college persistence and goal attainment*. Princeton, NJ.

Cabrera, A. F., Nora, A., & Castañeda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education*, *64*(2), 123–139.

Campus Labs. (2012a). *Theoretical and statistical foundations of the Student Strengths Inventory*. Buffalo, NY.

Campus Labs. (2012b). Using the non-cognitive factors of Beacon in advising. Buffalo, NY: Campus Labs, Inc. Retrieved from http://beaconsupport.campuslabs.com/hc/en-us/article_attachments/201127608/Using_the_Non-cognitive_Factors_of_Beacon_in_Advising.pdf

Carnevale, A. P., Smith, N., & Strohl, J. (2013). *Recovery: Projects of jobs and education requirements through 2020*. Washington, DC.

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, *104*(2), 407–420. http://doi.org/10.1037/a0027180

Choy, S. (2001). *Students whose parents did not go to college: Postsecondary access, persistence, and attainment*. Washington, D.C.

Clinedinst, M., Koranteng, A.-M., & Nicola, T. (2015). *2015 state of college admission*. Arlington, VA.

Conley, D. T. (2007). *Redefining college readiness*. Eugene, OR. Retrieved from www.epiconline.org

Creswell, J. W. (2015). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (5th ed.). Boston, MA: Pearson Education, Inc.

Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology*, *101*(4), 948–963.

DeBard, R., & Sacks, C. (2011). Greek membership: The relationship with first-year academic performance. *2Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 109–126.

DeBerard, M. S., Spielmans, G. I., & Julka, D. C. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal*, *38*(1), 66–80.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores*. New York, NY.

Elder, A. C., & Radunzel, J. (2017). Multidimensional models of who goes to college and who stays. In *Association for Institutional Research Annual Forum*. Washington, D.C.

Elias, S. M., & MacDonald, S. (2007). Using past performance, proxy efficacy, and academic self-efficacy to predict college performance. *Journal of Applied Social Psychology*, *37*(11), 2518–2531. http://doi.org/10.1111/j.1559-1816.2007.00268.x

Ene, M., Leighton, E. A., Blue, G. L., & Bell, B. A. (2015). *Multilevel models for categorical data using SAS PROC GLIMMIX: The basics*. Cary, NC.

Fisher, S., & Hood, B. (1987). The stress of the transition to university: A longitudinal study of psychological disturbance, absent-mindedness and vulnerability to homesickness. *British Journal of Psychology*, *78*(4), 425–441.

Gallagher, A. M., & Kaufman, J. C. (Eds.). (2005). *Gender differences in mathematics: An integrative psychological approach*. Cambridge, UK: Cambridge University Press. http://doi.org/doi:10.1017/CBO9780511614446.013

Geiser, S., & Santelices, M. V. (2004). *The role of advanced placement and honors courses in college admissions*. Berkeley, CA.

Geiser, S., & Santelices, M. V. (2007). *Validity of high school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes*. Berkeley, CA.

Godfrey, K., Matos-Elefonte, H., Ewing, M., & Patel, P. (2014). *College completion: Comparing AP, dual-enrolled and non-advanced students*. New York, NY.

Gore, P. A. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*, *14*(1), 92–115.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. http://doi.org/10.1007/b94608

Howard, C. (2016). Top colleges ranking 2016: The full methodology. Retrieved October 17, 2016, from http://www.forbes.com/sites/carolinehoward/2016/07/06/top-colleges-ranking-2016-the-full- methodology/#3dbe06ed59a8

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. http://doi.org/10.1007/978-1-4614-7138-7

Johnson, I. Y. (2006). Analysis of stopout behavior at a public research university: The multi-spell discrete-time approach. *Research in Higher Education*, *47*(8), 905–934. http://doi.org/10.1007/s11162-006-9020-9

Johnson, I. Y. (2008). Enrollment, persistence and graduation of in-state students at a public research university: Does high school matter? *Research in Higher Education*, *49*(8), 776–793. http://doi.org/10.1007/s11162-008-9105-8

Johnson, R. B., & Christensen, L. (2014). *Educational research: Quantiative, qualitative, and mixed approaches* (5th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*(2003), 417–442. http://doi.org/10.1146/annurev.soc.29.010202.100019

Kena, G., Hussar, W., McFarland, J., de Brey, C., Musu-Gillette, L., Wang, X., … Dunlop Velez, E. (2016). *The condition of education 2016 (NCES 2016-144)*. Washington, DC.

Kim, J. (2015). Predictors of college retention and performance between regular and special admissions. *Journal of Student Affairs Research and Practice*, *52*(1), 50–63.

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences*, *25*, 67–72.

Komarraju, M., Ramsey, A., & Rinella, V. (2013). Cognitive and non-cognitive predictors of college readiness and performance: Role of academic discipline. *Learning and Individual Differences*, *24*, 103–109.

Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, *22*(4), 497–519. http://doi.org/10.1093/pan/mpu007

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, *79*(5), 540–563.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*(1), 63–82. http://doi.org/10.3102/00346543075001063

Lin, S., & Sanchez, E. I. (2017). Effects of discrepant ACT and high school GPA on enrollment and persistence. In *American Educational Research Association*. San Antonio, TX.

Liu, W. M., Ali, S. R., Soleck, G., Hopps, J., Dunston, K., & Pickett, T. (2004). Using social class in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 3–18. http://doi.org/10.1037/0022-0167.51.1.3

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, *49*(2), 285–322. Retrieved from http://aer.sagepub.com/cgi/doi/10.3102/0002831211431952

Lotkowski, V. A., Robbins, S. B., & Noeth, R. J. (2004). *The role of academic and non-academic factors in improving college retention*. *ACT*. Iowa City, IA.

Luthar, S. S., & Cicchetti, D. (2000). The construct of resilience: implications for interventions and social policies. *Development and Psychopathology*, *12*(4), 857–885. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1903337&tool=pmcentrez&rendertype=abstract

Mattern, K. D., Burrus, J., Camara, W. J., O'Connor, R., Hanson, M. A., Gambrell, J., … Bobek, B. (2014). *Broadening the definition of college and career readiness: A holistic approach*. Iowa City, IA.

Mattern, K. D., & Patterson, B. F. (2009). *Is performance on the SAT related to college retention? (Research Report 2009-7)*. New York, NY. Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2009-7-is-performance-on-sat-related-college-retention.pdf

Mattern, K. D., & Patterson, B. F. (2010). *The relationship between SAT scores and retention to the third year: 2006 cohort*. New York, NY.

Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2010). Academic fit: Is the right school the best school or is the best school the right school? *Journal of Advanced Academics*, *21*(3), 368–391.

Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2011). An alternative presentation of incremental validity: Discrepant SAT and HSGPA performance. *Educational and Psychological Measurement*, *71*(4), 638–662. http://doi.org/10.1177/0013164410383563

Mattern, K. D., Wyatt, J. N., & Shaw, E. J. (2013). College distance from home: Implications for student transfer behavior. *Journal of The First-Year Experience and Students in Transition*, *25*(1), 77–92.

McGlade, C. R., & Travis, S. (2016, October 21). Minorities, poor hit hardest by stricter Bright Futures requirements. *Sun Sentinel*. Miami, FL. Retrieved from http://www.sun-sentinel.com/news/education/sfl-minorities-poor-hit-hardest-by-bright-futures-changes-20161021-htmlstory.html

Morse, R., Brooks, E., & Mason, M. (2016). How U.S. News calculated the 2017 best colleges rankings. Retrieved October 17, 2016, from http://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings

Munt, J. A., & Merydith, S. P. (2012). The relationship of students' personality traits and psychosocial characteristics with academic retention. *Journal of College Student Retention: Research, Theory and Practice*, *13*(4), 457–478.

Nagaoka, J., Farrington, C. A., Roderick, M., Allensworth, E., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2013). Readiness for college: The role of noncognitive factors and context. *Voices in Urban Education*, (38), 45–52.

Nardi, P. M. (2006). *Doing survey research: A guide to quantiative methods* (2nd ed.). Boston, MA: Pearson Education, Inc.

Nelson, S. M., Halperin, S., Wasserman, T. H., Smith, C., & Graham, P. (2006). Effects of fraternity/sorority membership and recruitment semester on GPA and retention. *Oracle: The Research Journal of the Association of Fraternity Advisors*, *2*(1), 61–73. Retrieved from http://c.ymcdn.com/sites/www.afa1976.org/resource/collection/0E038C73-3450-45DA-A3B4-C5C64D5ED39B/V2_Iss1_Oracle_Feb_2006.pdf

Nguyen Barry, M., & Dannenberg, M. (2016). *Out of pocket: The high cost of inadequate high schools and high school student achievement on college affordability*. Washington, D.C.

Nieuwenhuis, J., Hooimeijer, P., & Meeus, W. (2015). Neighbourhood effects on educational attainment of adolescents, buffered by personality and educational commitment. *Social Science Research*, *50*, 100–109. http://doi.org/10.1016/j.ssresearch.2014.11.011

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199–242). Charlotte, NC: Information Age Publishing, Inc.

Pascarella, E. T., & Chapman, D. W. (1983). A multiinstitutional, path analytic validation of Tinto's model of college withdrawal. *American Educational Research Journal*, *20*(1), 87–102.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *The Journal of Higher Education*, *75*(3), 249–284.

Pascarella, E. T., & Terenzini, P. T. (1998). Studying college students in the 21st century: Meeting new challenges. *Review of Higher Education*, *21*(2), 151–165.

Paulsen, M. B., & St. John, E. P. (2002). Social class and college costs. *The Journal of Higher Education*, *73*(2), 189–236.

Pew Research Center. (2014). *The rising cost of not going to college*. Washington, DC.

Porchea, S. F., Allen, J., Robbins, S., & Phelps, R. P. (2010). Predictors of long-term enrollment and degree outcomes for community college students: Integrating academic, psychosocial, socio-demographic, and situational factors. *The Journal of Higher Education*, *81*(6), 680–708.

Pryor, J. H., Hurtado, S., DeAngelo, L., Palucki Blake, L., & Tran, S. (2009). *The American freshman: National norms fall 2009*. Los Angeles, CA: Higher Education Research Institute, UCLA. Retrieved from http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf

Radunzel, J., & Noble, J. (2012). *Predicting long-term college success through degree completion using ACT composite score, ACT benchmarks, and high school grade point average. ACT Research Report Series, 2012 (5)*. Iowa City, IA. Retrieved from http://ezproxy.umsl.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED542027&site=ehost-live&scope=site

Raines, L. (2013, March 1). More changes proposed to keep HOPE alive. *Atlanta Journal Constitution*. Atlanta, GA. Retrieved from http://www.ajc.com/business/more-changes-proposed-keep-hope-alive/z5bg20LDrlL5yRdHo4NPiL/

Raisman, N. A. (2013). *The cost of college attrition at four-year colleges and universities*. Virginia Beach, VA. Retrieved from http://www.educationalpolicy.org/pdf/1302_PolicyPerspectives.pdf

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*(1), 141–157.

Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments. *NASPA Journal*, *40*(4), 172–191. http://doi.org/10.2202/1949-6605.5022

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. http://doi.org/10.1037/a0026838

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology*, *98*(3), 598–616.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14979772

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*(6), 1479–1497. http://doi.org/10.1037/a0016810

Schneider, M. (2010). *Finishing the first lap: The cost of first-year student attrition in America's four-year colleges and universities*. Washington, DC. Retrieved from http://www.air.org/files/AIR_Schneider_Finishing_the_First_Lap_Oct101.pdf

Sentell, W. (2017, January 9). Permanent cuts to TOPS , more hours per year recommended in new Louisiana Board of Regents report. *The Advocate*. Baton Rouge, LA. Retrieved from http://www.theadvocate.com/baton_rouge/news/education/article_91b3aee4-d67c-11e6-9723-478cfe7829f1.html%0APermanent

Shapiro, D., Dundar, A., Khasiala, P., Yuan, W. X., Nathan, A., & Hwang, Y. (2015). *Completing college: A national view of student attainment rates – Fall 2009 cohort (Signature Report No. 10)*. Herndon, VA.

Shapiro, D., Dundar, A., Yuan, X., Harrell, A. T., & Wakhungu, P. K. (2013). *Completing college: A national view of student attainment rates - Fall 2008 cohort (Signature Report No. 8)*. Herndon, VA.

Shaw, E. J., & Mattern, K. D. (2013). Examining student under- and overperformance in college to identify risk of attrition. *Educational Assessment*, *18*(4), 251–268. http://doi.org/10.1080/10627197.2013.846676

Singell, L. D., & Waddell, G. R. (2010). Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education*, *51*(6), 546–572.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

St. John, E. P., Hu, S., Simmons, A. B., & Musoba, G. D. (2001). Aptitude vs. merit: What matters in persistence. *The Review of Higher Education*, *24*(2), 131–152. http://doi.org/10.1353/rhe.2000.0031

Sullivan, D. F. (2010). The hidden costs of low four-year graduation rates. *Liberal Education*, *96*(3). Retrieved from http://www.aacu.org/publications-research/periodicals/hidden-costs-low-four-year-graduation-rates

Svanum, S., & Bigatti, S. M. (2009). Academic course engagement during one semester forecasts college success: Engaged students are more likely to earn a degree, do it faster, and do it better. *Journal of College Student Development*, *50*(1), 120–132.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education, Inc.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago, IL: The University of Chicago Press.

Venezia, A., & Jaeger, L. (2013). Transitions from high school to college. *The Future of Children*, *23*(1), 117–136. http://doi.org/10.1353/foc.2013.0004

Vuong, M., Brown-Welty, S., & Tracz, S. (2010). The effects of self-efficacy on academic success of first-generation college sophomore students. *Journal of College Student Development*, *51*(1), 50–64.

Wyatt, J. N., Wiley, A., Camara, W. J., & Proestler, N. (2011). *The development of an index of academic rigor for college readiness. Research Report No. 2011-11*. New York, NY. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED561023&site=ehost-live

Yorke, M., & Longden, B. (2004). *Retention and student success in higher education*. Maidenhead, UK: Open University Press.

Zajacova, A., Lynch, S. M., & Espenshade, T. J. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education*, *46*(6), 677–706.

Zhang, Q., & Sanchez, E. I. (2013). *High school grade inflation from 2004 to 2011. ACT Research Report Series, 2013 (3).* Iowa City, IA. Retrieved from http://act.org/research/researchers/reports/pdf/ACT_RR2013-3.pdf

Zwick, R. (2007). *College admission testing*. Arlington, VA. Retrieved from http://stage.nacacnet.org/research/research-data/nacac-research/Documents/StandardizedTesting.pdf

APPENDIX A
LOUISIANA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPLICATION FOR EXEMPTION FROM INSTITUTIONAL OVERSIGHT

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu | lsu.edu/irb

TO: Adam Elder
Education

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: December 9, 2016

RE: **IRB#** E10251

TITLE: Who Stays and Who Leaves? Predicting College Student Persistence Using Comprehensive Retention Models

New Protocol/Modification/Continuation: New Protocol

Review Date: 11/28/2016

Approved X **Disapproved** _____

Approval Date: 12/9/2016 **Approval Expiration Date:** 12/8/2019

Exemption Category/Paragraph: 4a

Signed Consent Waived?: N/A

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable):

Protocol Matches Scope of Work in Grant proposal: (if applicable)

By: Dennis Landin, Chairman _____

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. **SPECIAL NOTE: When emailing more than one recipient, make sure you use bcc. Approvals will automatically be closed by the IRB on the expiration date unless the PI requests a continuation.**

* *All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

APPENDIX B
SUMMARY OF VARIABLES

Variable	Source	Possible Values
	Academic Factors	
High School GPA	IR	0.00 - 4.00
ACT Composite Score	IR, ACT	1 - 36
College Preparatory Curriculum	ACT	Yes / *No*
Registered Credit Hours	IR	12 - 19
	Psychosocial Factors	
Academic Engagement	SSI	1 - 99
Academic Self-Efficacy	SSI	1 - 99
Campus Engagement	SSI	1 - 99
Educational Commitment	SSI	1 - 99
Resiliency	SSI	1 - 99
Social Comfort	SSI	1 - 99
	Noncognitive Factors	
Participated in Extracurricular Activities	ACT	Yes / *No*
Out-of-Class Accomplishments	ACT	Yes / *No*
Plan to Participate in Extracurricular Activities in College	ACT	Yes / *No*
Plan to Work in College	SSI	Yes / Unsure / *No*
Hours Anticipated Studying	SSI	*Less than 10* / 10-19 / 20-29 / 30 or More
Expected Educational Attainment	SSI, ACT	*Bachelor's Degree* / Graduate Degree
Institutional Fit	ACT	0 - 4
	Demographic Factors	
Gender	IR	Male / *Female*
Ethnicity	IR	*White* / African American / Asian / Hispanic / Other
First Generation Status	ACT, IR, SSI	Yes / *No*
	Financial Factors	
Scholarships	IR	0 - 61089
Financial Aid	IR	0 - 57340
Family Income	IR	0 - 6680972

(continued)

Appendix A (continued)

Variable	Source	Possible Values
	High School	
High School Status	NCES	*Public* / Private
Percentage of Minority Students	NCES	0 - 100
Enrollment	NCES	33 - 5457
Residency Status	NCES	*In State* / Out of State
Distance from Home	NCES	0 - 4152
	College	
GPA Differential	IR	0.00 - 4.00
Greek Life	IR	Yes / *No*
Earned Credit Hours	IR	0 - 19

Note. IR = Institutional Record, SSI = Student Strengths Inventory, NCES = National Council of Education Statistics. The reference group for categorical variables is italicized.

VITA

Adam Christopher Elder was born in Birmingham, Alabama to Bruce and Dawn Elder in 1988. He graduated from Clay-Chalkville High School in Pinson, Alabama in 2006, then went on to earn his Bachelor of Science in Secondary Mathematics Education from Auburn University in 2010. Following the completion of his baccalaureate studies, he taught mathematics at Opelika Middle School in Opelika, Alabama before moving to Baton Rouge, Louisiana where he taught mathematics at Dutchtown High School in Geismar, Louisiana. Adam earned his Master of Arts in Education with a concentration in applied research, measurement, and evaluation from Louisiana State University in 2014. He then went on to become a graduate research and teaching assistant as he pursued his doctorate. During this time, he also completed a summer internship in the Statistical and Applied Research division of ACT in Iowa City, Iowa. Adam anticipates graduating from Louisiana State University in May 2017 with his Doctor of Philosophy in Educational Leadership and Research with a concentration in educational research methodology and a minor in applied statistics. He plans to join the faculty at Southeastern Louisiana University as an Assistant Professor in the Department of Educational Leadership and Technology.