2008

A multilevel discrete - time hazard model of retention data in higher education

Christopher W. Guillory
Louisiana State University and Agricultural and Mechanical College, cguill8@lsu.edu

A MULTILEVEL DISCRETE - TIME HAZARD MODEL OF RETENTION DATA IN HIGHER EDUCATION

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 Department of Educational Theory, Policy, and Practice

by
Christopher W. Guillory
B.S. University of Southwestern Louisiana, 1997
M.S. Louisiana State University, 2000
M.Ap.Stat. Louisiana State University, 2003
May 2008

Acknowledgements

I would like to begin by thanking my heavenly Father for the blessings and abilities that made this accomplishment possible. I would like to thank Him for his help and guidance in enabling me to accomplish this life long dream.

I would next like to express me heartfelt thanks to my major professor Dr. Eugene Kennedy for his support, his help, and his guidance in completing this work. Dr. Kennedy helped me to get back on the path to completing a dream I thought I would not be able to fulfill. I would like to thank him for believing in my abilities, when others would not and giving me a chance to prove I had what it takes to be a research statistician. I will always be thankful for the opportunities his hard work, training, and belief in me will make possible in the future.

I would also like to thank all the members of my dissertation committee. Dr. Kim MacGregor, Dr. Janice Hinson, and Dr. Charles J. Monlezun for their valuable input and suggestions. I would also like to thank Dr. Helena Verrill for agreeing to serve as the dean's representative. I would like to thank Dr. Christine Distefano, Dr. Charles Teddlie, and Dr. Barry Moser for the time they spent as members of my dissertation committee.

I would now like to thank my family, who without all of their love and support none of this would have been possible. My wife Melanie whose wealth of support and patience in this adventure, I am eternally grateful. My parents Carroll and Gracie who give me the necessary love and support to help me get to where I am today. I would next like to thank me son, Grant who was my writing partner on several occasions without his help I would not have gotten through some of the tough spots. I would like to thank Cedric Banks for checking in from time to time to see how my progress was going.

Lastly, I would like to dedicate this work to the memories of Dr. Carroll J. Guillory and Dr. E. Barry Mose. Two wonderful men and great teachers who were taken from us much to early.

Table of Contents

List of Tables

List of Figures

Abstract

College student retention rates are often used as a measure of institutional accountability, institutional success, and are used more frequently as a means of determining resource allocation. Understanding what factors impact the retention of college students has become critical to institutions of higher education. The study of the factors that impact student retention has been plagued with methodological concerns, especially the longitudinal and hierarchical nature of retention data. The purpose of this study was to investigate college student retention using a multilevel discrete time hazard model. A multilevel discrete time hazard model deals with many of the concerns associated with analyzing college student retention data, such as censored observations, the multilevel nature of the data, and variables that change over time. Gender, ethnicity and school-type were used to model the timing of students leaving a university from a cohort of first-time freshmen over five year period.

1. Introduction

Importance of Studying Retention

A student retention rate, in higher education, is the percentage of students who complete a semester at a university and return to the university the next semester. A university's student retention rate not only has an impact on the university, but it also impacts the surrounding communities where institutions of higher education are located. An institution's retention rate influences student recruitment, funding, and public perception. Universities study student retention to understand why students leave and how to improve the retention of students. Studying student retention allows institutions to improve diversity in higher education, learn how the choice of major, and the amount of financial assistance change student retention rates, and then use this information to improve students' rate of persistence (St. John, Shouping, Simmons, Carter, & Weber, 2004). Examining retention also allows institutions to see the different trends that effect student retention, such as why some of the best students do not return to the university, how financial hardships effect students, or why students may not be satisfied with their college experience and transfer to another school (Tinto, 1990). It is important for a university to study retention to determine why students are leaving, but it is also important when dealing with state legislators, and understanding the impact student retention has on the amount of funding colleges and universities receive from state governments.

The number of state legislatures, who want to connect university funding with the number of students that graduate is on the rise. This decision by policymakers is of great concern to public colleges and universities that have large enrollments of at-risk students (Barefoot, 2004). Also universities are now being held accountable for the attrition rate of their students by their governing boards and a higher attrition rate can cause universities to have a poor public image (DesJardins, Ahlburg, & McCall, 1999).

The decrease in federal and state funding of higher education has caused universities to use those funds more efficiently (DesJardins et al., 1999). The amount of funding a public university receives is based on its enrollment. Thus when students do not return, the amount of funding a university receives decreases. The opposite is also true when students return to the university, the university's enrollment grows, and the university receives more funding (Bowen, 1980). A university may have a disproportional lose in funding, when the

enrollment of an institution decreases, due to the fact that students are not persisting (DesJardins et al., 1999). The decrease in funding has also caused universities to study student retention because the increase in cost of getting a higher education has been passed on to students.

Hu and John (2001) discovered that in recent years the cost of higher education has increased and the amount of funding received by public colleges and universities from states has been on the decline. Thus the increased cost of higher education has been passed to college students and their families. This increase in cost has caused universities to be more concerned with getting students to persist at the institutions.

Introduction of Analysis and Data Source Used in This Study

Longitudinal Data Analysis

Longitudinal data are collected from the same population over a length of time. Longitudinal data allow a researcher to follow patterns of change in the same population over a period of time (Creswell, 2002). Willett and Singer (1991) gave the following benefits of using longitudinal data: collecting longitudinal data allow a researcher to have a better understanding of a student's college career, a way to follow factors that impact a student's decision to stay or leave a university, and to "increase statistical power".

Hedeker and Gibbons (2006) illustrated several advantages to using longitudinal data analysis. The first advantage is the economical use of information gathered on subjects. The next advantage is the ability to use the information gathered on subjects as the control for the subjects. Another advantage is the between-subject variation is omitted from the error. When the patterns and observations are the same, longitudinal designs provide better estimators than with cross-sectional designs. Longitudinal data analysis allows a researcher to determine the change in an individual over the length of a study. One other advantage of longitudinal data analysis is that it provides information on the change that occurs in the subjects.

Hierarchical Linear Modeling

A hierarchical linear model is a model that consists of nested data, for example the productivity of workers may be influenced by workplace characteristics. In this example data are gathered on the workplace and the workers with analysis being done on both levels. There is a hierarchy to the data in this example the workers are nested within the

workplace, since workers are nested within the workplace the variation the workplace causes on the workers must be accounted for in the study (Raudenbush & Bryk, 2002).

Raudenbush and Bryk (2002) showed several advantages to using multilevel data analysis. First, it allows a researcher to determine the amount of variability caused by each level of data hierarchy. Second, a researcher is able to model the first level of data analysis in terms of the effects at all levels. Third, by using a multilevel model a researcher is able to test the possible interactions between each level of data. Finally, the subjects within the data set are similar because they come from similar environments because of this subjects are not independent, and multilevel data analysis is able to handle the absence of independence in the subjects.

Hazard Modeling

Willett and Singer (1991) define hazard modeling as "the population hazard function describes the risk of an event's occurrence in each time period, the probability that a randomly selected population member will experience the event in the period given that the event has not already occurred" (p.954). Discrete-time survival analysis has several advantages. First, discrete-time survival analysis is suited to analyze longitudinal data. Second, discrete-time survival analysis can handle time-invariant and time-variant predictors. Third, violations of the model can easily be tested and corrected. Finally, censored observations can be handled with discrete-time survival analysis.

Data Source and Target Population

The data used in the study came from the National Longitudinal Survey of Youth, 1997 (NLSY97). The National Longitudinal Surveys (NLS) was a group of surveys supported by the Bureau of Labor Statistics (BLS), U.S. Department of Labor. These surveys were used to gather data about the labor market and the different individuals who made up the labor market at several different points in time. The National Longitudinal Survey of Youth, 1997 gathered information relating to the 1997 population born between the years 1980 to 1984. The NLSY97 survey consists of $8,984$ respondent with information on education, employment opportunities, family background, and environmental information (BLS, 2003).

The target population for this study consisted of individuals born between the years 1980 and 1984 and who attended college between the years 1999 and 2004.

Brief Introduction to Retention Literature

Tinto's Student Integration Model

Vincent Tinto's Student Integration Model was based on the works of both Van Gennep's Rite of Passage (Van Gennep, 1960) and Spady's work in student retention (Spady, 1970, 1971). The Student Integration Model was designed to explain the stages students go through when starting college. Tinto's Student Integration Model was divided into three stages: separation, transition, and incorporation. The separation stage was characterized by students decreasing the amount of time spent with individuals they associated with before college. The transition stage consisted of students starting to interact with their new environment. Students learned the skills necessary to function in their new setting. In the incorporation stage, students became full participants of their new environment. The length of time a student continued at the university was based on how well they maneuvered through the three stages. If a student was able to separate themselves successfully from their pre-college environment, successfully learn the new skills needed in the college environment, and successfully incorporate those skills into their position in the college environment these students were more likely to persist (Tinto, 1982).

Bean's Student Attrition Model

Bean (1982) explored the relationship between the dependent variable dropout and the independent variables: practical value, intent to leave, loyalty, certainty of choice, courses, grades, educational goals, opportunity to transfer, major and job certainty, and family approval of the institution. When the gender of the student was unknown, Bean discovered the mean rank of the ten independent variables to be in the following order from first to tenth: intent to leave, grades, opportunity to transfer, practical value, certainty of choice, loyalty, family approval, courses, student goals, major and occupational certainty. Thus Bean determined that intent to leave, grades, and opportunity to transfer to be the most significant variable in determining a student's decision of returning to the university.

Bean (1983) used Price and Mueller's model of turnover in the work place (Price & Mueller, 1981) for the basis of his Student Attrition Model. In Bean's Student Attrition Model, he explained a "student's interaction with a university." The Student Attrition Model was used to determine "student satisfaction." The following variables were used to measure student attrition: intent to leave, grades, practical value, opportunity, marriage,

satisfaction, campus organizations, courses, and participation. Four of the variables were significantly related to dropout, they were: intent to leave, grades, courses, and marriage.

Cabrera's Work with Tinto's and Bean's Models

Cabrera worked to determine the similarities between Tinto's Student Integration Model (Tinto, 1982) and Bean's Student Attrition Model (Bean, 1983). Cabrera tested the independence and correlation of the constructs in the two models. They also tested if the constructs represented the same concepts. They found *Courses*, a construct in the Student Attrition Model, and *Academic Integration*, a construct in the Student Integration Model, were similar constructs "and provided a perfect fit for the data." They also determined that *Institutional Commitment*, a construct in the Student Attrition Model, and *Institutional Fit and Quality*, a construct of the Student Integration model were similar constructs. Their findings suggested that by using both models to study student retention a researcher developed a better understanding of why students were not persisting (Cabrera, Castaneda, Nora, & Hengstler, 1992).

Willet and Singer's Work with Survival Analysis

Willett and Singer (1991) placed the focus of their study on the importance of when an event occurred instead of whether an even occurred. They asked when a student was at the greatest risk of not returning to the university instead of whether a student returned after the first semester. By asking when a dropout occurred instead of asking whether a dropout occurred a researcher discovered more than just whether the event occurred, but can learn if the event occurred at more than one point in time.

Willett and Singer (1991) estimated student retention using a survival rate, which is the proportion of students in one semester who persisted in school to the next semester. By using survival analysis, Willet and Singer were able to incorporate censored observations into their analysis to estimate the median time until students left the university. They also suggested when studying retention the group of students still available at the end of each semester should be analyzed. This analysis could be done by using the sample hazard probabilities. The sample hazard probability was the conditional probability that a student would return to the university the next semester given that he or she survived the previous semester.

Statement of the Problem

A university's student retention rate influences how university officials can improve their university, create programs to improve the retention rate of different student populations, and to obtain funding from the state. Past research has shown that retention rates are influenced by such factors as students' college grade point average, socioeconomic status, or the amount of financial aid a student receives (DesJardins et al., 1999; DesJardins, Ahlburg, & McCall, 2002b; DesJardins, Kim, & Rzonca, 2002 - 2003; DesJardins, Ahlburg, & McCall, 2006; Hu & John, 2001). These are only some of the factors that influence a student's decision to return to a university at the beginning of each semester. However, these factors are all individual student-level variables.

There was a need to study both student-level variables and other levels of the hierarchy of student retention in higher education. There were few multilevel discrete hazard studies of retention in higher education and there were even fewer studies that look at the hierarchical data of student retention, and how those different levels effect student retention. This study was one of the first to use multilevel discrete-time hazard models to explore the impact of the hierarchical data structure found in student retention data.

Objectives of the Study

The primary objective of this study was to use a multilevel discrete-time hazard model to determine what impact the different nested levels of higher education had on the retention of students. This study also explored the development of a multilevel discrete-time hazard model. In addition this study illustrated the use of statistical software to estimate models of how the school-level variable, the type of school, whether a student was enrolled in a public or private four-year university, a student was attending effects retention and how the student-level variables influenced retention.

The specific objectives of this study were the following:

1. Describe and analyze retention in higher education with a multilevel discrete-time hazard model.

2. Explore the likelihood a student left a university during a year.

3. Explore what individual level factors were most influential in a student's decision to leave a university during a year.

4. Explore the extent the type of school a student attended effected the risk of a student leaving a university during a year.

Significance of the Study

This study was significant because it was one of the first to use multilevel discrete-time hazard models to analyze student retention data in higher education. The significance that this was the first use of a multilevel discrete-time hazard model was very important because this study established a new method that can be used to study retention data. This study helped to establish a very important analytical tool in the study of retention data.

There were several student-level variables that have been found to effect a student's decision to leave a university during the year and/or fail to return to a university at the beginning of the next year. There have been many studies done on retention in higher education, but none of these studies have looked at the multilevel structure of the data found in higher education.

This study also used a data analysis method not commonly found in retention studies. This method was appropriate because of the structure of the data and when the data were collected. This method allowed for the evaluation of the predictor variables and took into account the longitudinal nature of the data.

Past research has shown that retention in higher education was shaped by different student-level variables. This study explored the relationship between the different hierarchical levels found in retention data in higher education.

This study also explored the use of multilevel discrete-time hazard models, which were a combination of survival analysis and hierarchical linear modeling. This was a relatively new method to analysis data and to the researcher's knowledge has not been used to study retention in higher education. The use of this method allowed for exploring the relationship between student-level variables, but it also allowed for the exploring of the relationship between the nested levels found in retention data.

Definition of Terms

Student: An individual enrolled at a four-year university.

Hierarchical Linear Modeling: the use of data from nested levels to determine the impact of individual level and group level factors (Raudenbush & Bryk, 2002).

Survival Analysis: Survival analysis is a method of statistical modeling that deals with the occurrence of events in a longitudinal data set or the timing of events (Allison, 1995).

Discrete-Time Hazard: the conditional probability that individual experienced the event of interest in time interval t_j given that the individual has not experienced the event in any earlier time intervals (Singer & Willett, 1993). That is

$$h(t) = \Pr\{T = t_j | T > t_j\}.$$

Survival Function: The **cumulative distribution function** (cdf) of T is defined by $F(t) = Pr(T \leq t)$, gives the probability that a student will drop out before time t, $F(t)$ can also be thought of as the proportion of students in the population that will drop out before time t. The **survival function**(sf) is the complement of the cdf (Meeker & Escobar, 1998),

$$S(t) = Pr(T > t) = 1 - F(t) = \int_t^\infty f(x)dx$$

and gives the probability of surviving beyond time t.

Censoring: occurs when an observation's exact failure time was not known (Allison, 1995).

Longitudinal Study: A study in which the same population is observed over a period of time. Longitudinal study allowed a researcher to follow patterns of change in the same population over a period of time (Creswell, 2002).

Time-Variant Variables: variables that can have different values in each time period (Singer & Willett, 1993).

Time-Invariant Variables: variables that remain the same in each time period (Singer & Willett, 1993).

Person-Period Data Set: a data set where each individual in the study had multiple lines of data. One line of data for each period the individual was observed (Singer & Willett, 1993).

Person-Oriented Data Set: a data set where each individual in the study had one line of data (Singer & Willett, 1993).

Maximum Likelihood Estimators: Let $X_1, \ldots, X_n$ be a sample from a population with pdf $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function was defined by

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k).$$

Let $\hat{\theta}(x)$ be a parameter value, for any sample point $\mathbf{x}$, at which $L(\theta|x)$ was its maximum as a function of θ, where $\mathbf{x}$ was held constant. "A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample of $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$" (Casella, 2002, pg. 316).

Assumptions and Limitations of the Study

The first limitation of this study involved the time intervals used in each period of observation. In this study, one period of observation spanned a year, but there were two semesters in one year. A student could have left a university at the end of the fall semester, however they were considered enrolled for the complete year.

Another limitation had to do with a student leaving a university. There was no difference between a student dropping out of a university and a student graduating from a university. There was no way to determine if a student graduated or if a student did not return to the university. Thus if a student graduated from a university he/she was considered as not retained.

The next limitation dealt with the variables used in the study. This study used an extant database, and because of this the variables that were available in that database had to be used to determine the effect on retention. The available variables did not allow for the exact match of the desired variables wanted for this study.

The final limitation has to do with the problem of missing data in the data set. The amount of missing data caused several limitations in this study. The first limitation was the amount of missing data caused a degree of bias in the sample used, and caused problems in the analysis of the data. The next limitation dealt with the ethnicity variable. The ethnicity variable separated the students in this study to white and non-white students. The ethnicity variable was created because of the smaller number of non-white students compared

to the number of white students. This grouping of the data caused a bias in the parameter estimation of ethnicity because African American, Native American students historically had lower retention rates than white students and Asian American students historically had higher retention rates than white students. The final limitation caused be the missing data dealt with the school-level variable school type. School type was a dichotomous variable that indicated if a student attended a public four-year university or a private four-year university. The school type variable was separated into these two groups because of the large number of students enrolled at public four-year universities compared to the other type of schools in the study. This grouping of the data also caused a bias in the parameter estimate of school type.

The above limitations caused the model used in this study to not be as a good as original intended. That being said, the demonstration of the multilevel discrete-time hazard model was still important to included in the study of retention data.

2. Review of the Literature

This chapter presents a review of the literature related to retention in higher education. The beginning of this chapter is a review of the major theories of retention in higher education. Next is a discussion of the importance of using survival analysis to analyze retention data. This is followed by a discussion of the methods used to study retention in higher education. The final section summaries the results of retention studies in higher education.

Search of Relevant Literature

This literature review was conducted using the following databases: ERIC, Infotrac, and Academic Search Complete, and spanned the period 1960 to 2008.

Spady's, Tinto's, Bean's, and Cabrera's Work in Retention

Spady's Work in Retention

In Spady's (1970) review and synthesis of retention, he found the studies of retention lacked "both theoretical and empirical coherence." He called for an end to the "bivariate research on the 'correlates' of dropping out." Spady suggested the use of Durkheim's theory of suicide (Durkheim, 1961) to provide an interdisciplinary approach to exploring the relationship between the abilities of students and the social and academic aspects of a university. Spady suggested that students succeed in the social and academic systems of a university by successfully integrating into these two systems. A student's successful integration was based on receiving rewards in each system. In the academic system, the rewards were grades and intellectual development. For the social system, success was measured by "having attitudes, interests, and personality dispositions that were basically compatible with the attributes and influences of the environment. This condition we call *normative congruence*" (p. 83). Success in the social systems was also based on developing a "friendship support" network that allowed students to become part of the social system of the university. Spady later expanded on his work here by developing a model of student retention based on Durkheim's theory of suicide.

Spady (1971) attempted to create an empirical model that could deal with the inconsistencies he put forth in Spady (1970). Spady's empirical model consisted of the following ten variables: institutional commitment, satisfaction, social integration, intellectual development, grade performance, friendship support, normative congruence, academic po-

tential, family background, and previous educational background. Spady's used principal component analysis and multiple regression analysis to analyze the data. Of the ten variables, Spady found that grade performance, institutional commitment, and social integration explained the greatest amount variance in student retention rates when looking at men, and from women a general commitment to the university explained the most variance in student retention rates. When Spady tried to explain the inconsistences that were present in student retention research at that time he discovered the variables in his model, "did not provide an unambiguous basis for revising the model in some particular way, nor do they automatically resolve many of the ambiguities and inconsistencies in the literature" (p. 57). Vincent Tinto used Spady's work in student retention to begin building what was to become one of the cornerstones in studying student retention in higher education.

In summary, Spady wanted to develop a theoretical model for studying student retention that did more than just describe the "bivariate research on the 'correlates' of dropping out." To develop this theoretical model, Spady used Durkheim's work in suicide. Spady discovered that by using grade performance, institutional comment, and social integration one was better able to understand student retention in higher education.

Tinto's Work in Retention

Tinto (1975) further developed Spady (1971) model for studying student retention in higher education. Tinto wanted to extend Spady's model to explain the "processes of interaction between the individual and the institution that lead differing individuals to drop out from institutions of higher education, and that also distinguishes between those processes that result in definable different forms of dropout behavior" (p. 90). To explain the different types of dropout behavior Tinto added the following variables to Spady's model: educational goal commitment, which represented both the desire for one's education and the expected level of degree completion; external forces, which represent such things as possible employment opportunities, and possible future opportunities of advancement denied to individuals based on sex or ethnic group membership; perceptions of reality, individuals may view the same situations differently because of their background and experiences. Tinto determined that a student's commitment to both the institution and to completing college greatly impacted if an individual dropped out of the university and the type of "dropout behavior the individual adopts." Tinto would later add the importance of the timing of

when a student left the university to get a better understanding of why students completed a college education or did not complete a college education.

Tinto (1988) added a new dimension to the study of student departure by recognizing that students leave a university at different times for different reasons. "The forces that shape departure during the first year of college, especially during the first six weeks of the first semester, are qualitatively different from those that mold departure in the latter years of college" p(439). Tinto extended his student retention model by using Van Gennep's social anthropology study of the process of becoming part of a tribal society (Van Gennep, 1960). This extended model has become known as the Student Integration Model. The Student Integration Model was designed to explain the stages students go through when entering college. Tinto's Student Integration Model was divided into three stages: separation, transition, and incorporation. The separation stage was characterized by students decreasing the amount of time spent with individuals they associated with before college. The transition stage was when students started to interact with their new environment. In this stage, the students learn the skills necessary to function in their new setting. In the incorporation stage, students became full participating members of their new environment. The amount of time a student continued at the university was based on how well they maneuvered through the three stages. If a student was able to separate themselves from their pre-college environment, learned the new skills needed in the college environment, and incorporated those skills into their position in the college environment these students were more likely to persist (Tinto, 1982). Looking at the different stages of student departure Tinto was able to take into account the longitudinal process students go through to either become part of the college community or leave the college community.

In summary, Tinto extended Spady's model of student retention by explaining how the interaction between individuals and a university can cause different individuals to leave a university, and how students could leave the university through different types of dropout behaviors. In order to explain these two concepts, Tinto added education goal commitment, expected level of degree completion, and possible future opportunities denied to individuals because of gender or ethnic background. Tinto also added a time dimension to his model were students go through different stages to become integrated into a university. Tinto used Van Gennep's model of tribal society (Van Gennep, 1960) to develop the following

stages a student goes through when starting at a university: separation, transition and incorporation.

The works of Spady and Tinto did have shortcomings and these shortcomings were pointed out by other researchers. One of those individuals was John Bean, whose work in retention lead to the development of the Student Attrition Model.

Bean's Work in Retention

Bean (1980) suggested that Spady (1970, 1971) and Tinto (1975) models of student retention were insufficient. First, he suggested there was no evidence to support a connection between student retention and Durkheim's theory of suicide (Durkheim, 1961). Secondly, he felt the operational definition of the variables in the two models did not allow for the use of path analysis to analyze the data. Bean applied Price's model of turnover in the work place (Price, 1977) to explore what factors caused students to leave college. Bean used a causal model with three types of independent variables: satisfaction and institutional commitment, organizational determinants, and background variables. Bean used multiple regression and path model analysis to determine that institutional commitment was the most influential variable to explain dropping out of institutions of higher education.

Bean (1982) reduced Bean (1980) model of over twenty independent variables to ten independent variables, and the sample was divided into high confidence and low confidence men and women. The ten independent variables were: intent to leave, practical value, certainty of choice, loyalty, grades; courses, educational goals, major and job certainty, opportunity to transfer, and family approval of the institution. Four of the ten variables were found to significantly effected dropout, they were in order of significance: intent to leave, grades, opportunity to transfer, and loyalty.

Bean (1983) used Price and Mueller's model of turnover in the work place (Price & Mueller, 1981) for the basis of his Student Attrition Model. In Bean's Student Attrition Model, he explains "student's interaction with a university." The Student Attrition Model was used to determine "student satisfaction." The following variables were used to measure student attrition: intent to leave, grades, practical value, opportunity, marriage, satisfaction, campus organizations, courses, and participation. Four of the variables were significantly related to dropout, they were: intent to leave, grades, courses, and marriage.

In summary, Bean felt that Spady's and Tinto's works in retention were insufficient, because of this he used Price and later Price and Mueller's models of work place turnover to explore student retention in higher education. He discovered that institutional commitment, intent to leave, grades, opportunity to transfer, loyalty, and marriage were important predictors in understanding retention in higher education.

Tinto's Student Integration Model and Bean's Student Attrition Model are the two cornerstones of the modern day theory of student retention. Because of the importance of these two theories it is important to understand how they are similar, dissimilar, and how they can be used together to get a better understanding of student retention.

Cabrera's Work in Retention

Cabrera worked to determine the similarities between Tinto's Student Integration Model (Tinto, 1988) and Bean's Student Attrition Model (Bean, 1983). Cabrera et al. tested the independence and correlation of the constructs in the two models. They also tested if the constructs represented the same concepts. They found *Courses*, a construct in the Student Attrition Model, and *Academic Integration*, a construct in the Student Integration Model, to be similar "and provided a perfect fit for the data." They also determined that *Institutional Commitment*, a construct in the Student Attrition Model, and *Institutional Fit and Quality*, a construct of the Student Integration model were similar. Their findings suggested by using both models to study student retention a researcher obtains a better understanding of why students are not persisting (Cabrera et al., 1992).

Cabrera, Nora, and Castaneda (1993) extended Cabrera et al. (1992) work by exploring the extent to which Tinto's Student Integration Model (Tinto, 1988) and Bean's Student Attrition Model (Bean, 1983) could be used together to explain student retention. The merger between the two theories was done "by simultaneously testing all non-overlapping propositions" of the two models. The study found that environmental factors played a significant part in understanding student retention. This showed that environmental variables were more significant than what was found in the Student Integration Model. The environmental variables effected academic experience and socialization of the students.

In summary, Cabrera et al. determined that both models had some constructs in common and some constructs that were different and thus gave a researcher different types of insight into retention in higher education. Cabrera's research offered an individual a better

understanding of retention by using both the Student Integration Model and the Student Attrition Model together.

Survival Analysis Used in Student Retention

Why Use Survival Analysis

Willett and Singer (1991) determined that "dropout rates calculated using aggregate enrollment figures are among the most misleading educational statistics published today" (p. 429). Analyzing retention data with traditional methods of data analysis had several disadvantages. Traditional methods of data analysis provided results that could be misleading. It was determined that the enrollment figures were used for political gain, and did not provide insight to why students were leaving the university. These misleading figures did not provide a true understanding of what type of students were leaving universities or why these students were leaving universities before completion of a degree. These statistics also do not take into account censoring, do not detail the risk over time, or how risk changes over time. Survival analysis provides several advantages, and gave further credence to the difficulties traditional methods of analysis encountered when analyzing retention data.

One of the problems in analyzing retention data was what to do with the censored observations. Censoring occurred when a subject did not experience the event of interest by the time the study was concluded. The question then became what to do with the censored observations. One suggested method of dealing with censored observations, when using traditional data analysis, was to separate the sample into subjects that have experienced the event and those that have not experienced the event. This method allowed logistic regression to be used on the data set. This dichotomized sample could hide knowledge about "educational transitions." The splitting of the sample into those who have experienced the event and those who have not experienced the event could eliminate possible important variation because of the clustering nature of the split data set. Another method of dealing with censored observations, when using traditional methods of data analysis, called for researchers to give the censored observations an event time. This event time was usually the time the observation ended. This method caused an underestimation of results. One last method of dealing with censored observation, when using tradition methods of data analysis, had researchers try to design the model to handle the censored data. This method called for the researcher to design the experiment only to look at individuals who have experienced the

event of interest. This method changed the target population and thus changed the research questions (Willett & Singer, 1991). The question of what to do with censored data can be handled with survival analysis.

Willett and Singer (1991) described several benefits of using survival analysis to analyze retention data. Survival analysis provided researchers with a large class of methods that allowed for the description of "temporal patterns of occurrence, compared these patterns among groups, and build statistical models of the risks of occurrence over time" (p. 409). Survival analysis did not have the same shortcomings as found with traditional data analysis, when analyzing retention data. The data gathered by researchers were linked to a certain point in time. In the case of students in higher education, these points in time were usually per semester, one year, four years, or six years. Traditional forms of data analysis did not take into account the changes in risk over time. By not accounting for the change researchers could not discover what predictors effected the risk of the event of interest occurring. Survival analysis' primary focus was analyzing the changes in risk over time. The results generated from traditional forms of data analysis could not change with the point in time that was being examined. "In survival analysis, the time itself is an integral part of the answer; it highlights, rather than obscures, variation over time" (p.426). Traditional methods of analysis had no systematic way of bringing censored observations into the model. Finally when using traditional methods of analysis it was difficult to incorporate time varying predictors into the analysis. When using survival analysis, the analysis was the same whether the predictors were time-invariant, time varying predictors, or both. When studying longitudinal data, it was possible for some observations to be followed for different lengths of time, and thus may cause observations to have different censoring times, and this may cause observations to have different risk periods. Traditional methods of analysis were unable to take different time periods into account. An area of survival analysis that was a useful analytic tool in the area of student retention was discrete-time survival analysis.

Willett and Singer (1991) provided five reasons to use discrete-time survival analysis when analyzing retention data. First, retention observations were recorded in discrete time. Second, it provided a practical way to introduce an individual to using survival analysis. Third, both time-invariant and time-varying predictors could be used in a discrete-time model. Fourth, violations of the assumptions of the model could easily be tested and cor-

rected when using a discrete-time model. Finally, analysis done with discrete-time models could be done with standard statistical software. The hazard functions was an important tool in analysis data with discrete-time survival analysis.

Willett and Singer (1991) suggested several advantages to using hazard probabilities. By computing hazard probabilities, researchers were able to statistically determine when students were most likely not to return to the university. Using hazard probabilities also allowed researchers to develop statistical models that could find important predictors of student profiles. The hazard model allowed analysis of data that was "powerful, flexible, and sensitive approach" when looking at student retention "that allows simultaneous inclusion of both censored and uncensored" students. The hazard model allowed for the inclusion of both time-invariant predictors and time-variant predictors.

(Willett & Singer, 1993) put forth several advantages to using the hazard function when analyzing survival data. The hazard function described the risk of an event's occurrence in each time period—the probability that a randomly selected student would not return to the university in that semester, given that the student had not already dropout of the university. The hazard function provided many benefits when using survival analysis. A researcher could tell if an event occurred by using the hazard function. Censored observations did not effect the hazard function. Hazard functions could be determined in all periods of time where the events occurred which implied "no information is ignored or pooled." The survival function could be determined by using the hazard function in cases where censoring prevented the direct calculation of the survival function.

In summary, survival analysis over came many of the difficulties that arise when retention data were analyzed using traditional methods of analysis. Survival analysis could handle censored observations, changes in risk over time, time-varying and time-invariant variables at the same time as well as observations with different start and stop times. Researchers could determine when an event of interest was most likely to occur in each time period examined.

Methods of Studying Retention

This section of the literature review deals with the different methods used to study retention data. In the beginning, retention data were analyzed using bivariate research and correlation analysis (Spady, 1970). As the years progressed and the advancements in

computer technology, there have been many new statistical techniques employed to analyze retention data. What follows is a brief description of the methods used in studies dealing with retention in higher education.

Bayer (1968) used multiple regression and correlation analysis to study $8,567$ students who started college within five years of completing high school. Bayer used 38 background and personal factors to determine their impact upon a student's ability to complete college.

Pascarella and Terenzini (1977) used discriminant analysis to examine the relationship between faculty and student interactions and freshman attrition. They looked at students who persisted and those who dropout to determine how the different types of interactions with faculty members effected student retention.

Munro (1981) analyzed the National Longitudinal Study of the High School Class of 1972 with path model analysis to test Tinto's model of college retention. Munro looked at $6,018$ full time students entering four-year universities in the fall of 1972.

Willett and Singer (1991) was one of the first to suggest the use of survival analysis to analyze retention data. They put forth the notion of answering the following question "whether events occurred by trying to determine when the events occurred." They also suggested that researchers should examine when was an individual at the greatest risk of experiencing the event of interest. Willett and Singer (1991) analyzed retention data of both teachers and students using survival analysis.

Willett and Singer (1993) illustrated the advantages of analyzing retention data with the hazard function and discrete-time survival analysis. They illustrated how survival analysis dealt with some of the problems of time-to-event data, such as censoring, the different entry points, and exit points of some individuals into a study. One other important illustration was the creation of a person-period data set. A person-period data set contained a record of data for each individual with information on each predictor at each time period the data was recorded.

Willett and Singer (1995) expanded on Willett and Singer (1993, 1991) by showing how survival analysis can be used to model a multiple-spell data set. A multiple-spell data set consists of some observations where one individual experienced the event of interest on several occasions or an individual experienced different forms of the event of interest. The

multiple-spell model had several advantages over the sequential single-spell model. Those advantages were: the simultaneous analysis of multiple-spells was more efficient, a researcher was able to test for differences in the effect of independent variables across spells, and allowed a researcher to test for the delayed effect of independent variables "from one spell to the next."

Allen (1999) study of retention used structural equation modeling to examine the impact of precollege background variables, motivation, and persistence behavior in the retention of minority and non-minority students. Allen was interested in the direct and indirect effects of motivation on academic performance in college and on persistence behavior. The study also wanted to determine to what extent motivation differed in the retention and academic performance for minority and non-minority students.

Murtaugh, Burns, and Schuster (1999) was one of the first to use survival analysis to study retention data in higher education. They showed advantages of using survival analysis to analyze retention data. The study also determined some of the factors that effected student retention at Oregon State University. The study analyzed a longitudinal data set that consisted of $8,867$ first-time freshmen who were enrolled starting in the fall quarter of 1991 through 1995. The predictors in the study were: age, sex, ethnicity/race, residency, college at first enrollment, high school GPA, SAT score, first quarter GPA, participation in Educational Opportunities Program, and enrollment in Freshman Orientation Course. They performed an univariate analysis, the Kaplan-Meier Method, and a multiple-variable analysis, the Cox proportional hazards regression model.

DesJardins et al. (1999) wanted to get a better understanding of student departure from a university by using event history modeling, another name for survival analysis. The study was begin done to get "a more exact timing of departure into the estimation of student exits from college and permits a more appropriate utilization of longitudinal data" (p 376). They gathered data on $3,975$ first-time freshmen from the University of Minnesota starting in the fall 1986. The data set covered twenty-two terms. The independent variables used in the model were: race, gender, age, initial home location, the presence of a disability, composite ACT score, high school rank percentile, college major, college GPA for each term, and financial aid. They constructed two models a time-constant coefficient model and a time-varying coefficient model.

Elkins, Braxton, and James (2000) used path analysis to analyze the impact of student pre-entry characteristics, initial institutional commitment, separation and first- to second-semester persistence on student persistence during their first semester.

Hu and John (2001) explored the effects of federal and state aid programs on the retention of minority students and to understand the differences in retention rates between different racial/ethnic groups. The following predictor variables were used in this model: age, ethnicity, dependency status, income, college grades, types of institution attended, housing status, year in college, type of financial aid.

DesJardins, McCall, Ahlburg, and Moye (2002) further developed the work of Adelman (1999) and DesJardins et al. (1999) to determine what factors effect a students ability to complete a college degree. They used a flexible time varying coefficient model which controlled for unobserved heterogeneity to study the impact of time on the results of Adelman (1999) *Tool Box* study. The *Tool Box* study looked at factors that effected the time it took a student to complete a college degree. Adelman's *Tool Box* looked at the effect of academic resources and academic patterns on the time to complete college. Desjardins et al. used event history modeling to see if they found similar results to those found by Adelman. Desjardins et al. used three variables to measure academic resources: academic intensity/quality, high school rank, and senior year test score. The other variables included in the model were: a variable to tell if a student was a parent or not, the socioeconomic status of the student, a variable to measure student anticipations, whether or not they believe they would complete college, gender, race, financial aid package information; work-related information, and college GPA.

DesJardins, Ahlburg, and McCall (2002a), extend the work of DesJardins et al. (1999), by using a hazard model to study the longitudinal effects of financial aid on the retention of students in higher education. They were able to separate the financial aid received by students into its component parts, loans, grants, scholarships, work-study, and other forms of on-campus student employment, and allowed these components to vary over time.

DesJardins, Ahlburg, and McCall (2002b) used a time-varying event history model with flexible controls to determine the length of time it took students to complete an undergraduate college degree. The analysis also used a competing risks model which

takes into account the interdependence between different outcomes such as stop out and graduation. They investigated the impact of student's choice of major, student demographic characteristics, academic performance, student ability, attitudinal variables, and financial aid. They used a sample of 2,373 freshmen from the University of Minnesota - Twin Cities campus that covered a nineteen-term period.

Ishitani and DesJardins (2002 - 2003) explored if student dropout rates changed over time, the variables that effect dropout rates, and if the time dropouts occurred remain the same over a period of time. They used an event history model to study a sample of 3,450 U.S. citizens between the ages of 18 to 25 for the time period of August 1989 to June 1994. They used the following variables: gender, race, family income, mother's education, father's education, educational aspiration, GPA, SAT total, institutional types and sizes, career identity, academic integration, social integration, financial aid, and employment.

St. John et al. (2004) used logistic regression to study the difference in retention rates between white and African American students in high demand areas of study. They also looked at the different retention rates in the freshmen and sophomore classes. They used student background variables, college experience variables, student major, and financial aid to determine the retention rate in this study.

Glynn, Sauer, and Miller (2005-2006) used logistic regression to predict the probability that a student would drop out of college. The independent variables used in the model were demographic variables, high school experience, and measures of attitudes, opinions, and values.

Gansemer-Topf and Schuh (2006) used multiple regression to determine the effect of institutional grants, institutional support, facilities, student services, academic support, and expenditures for instruction could predict retention and graduation rates.

DesJardins et al. (2006) examined the different stages of student enrollment: enrolled, stop out, and graduate, and the impact the length of time of a stop out had on the completion of a degree. A multiple spells model was used because of the repeating nature of the data. The sample consisted of 12,648 students who were enrolled at the University of Minnesota - Twin Cities campus as first-time freshman. Three cohorts of students were examined those entering in the fall of 1984, the fall of 1986, and the fall of 1991 and each were observed for more than six years.

Table 1: Summary of the Methods Used to Analyze Retention Data

Bivariate Analysis
Correlation Analysis
Multiple Regression
Path Model Analysis
Discriminate Analysis
Survival Analysis
Structural Equation Modeling
Logistic Regression
Trend Analysis
Ordinary Least Squares Regression
Multivariate Analysis

Ishitani and Snider (2006) used survival analysis to determine the longitudinal effect of "high school programs on college retention." They used a sample of $4,445$ first-time freshmen who were attending four-year universities from 1992 to 1994. The data were gathered from the National Education Longitudinal Study: $1988 - 2000$ and the NELS:88/2000 Postsecondary Education Transcript Study. Ishitani and Snider employed both an exponential survival model and a period-specific model to analyze the data.

Noble, Flynn, Lee, and Hilton (2007 - 2008) used multivariate analysis to study the impact of the ESSENCE program on the GPAs and retention rates of first year students. The dependent variables used in this study were first year GPA, four year graduation rate, and five year graduation rate. The only independent variable was participation in the ESSENCE program.

Table 1 summarizes the different methods used to analyze retention data, but these studies were missing the advantages available in using a multilevel discrete-time hazard model. These studies did not take into account the nested nature of retention data. These studies also did not take into account the possibility of observations having similar environments and similar behaviors. This study used a multilevel discrete-time hazard model which took into account what was missing from previous studies.

Results of Retention Studies

The following section summarizes the results found in previous studies of retention data. This section shows the benefits of programs used to increase retention rate (Noble et al., 2007 - 2008). This section also shows the impact of ethnicity and the receipt of the dif-

ferent types of financial aid had on retention rates (Hu & John, 2001; DesJardins, Ahlburg, & McCall, 2002a). This section also demonstrates the importance of understanding what factors cause students to leave a university in order to better assist them (Gansemer-Topf & Schuh, 2006; Ishitani & DesJardins, 2002 - 2003; Glynn et al., 2005-2006; St. John et al., 2004).

Bayer's multiple regression model of 38 factors accounted for less than 20 percent of the variance in completing college or dropping out of college for men, and less than 30 percent of the variance in completing college or dropping out of college for women. Bayer found that marriage, family variables, and aptitude measures had the greatest impact on students not completing college (Bayer, 1968)

Pascarella and Terenzini (1977) discovered that students who interacted more frequently with faculty members to discuss course related information, academic programs, and future career opportunities were more likely to persist than students who had fewer interactions with faculty members.

Munro's path analysis of Tinto's model of college persistence accounted for 14 percent of the variation of a student not returning to a university. Munro determined that ethnicity, SES, and sex only had an indirect effect on student retention. Munro also determined that a student's commitment to graduate from college had the strongest effect on a student's persistence (Munro, 1981).

Willett and Singer (1991) in their analysis of student retention, used an hypothesized school district made up of 64,106 students ranging from kindergarten to twelfth grade. They used survival analysis to do a grade-by-grade determination of the dropout rate of students. They determined that in kindergarten to the seventh grade the dropout rate for students was below one percent. In the eighth grade the dropout rate increased to just above one percent, and in the ninth grade the dropout rate was above five percent. The highest dropout rate was found in the tenth grade, where it was thirty-one percent. In the eleventh and twelfth grades, the dropout rates were twenty-five percent and twenty percent respectively.

In Allen (1999)'s study of academic performance and persistence, he determined that background variables and the desire to finish college had the largest impact on persistence. He also discovered that financial aid had an impact on college GPA and persistence.

When looking at non-minority students, he determined that academic performance during freshmen year of college, high school rank, and parent's education level had the greatest impact on persistence. The only variable to impact persistence of minority students was high school rank.

Murtaugh et al. (1999) determined that Oregon State University had a retention rate of approximately 60%. Multiple-variable analysis was performed to determine hazard ratios, which is a "student's hazard of withdrawal is multiplied by a unit increase in the predictor" (p 361). The independent variables that were part of the final model are: age (1.05, $p = 0.0075$), residency (1.29, $p < 0.0001$), high school GPA (0.73, $p < 0.0001$), first-quarter GPA (0.49, $p < 0.0001$), freshman orientation (0.79, $p < 0.0001$).

In DesJardins et al. (1999) study of retention in higher education they determined that the time-varying model was better able to predict stop out. A stop out occurred when students did not return to the university at the beginning of the next semester given they completed the previous semester. The time-varying model also showed the effect of time-varying independent variables on stop out over a period of time. It was determined that white students were more likely to stop out than Asian-American students in their first year, and in the third year African-American students are more likely to stop out than white students. No significant difference was found in gender. Students with a disability were more likely to stop out in their fourth year than the general student population. Students with higher GPAs were less likely to stop out, but the effect of higher GPAs on stop out decreases over time. Students with an on campus job were less likely to stop out than students who were employed off campus. Students who received scholarships were less likely to stop out in their first three years. Students, with loans, were more likely to stop out after their first year. High school rank and composite ACT scores had a negative impact on stop out, that is as high school rank and ACT composite scores increased the chance of a student stopping out decreased.

Elkins et al. (2000) discovered that students who made it through what Tinto (1988) referred to as the separation stage were more likely to return to the university for the second semester. They determined that a strong support system and "rejection of attitudes and values" were key to helping a student persist to the second semester. They also found that minority students who did not have a strong support system were less likely to persist

into the second semester. The strong support system necessary for minority students allowed them to better deal with the separation stage.

Hu and John (2001) determined that African American female students were more likely to persist than African American male students. African American student from a lower socioeconomic status were less like to persist. African American students enrolled at research universities were less likely to persist than those enrolled an non research universities. College grade point average was positively related to persistence, that is as GPA increased the chances of student persisted in school increased. The results, when looking at Hispanic students, were similar to those of African American students. When looking at White students, the results where similar to the other two ethnic groups except when looking at students enrolled at research universities. White students enrolled at research universities where more likely to persist. When looking at financial aid they discovered that those who received some type of financial aid persisted longer than those students who did not receive any financial aid.

DesJardins, McCall, et al. (2002) determined that male students were less likely than female students to graduate in year four, but this effect was reversed with the passage of time. Students with lower levels of academic resources and students that were parents were less likely to graduate. A student's desire to get a bachelors degree was positively related to graduating, but the effect of this desire reduced over time. When socioeconomic status was taken into account, minority students were less likely to graduate. Students who received loans and grants were more likely to graduate, but the positive effect of loans and grants decreased over time. Initially student who received work-study were less likely to graduate, but in year 6 this effect was reversed, and work-study had a positive effect on students graduating. College GPA was found to be a very important indicator for college graduation. The effect of college GPA was stronger in the beginning of a student's college career but declined over time.

DesJardins, Ahlburg, and McCall (2002a) found that scholarships had the greatest effect on student retention, followed by work study. The greatest effect of work study was within the first two years, after the second year the effect of work study decreased over time. Receiving loans were then found to have the most effect on retention. Grants were found to have little or no impact on retention. DesJardins et al. then performed simulations

to see how changes in financial aid packages effected student retention. They explored three simulations: students received no financial aid, student loans were converted into scholarships, and students received front-loaded scholarships. The students who did not receive financial aid were less likely to complete a college degree. The first stop out for students who did not receive financial aid occurred sooner then if a student received financial aid. In the second simulation all student loans were converted to scholarships. If all student loans were converted to scholarships, students would be more likely to survive to graduation. When student loans were converted to scholarships the median first stop out time was 11.13 academic terms. In the third simulation students were provided with scholarships and grants in the first two years of college only. The time to first stop out for this simulation was 10.13 academic terms.

DesJardins, Ahlburg, and McCall (2002b) determined that 36.5 percent of students graduated without a stop out. Those students were characterized as white females, who were from out-of-state, had high college GPAs, received less financial aid, had higher high school percentile ranks, and higher ACT scores than students who stopped out. Of the students who had stop outs, 4.8 percent with one stop out graduated, 0.76 percent with two stop outs graduated, and less than 0.01 percent with more than two stop outs graduated. Grade point average and graduation had a positive relationship, and students with higher GPAs were less likely to stop out. When looking at stop out and graduation at the same time, income had little or no significant impact on graduation. The only component of financial aid to have a significant impact on graduation was work study, but receiving some form of financial aid did help prevent stop outs.

Ishitani and DesJardins (2002 - 2003) determined the greatest risk of dropping out came "at the end of the spring semesters in a student's first and third years." No significant difference was found in gender, but they did discover that white student were more likely to dropout in their first year than Asian students which was similar to DesJardins et al. (1999). Students from higher income family were less likely to dropout than students from lower income families. Students with low educational aspirations were more likely to dropout of school; however, this was only true for year one. After year one there was very little significant difference. Students with higher GPAs and SAT scores were more likely to stay in school. Financial aid had a negative impact on dropout, that is the more financial

aid a student received the less likely they were to dropout of school. They also determined that employment did not have a significant impact on dropout.

St. John et al. (2004) found that African American students majoring in business, health, and engineering/computer science were more likely to persist than African American students who majored in other fields. They also found that white freshmen who were undecided or who were majoring in social sciences were less likely to persist.

The model put forth by Glynn et al. (2005-2006) was able to accurately predict students who would drop out 80 percent of the time. The probability of a student dropping out was based on time of matriculation.

Gansemer-Topf and Schuh (2006) determined that institutional selectivity and institutional expenditures, especially those that supported students' academic integration were found to help improve retention and graduation rates.

DesJardins et al. (2006) discovered that of the $4,490$ students who graduated 76 percent graduated without have a stop out, 12 percent graduated with one stop out, 7.7 percent graduated with two stop outs, and 4.6 percent graduated with three stop outs. Without looking at the student who graduated, the average time until a students first stop out was 7 terms, and excluding those who dropped out of the university the average length of a stop out was 2.6 terms. African American students were more likely to dropout and less likely to graduate than white and Asian American students. Students from families with lower incomes were more likely to dropout and less likely to graduate than students from middle income families and students from higher income families. Students, in the higher percentile of high school rank, were more likely more to graduate without a stop out. Students, with higher ACT scores, were more likely to not have a stop out and more likely to graduate. If students had a first stop out they were more likely to return if they had higher college GPAs and ACT scores and came from middle income or higher income families. The length of time of a stop out had a negative impact on reenrolling at the university.

Ishitani and Snider (2006) discovered that students who were the first to go to college in their families were more likely to dropout than students where parents both had graduated from college. When looking at ethnic groups, using Caucasian students as the reference group, Ishitani and Snider discovered that Hispanic American students were 32 percent more likely to dropout, Native American students where 42 percent more likely

to dropout, African American students were 32 percent more likely to dropout, and Asian American students were 32 percent less likely to dropout. Parents' educational level and students' income level had a positive effect on student retention. Students in lower high school ranking quintile were more likely to dropout of college. Students who participated in ACT/SAT preparation classes were 33 percent more likely not to dropout, but students who received assistance in preparing a financial aid forms were 21 percent more likely to dropout of school. Students, whose parents were involved in helping to choose a college, were more likely to not dropout of school.

Noble et al. (2007 - 2008) determined that the ESSENCE program had a positive impact on GPAs and five-year graduation rates of first year students who took part in the program. In particular, the ESSENCE program raised the GPAs and increased the graduation rates of females and minority students.

Much has been done to study the retention of students in higher education. This study used a multilevel discrete-time hazard model which brought new insight to why students are not persisting.

Recommendation Found in the Literature to Improve Retention

This section summarizes the recommendations found in the literature that can be used to improve retention in higher education.

In Bayer's study done in 1968, he recommended the studying of the various subgroups within the dropout group. He suggested by studying these groups a university would be able to better counsel individuals about their educational future. His study further indicated by looking at the subgroups in question a university would be able to adjust its admission standards (Bayer, 1968).

Pascarella and Terenzini (1977) suggested more informal contact between faculty and students that dealt with course related material to increase persistence.

Elkins et al. (2000) found the importance of support along with the "rejection of attitudes and values" were important factors to improve student retention because students were better able to make it though the separation stage. They suggested that universities create orientation programs that lasted throughout the year. They recommended that students and parents visit college campuses together and to involve students on campus at an early age.

Hu and John (2001) suggested that public policy should focus on making college more affordable. They also recommended the creation of intervention programs that focused on enhancing the college achievement of minority students. The intervention programs would give minority students the opportunity to enhance his/her academic skills.

Ishitani and DesJardins (2002 - 2003) advocated the need for more cooperative work between the K-12 sector and higher education. They suggested institutions develop models for the different factors the led to the risk of dropping out. The development of these models would help administrators responsible for retention, to develop profiles of at-risk students.

St. John et al. (2004) advised policymakers to consider the importance of student aid in ensuring diversity. They proposed that campus leaders consider how the positive link between academic programs and employment opportunities be used to increase the persistence of African American students. They also advocated the use of minority programs in science and engineering, and the hiring of more minority faculty to improve diversity on predominately white campuses.

Daempfle (2003-2004) suggested that students should be actively involved in lectures in order to improve student attitudes, achievement, and retention. He also recommended the use of workshops to overcome the academic shortcomings of some students. He also advocated the hiring of faculty who would work to improve communication with students.

Glynn et al. (2005-2006) recommended that data be collected on student matriculation and use this information to identify those students at risk of dropping out; in order to facilitate contact and offer assistance to those students.

Gansemer-Topf and Schuh (2006) suggested that universities allocate funds that support the institution's mission, instruction, and academic support services in order to improve retention rates.

Ishitani and Snider (2006) recommended that students take part in ACT/SAT preparation classes. They also recommended that high school guidance counsellors determine the risk of a student not persisting in college in order to better advise these students. Finally,

they suggested that universities have various levels of interventions for incoming freshmen based on the risk of those students not persisting.

Multilevel Survival Analysis

Traditional methods of statistical analysis assume that observations behaved independently, but it was possible that observations from similar environments showed similar behaviors as opposed to observations from a different environment. Thus traditional methods of statistical analysis ignored the hierarchical structure of the data which caused underestimation of standard error, which in turn led to a greater chance of making a Type I error. If traditional methods of data analysis were used to evaluate observations with different characteristics from similar environments a researcher got results that were not what they were looking for, because the multilevel nature of the data was not account for in the traditional method of analysis (Barber, Murphy, Axinn, & Maples, 2000). These were two reasons why multilevel analysis should be done on multilevel data sets. Little work has been done to explore the multilevel nature of retention data, when using discrete-time hazard model. What follows are three examples of multilevel discrete-time hazard models.

Reardon, Brennan, and Buka (2002) used multi-level models and event history analysis to study when individuals between the ages of eleven and eighteen began smoking cigarettes. The sample consisted of $1,979$ individuals from seventy-nine neighborhoods in Chicago. They used a multilevel model because the racial makeup of a neighborhood "accounts for roughly half of the difference in age of smoking initiation between Black and White teenagers" (pg 297). The data were gathered from the Project on Human Development in Chicago Neighborhoods, this data set was designed to give researchers the opportunity to study effects of social context and neighborhood characteristics. They used several models. The first model was a simple discrete time model that ignored neighborhood clustering. The later models include the neighborhood variables thus giving the researchers multilevel models.

Ma and Willams (1999) wanted to determine the likelihood that a student would drop out of advanced mathematics, what factors influenced a student decision to drop out of advanced mathematics at each grade level. The study also looked at what was the difference between schools when looking at dropout rates of advance mathematics at each grade level. The study also examined the difference in school dropout rates related to the demographical

characteristics of the school, and could some of the differences between schools be explained by "schooling process." The data for the study were obtained from the Longitudinal Study of American Youth which consisted of $3,116$ students and 52 schools where information was gather from students, teachers, and principals. The student level variables were sex, socioeconomic status, prior mathematics achievement, and prior attitude toward mathematics. The school level variables were principal leadership, academic press, disciplinary climate, teacher autonomy, teacher commitment, material resources for mathematics, general support for mathematics, percentage of Black students, percentage of Hispanic students, and percentage of parents visiting the school. Ma and Williams first constructed a simple survival model and then developed a survival model for each school to determine the dropout rates of the students based on the "school level characteristics."

Barber et al. (2000) developed a discrete-time multilevel hazard model, to show how to use statistical software packages to estimate models that dealt with discrete-time multilevel hazard models. The study also provided information on the assumptions that allow regression coefficients when doing discrete-time multilevel hazard models. Barber et al. studied the timing of permanent contraceptive use. On the individual level they examined if educated women were more likely to use permanent contraceptive methods than uneducated women. On the neighborhood level they examined if the proximate of schooling influenced the hazard of using permanent contraceptive methods. Finally, they examined the relationship between education and permanent contraceptive use in relation to a school in close approximation. The data were collected from 171 neighborhoods in the Chitwan Valley in central Nepal, and all members of the neighborhoods were interviewed.

3. Methods and Procedures

Overview of the Data and Research Design

The Methods and Procedures chapter provides information on the data collection methods and the statistical analysis used in this study. The chapter begins with an overview of the data, the research design and the sampling procedures of how the data were collected. Next is a description of the variables used in this study, followed by the statistical models and finally the data analysis strategies.

The data used in this study was obtained from the National Longitudinal Survey of Youth, 1997 (NLSY97). The National Longitudinal Surveys (NLS) was a group of surveys supported by the Bureau of Labor Statistics (BLS), U.S. Department of Labor. These surveys were used to gather data about the labor market and the different individuals who made up the labor market at several different points in time. The National Longitudinal Survey of Youth, 1997 gathered information using the population of 1997 born between the years 1980 to 1984. The NLSY97 survey consisted of $8,984$ respondents with information on education, employment opportunities, family background, and environmental information (BLS, 2003).

Sampling and Data Collection Methods

The National Longitudinal Survey of Youth, 1997 followed $8,984$ individuals as they transitioned from children to adults and from school to employment. The survey used U.S. residents born between the years 1980 to 1984. The NLSY97 collected data on the following: educational experiences, employment behavior, family background, environmental factors, and participation in government programs. The first interview period for the NLSY97, had interviewers visit randomly selected households to determine all eligible individuals. All individuals, who had permanent residency established at the household, between the ages of 12 to 16 as of December 31, 1996, were eligible for the survey.

The $8,984$ individuals were selected using cluster sampling. The primary sampling units were 147 non-overlapping metropolitan areas, a single county or a group of counties, with $75,291$ households. The survey consisted of two samples, the first was a cross-sectional sample of $6,748$ individuals that represented the U.S. population in 1997, born between the years 1980 through 1984. The second sample consisted of a supplemental sample of $2,236$ individuals from African-American and Hispanic populations born between

Table 2: Demographic Information of the NLSY97 Data Set

	White	African American	American Indian	Asian	Other
Male	2072	1198	28	90	546
Female	2530	1190	33	70	517

the same years. The second sample was taken to get "more reliable statistical analysis of these two groups." The interviewers used a computer-assisted personal interviewing (CAPI) system to gather information. The CAPI system directed the interviews by asking the respondents questions based on their replies to past questions (BLS, 2003).

Table 2 presents demographic information about the NLSY97 sample. White males made up 30.4 percent of the sample and white females made up 28.4. African American males made up 13.5 percent and African American females made up 13.4. American Indian males and females made up 0.31 percent and 0.37 percent respectively. Asian American males and females made up 1.0 percent and 0.79 percent respectively. Males who were classified as other, made up 6.1 percent of the sample and females classified in the same category made up 5.8 percent.

Table 3 provides a list of the college majors available in the NLSY97 data set. Table 4 provides a list of the different types of universities and colleges available in the NLSY97 data set.

Random-Baseline Hazard Models

This study used random-baseline hazard models to determine the effect of level one (individual level) and level two (school level) variables on student retention in higher education. In a random-baseline hazard model, the random effects were found only on the coefficients that determined the shape and level of the baseline logit-hazard curve. This study used two types of random-baseline hazard models. The first was a random-level baseline hazard model. The random-level baseline hazard model had no random effects on the terms that determined the shape of the baseline logit-hazard curve. The second was the random-shape baseline hazard model. The random-shape baseline hazard model had random effects on the terms that determined the shape of the baseline logit-hazard curve.

Table 3: College Majors Available in the NLSY97

NLSY97 Code	College Majors
1	Agriculture and Natural Resources
2	Anthropology
3	Archaeology
4	Architecture and Environmental Design
5	Area Studies
6	Biological Sciences
7	Business Management
8	Communications
9	Computer or Information Science
10	Criminology
11	Economics
12	Education
13	Engineering
14	English
15	Ethnic studies
16	Fine and applied arts
17	Foreign Laguages
18	History
19	Home Economics
20	Interdisciplinary Studies
21	Mathematics
22	Nursing
23	Other Health Professionals
24	Philosophy
25	Physical Sciences
26	Political Science and Government
27	Pre-Dential
28	Pre-Law
29	Pre-Med
30	Pre-Vet
31	Psychology
32	Sociology
33	Theology and Religious Studies
99	0ther field (SPECIFY)

Table 4: School Types in the NLSY97

NLSY97 Code	Type of School
1	Public School
3	Catholic School
4	Private School - other religious affiliation
5	Private School - no religious affiliation

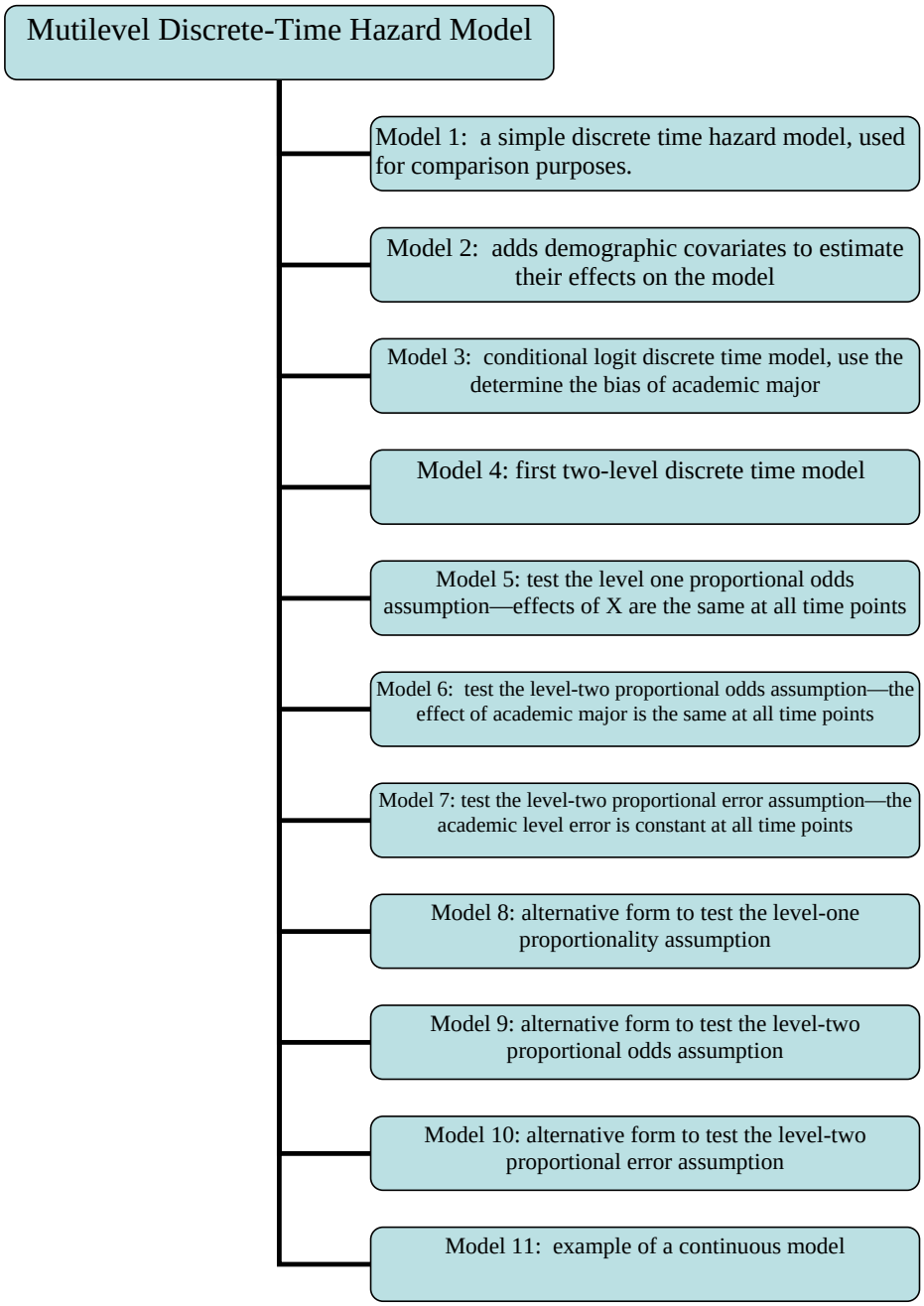

Figure 1: Diagram of the Models Used in This study

Variables

Dependent Variable

- Y: A dichotomous variable that indicated whether or not an individual returned to the university at the beginning of the year. This variable represented the event of interest. It should be noted that a student was considered enrolled as long as he/she was enrolled at a university. Thus if a student transferred to another university he/she was still considered enrolled for this study.

Pre-enrollment Characteristics

- Age: The age of the student when he/she enrolled at the university.
- Gender: The gender of a student was either male or female. Gender was coded 0 for male and 1 for female.
- High School GPA: The final high school grade point average a student earned when they graduated from high school. The high school GPA was on a 4.0 scale.
- ACT Score: The composite score a student received on the ACT test. The ACT composite score ranged from 0 to 36.
- Ethnicity: The ethnicity of a student. In this study ethnicity used were, African-American, American Indian, Asian, white, and other. Ethnicity was coded in the following manner: 0 for non-white and 1 for white.
- Residency: Was the student a resident of the state he/she attended school. Students were either residents of the state or nonresidents of the state where they were enrolled in school.

Enrollment Variables

- Duration: The number of years a student was enrolled at a university. A student's duration was measured as long as they were enrolled at a four-year university, that is, if a student transferred to another university they were still considered enrolled for this study.

- Semester GPA: The university grade point average a student received at the end of each semester. This was a time-varying independent variable.
- Cumulative GPA by Semester: The overall grade point average a student received for his/her entire university career. This was a time-varying independent variable.
- Academic College: The academic college a student was first enrolled in at the university.
- Academic Major: The academic major a student was first enrolled in at the university.
- Remedial Courses: Used to indicate if a student was enrolled in a remedial math or English course at one time during his/her university career. Zero was used to indicate not enrolled in a remedial course and 1 was used to indicate a student was enrolled in a remedial course.

Financial Aid Variables

- Total Financial Aid: The total amount of financial aid received by a student each semester. This was a time-varying independent variable.
- Financial Aid by Type: The amount of financial aid a student received each semester by type of aid. The types of aid were: grant, scholarship, loan, and work study.

School-Level Variable

- School Type: The second-level variable used in this study. School type was a dichotomous variable that indicated if a student was enrolled in a public or private four-year university.

Statistical Methods

Modeling Complex Surveys

Complex surveys are those surveys that have complex sampling strategies such as: stratified random sampling, cluster sampling, multi-stage cluster sampling, or multi-stage stratified random sampling. These types of sampling strategies provide a researcher with several advantages. A complex survey allows a researcher to study a large population in a

Table 5: List of Variables in the Current Study

Variable Status	Variable Name	Variable Type	Description
Dependent	Y_{ij}	Dichotomous	If a student was enrolled or not was enrolled at the university
Pre-enrollment Characteristics			
Time-Stable Covariates	AGE	Continuous	Age of student when first enrolled at the university
	Gender	Categorical	Male/Female
	High School GPA	Continuous	Final High School GPA (0.0 to 4.0)
	ACT Score	Categorical	Composite score on ACT test
	Ethnicity	Categorical	African-American/ Other/ American Indian/White
	Residency	Categorical	Resident of state where university is located
Enrollment Variables			
Time-Varying Covariates	Semester GPA	Continuous	GPA for each semester
Time-Stable Covariates	Academic College	Categorical	College enrolled
	Academic Major	Categorical	Student's Major
	Developmental Course	Categorical	Enrolled in a course
Financial Aid Variables			
Time-Varying Covariates			
	Financial Aid Type	Categorical	Aid by type

cost-effective manner. A complex survey design also incorporates the homogeneous nature of the data found in stratified samples and cluster samples (Lohr, 1999).

Stratification occurs when the population is divided into subpopulations or strata. These subpopulations do not overlap and an individual belongs to only one stratum. A stratified sample prevents a researcher from obtaining a poor sample of the target population. When conducting a stratified sample, a researcher can gain precision for subgroups. A stratified sample often more convenient to give and can lower the cost of the survey. Stratified sampling can result in more precise estimates of the population (Lohr, 1999).

Natural groupings such as households, neighborhoods, schools, or school districts can sometimes occur in a population. Cluster sampling can then be used to sample from these naturally occurring groups. Clusters are known as the primary sampling units. The secondary sampling units are the units of interest that are randomly sampled within each cluster. The data for a study are collected from the secondary sampling units. There are two advantages to cluster sampling. The first advantage is the creation of "a sampling frame list of observations that may be difficult, expensive, or impossible to find." The second advantage is that cluster sampling can be cheaper and more convenient than a simple random sample. A disadvantage of cluster sampling, is that one loses precision of estimates for the population parameters. This is caused because individuals in a cluster are more alike, than individuals selected at random (Lohr, 1999).

The effect of using a more complex sampling design over a simple random sample is measured by the design effect. Design effect provides a measure of the amount of precision gained or lost by using a more complex sampling design over a simple random sample. The formula for the design effect is the following:

$$\text{deff} = \frac{V(\text{estimate from sampling plan with n observations})}{V(\text{estimate from SRS with n observations})}.$$

Thus the design effect is the ratio of the variance of the more complex design over the variance of a simple random sample design. The design effect is important because it helps to protect against making a Type I error (Lohr, 1999).

Lohr (1999) states design-based sampling and model-based sampling are the two approaches used to analyze survey data. In the design-based model "the sampling design

determines how sampling variability is estimated" (p. 82). In model-based, "the *model* determines how variability is estimated, and the sampling design is irrelevant" (p.82). For a design-based model, the variance is found by taking the average squared deviation of the estimate from the mean, "averaged over all samples that could be obtained using a given design" (p. 97). In a model based design the variance is found by the average squared deviation of the estimate from the mean, "but here the average is over all possible samples that could be generated from the population model" (p. 97). The structure used to sample the data is part of the design-based model, and this takes into account the possible complex nature of the sample. In model-based sampling the complex nature of the sampling is not taken into account and each observation is given equal weight. Because of this a design-based model was used in this study.

For this study, a **person-period data set** was created from the NLSY97 data set. This was done because each student must have a record for each period of observation. Longitudinal data are usually stored as a **person-oriented data set.** This study followed the example of Willett and Singer (1993) to convert a person-oriented data set to a person-period data set. In a person-oriented data set, an individual's data are stored as a single record. A person-period data set has multiple lines of data. One line of data for each period the individual is observed. The student's records were distinguished for each time period, this was accomplished be creating a new set of variables. The newly created person-period data set had the following information, on the i^{th} individual with j records: The *time indicators*, a set of dummy variables which indicated the time period the record was taking place in; *the predictors* which were the covariates for individual i at time period t_j; and the *event indicator* which indicated if the event of interest had occurred for the i^{th} individual in time period t_j. The person-period data set also contained information about the amount of time before an individual experienced the event of interest, this was known as the *duration*. The person-period data set also contained information about the *censoring* status of the student and selected predictors.

Survival Analysis

Survival analysis is a method of statistical modeling that deals with the occurrence of events in a longitudinal data set or the timing of events. Survival analysis is known by different names in different areas of study: "event history (sociology), reliability analysis

(engineering), failure time analysis (engineering), duration analysis (economics), and transition analysis (economics)" (Allison, 1995). Survival analysis can be used to study events such as time to equipment failure, time to stock market crashes, or time to job terminations. Survival analysis can also be used to study promotions, births, marriages, earthquakes, and divorces. Survival analysis is such a powerful statistical method because it is able to work with the censoring of data and time-dependent covariates (Allison, 1995).

Willett and Singer (1991) suggested the following advantages to using discrete-time survival analysis. First, discrete-time survival analysis is suited to analyze longitudinal data. Second, discrete-time survival analysis can handle time-invariant and time-variant predictors. Third, violations of the model can easily be tested and corrected. Finally, censored observations can be handled with discrete-time survival analysis.

The **cumulative distribution function** (cdf) of T is defined by $F(t) = Pr(T \leq t)$, gives the probability that a student will drop out before time t, $F(t)$ can also be thought of as the proportion of students in the population that will drop out before time t. We have that $F(t)$ is a cdf if and only if the following conditions hold (Meeker & Escobar, 1998):

1. $\lim_{t \to -\infty} = 0$ and $\lim_{t \to \infty} = 1$.
2. $F(t)$ is a nondecreasing function of t.
3. $F(t)$ is right-continuous, that is, for every t_0, $\lim_{t \downarrow t_0} F(t) = F(t_0)$.

The **probability density function** (pdf) for a continuous random variable T is defined as the derivative of $F(t)$ with respect to t: $F(t) = dF(T)/dt$. The pdf can be used to represent relative frequency of students dropping out as a function of time. A function $F(t)$ is a pdf of a random variable T if and only if the following conditions hold (Meeker & Escobar, 1998):

1. $f(t) \geq 0$ for all t.
2. $\int_{-\infty}^{\infty} f(t)dt = 1$.

The **survival function**(sf) is the complement of the cdf (Meeker & Escobar, 1998),

$$S(t) = Pr(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx$$

and gives the probability of surviving beyond time t.

The **discrete-time hazard probability** is the conditional probability that an individual will experience the event of interest in time interval t_j given that the individual has not experienced the event of interest in any earlier time intervals (Singer & Willett, 1993). That is, $h(t) = \Pr\{T \leq t_j | T > t_{j-1}\}$.

One of the most important reasons for using survival analysis is the ability to censor an observation in a study. **Censoring** occurs when an observation's exact failure time is not known (Allison, 1995).

Discrete-Time Hazard Model

This section gives a brief overview of discrete-time survival analysis. When using a discrete-time hazard model, we must first record events in discrete intervals, that is, we must divide continuous time into an infinite set of contiguous time periods:

$$(0, t_1], (t_1, t_2] \ldots, (t_{j-1}, t_j] \tag{1}$$

where j is the index of the time periods and each time period begins as soon as the last previous period ends (Singer & Willett, 1993).

The discrete-time hazard probability is the conditional probability that an individual will experience the event of interest in time period j given that the individual has not experienced the event of interest in any earlier time intervals (Singer & Willett, 1993). That is,

$$h_j = \Pr\{T \leq t_j | T > t_{j-1}\}.$$

We next add a set of P predictors, Z_p $(p = 1, 2, \ldots, p)$ to the above definition. These predictors allow researchers to characterize the individuals in the population. We denote the P predictors in time period j for the i^{th} individual with the following vector $z_{ij} = [z_{1ij}, z_{2ij}, \ldots, z_{pij}]$. The discrete-time hazard function for individual i, in time period j, with p predictors is the following (Singer & Willett, 1993):

$$h_{ij} = Pr\{T \leq t_j | T > t_{j-1}, Z_{1ij} = z_{1ij}, \ldots, Z_{pij} = z_{pij}\}. \tag{2}$$

The discrete-time hazard function has two very important properties. First, it gives the baseline profile of risk. Second, a shift in the parameters of the discrete-time hazard shows

the effect of the predictors on the baseline profile. Since the h_{ij}s are probabilities, they can be reparameterized so that they have a logistic dependence on the predictors and the time periods. This model is the log-odds of event occurrence as a function of the predictors and has the same two properties, as mentioned above, as the discrete-time hazard function. The discrete-time hazard model can be rewritten as the following:

$$h_{ij} = \frac{1}{1 + e^{-[(\alpha_1 D_{1ij} + \cdots + \alpha_j D_{Jij}) + (\beta_1 Z_{1ij} + \cdots + \beta_P Z_{Pij})]}}, \tag{3}$$

where $[D_{1ij}, D_{2ij}, \ldots, D_{Jij}]$ are a series of dummy variables, with values $[d_{1ij}, d_{2ij}, \ldots, d_{Jij}]$ indexing time periods, J refers to the last time period anyone was observed, $[\alpha_1, \alpha_2, \ldots, \alpha_J]$ are the intercept parameters, and $[\beta_1, \beta_2, \ldots, \beta_P]$ are the slope parameters, which describe the effect of the predictors on the baseline model (Singer & Willett, 1993).

We will now take the logistic transformation of the both sides of 3 and get the following:

$$\ln\left(\frac{h_{ij}}{1 - h_{ij}}\right) = (\alpha_1 D_{1ij} + \cdots + \alpha_j D_{Jij}) + (\beta_1 Z_{1ij} + \cdots + \beta_P Z_{Pij}). \tag{4}$$

Now 4 is the conditional log-odds the event will occur in time period j, given that the event of interest did not occur in any previous time periods. We have that 4 is a linear function of the intercept parameters and the slope parameters (Singer & Willett, 1993).

Next we will let Y_{ij} be an dichotomous indicator variable of the occurrence of the event of interest, that is, y_{ij} is 0 if individual i in time period j does not experience the event of interest and y_{ij} is 1 if individual i in time period j does experience the event of interest. There will also be occurrences when an individual does not experience the event of interest before the observation time ends, and those individuals must be censored. Let C_i be a dichotomous indicator variable which tells if an individual was censored or not. Therefore we have, $c_i = 0$ if individual i has not been censored and $c_i = 1$ if individual i has been censored. Let j_i be the terminal time period, the subscript i indicates that the terminal time period may differ for each individual.

The final part of this overview of the discrete-time hazard model will deal with the construction of the likelihood function. Maximum likelihood functions are used to esti-

mate the parameter $[\alpha_1, \ldots, \alpha_J]$ and $[\beta_1, \ldots, \beta_P]$ in equation 3 and 4 and we therefore get an estimate for h_{ij}. The likelihood function must be constructed in two parts because of censoring. The two parts of the likelihood function deal with first the uncensored individuals, that is, the probability that the individual experienced the event of interest in time period j_i, and the censored individuals, that is, the probability that the individual experienced the event of interest after time period j_i.

We will first work with the probability that an uncensored individual experienced the event of interest in time period j_i. This conditional probability can be written as a product of terms.

$$\Pr\{T = t_{j_i}\} = \Pr\{T = t_{j_i} | T > t_{j_i - 1}\} \ldots \Pr\{T \neq 1 | T_i > 1\}. \tag{5}$$

We now rewrite equation 5 in terms of h_{ij}

$$\begin{aligned} \Pr\{T = t_{j_i}\} &= h_{ij_i}(1 - h_{i(j_i - 1)}) \cdots (1 - h_{i1}) & (6) \\ &= h_{ij_i} \prod_{j=1}^{j_i - 1} (1 - h_{ij}). & (7) \end{aligned}$$

We will now look at the probability that a censored individual will experience the event of interest after period j_i, the construction of this conditional probability is similar to equation 5.

$$\Pr\{T > t_{j_i}\} = \Pr\{T \neq t_{j_i} | T \geq t_{j_i}\} \ldots \Pr\{T \neq 1 | T \geq 1\}. \tag{8}$$

Now equation 8 can also be expressed in terms of h_{ij}

$$\begin{aligned} \Pr\{T > t_{j_i}\} &= (1 - h_{ij_i})(1 - h_{i(j_i - 1)}) \cdots (1 - h_{i1}) & (9) \\ &= \prod_{j=1}^{j_i} (1 - h_{ij}) & (10) \end{aligned}$$

which is the population survivor function. We assume the individuals in the sample are independent, the likelihood function is a product of equitations 5 and 8. Thus we have

$$\mathrm{L} = \prod_{i=1}^{n} [\Pr\{T = t_{j_i}\}]^{1 - c_i} [\Pr\{T > t_{j_i}\}]^{c_i} \tag{11}$$

where c_i takes values of 0 or 1. Now substituting from equation 7 and 10 into equation 11 we get

$$\mathrm{L} = \prod_{i=1}^{n} \left[h_{ij_i} \prod_{i=1}^{j_i-1} (1 - h_{ij}) \right]^{1-c_i} \left[\prod_{j=1}^{j_i} (1 - h_{ij}) \right]^{c_i}. \tag{12}$$

We now take the natural logarithms which gives the log-likelihood function:

$$l = \sum_{i=1}^{n} \left[(1 - c_i) \ln h_{ij_i} + (1 - c_i) \sum_{j=1}^{j_i-1} \ln(1 - h_{ij}) + c_i \sum_{j=1}^{j_i} \ln(1 - h_{ij}) \right], \tag{13}$$

or more simply

$$l = \sum_{i=1}^{n} \left[(1 - c_i) \ln \left(\frac{h_{ij_i}}{1 - h_{ij_i}} \right) + \sum_{j=1}^{j_i} \ln(1 - h_{ij}) \right]. \tag{14}$$

The event-history indicator Y_{ij} can be used with equation 14, and we have the following equation:

$$\sum_{j=1}^{j_i} y_{ij} \ln \left(\frac{h_{ij}}{1 - h_{ij}} \right) = \left\{ \begin{array}{ll} \ln \left(\frac{h_{ij_i}}{1 - h_{ij_i}} \right) & \text{when } c_i = 0 \\ 0 & \text{when } c_i = 0 \end{array} \right\} \tag{15}$$

$$= (1 - c_i) \ln \left(\frac{h_{ij_i}}{1 - h_{ij_i}} \right) \tag{16}$$

$$\tag{17}$$

if we replace the first term in the bracket with equation 14, we will eliminate the censoring indicator for the log-likelihood and replace it with the dichotomous indicator variable y_{ij}. By doing this we have the following equation:

$$l = \sum_{i=1}^{n} \left[\sum_{j=1}^{j_i} y_{ij} \ln \left(\frac{h_{ij_i}}{1 - h_{ij_i}} \right) + \sum_{j=1}^{j_i} \ln(1 - h_{ij}) \right]. \tag{18}$$

Equation 18 can be rewritten as:

$$l = \sum_{i=1}^{n} \sum_{j=1}^{j_i} \left[\ln \left(\frac{h_{ij_i}}{1 - h_{ij_i}} \right)^{y_{ij}} + \ln(1 - h_{ij}) \right]. \tag{19}$$

If we combine like terms and take the antilog we have

$$L = \prod_{i=1}^{n} \prod_{j=1}^{j_i} h_{ij}^{y_{ij}} (1 - h_{ij})^{(1-y_{ij})}. \tag{20}$$

Now by maximizing the likelihood in Equation 20 we get parameter estimates for $\alpha_1, \ldots, \alpha_J$ and $\beta_1, \ldots, \beta_P$.

The Proportional Hazards Model

The proportional hazard model has no assumptions about the nature or shape of the underlying survival distribution, because of this the proportional hazard model is the most general from of a survival regression model. The model uses the hazard rate, instead of the survival time, as a function of the covariates. The proportional hazard model maybe expressed as the following (Hill & Lewicki, 2005):

$$h\{(t), (z_1, z_2, \ldots, z_m)\} = h_0(t) \exp(b_1 z_1 + \ldots + b_m z_m) \tag{21}$$

where $h\{(t), (z_1, z_2, \ldots, z_m)\}$ is the resulting hazard function, with the values of the m covariates for the relevant case $(z_1, z_2, \ldots, z_m)$ and the relevant survival time (t). We have that $h_0(t)$ is an arbitrary and unspecified base-line hazard function. For the base-line hazard function, all the covariates have a value of zero. The proportional hazard model becomes a linear model if both sides of the equation are divided by the base-line hazard function, $h_0(t)$ and then we the natural logarithm of both sides. We then get the following:(Hill & Lewicki, 2005)

$$\log \left[\frac{h(t), (z_1, z_2, \ldots, z_m)}{h_0(t)} \right] = b_1 z_1 + \ldots + b_m z_m. \tag{22}$$

Thus we have a linear model that can be used to estimate parameters.

We have no assumptions about the shape of the underlying hazard function, but equation 21 and equation 22 do imply two assumptions. The first assumption has to do with the relationship between the log-linear function of covariates and the underlying hazard function. These two equations "specify a multiplicative relationship" between the log-linear function of covariates and the underlying hazard function. This assumption is known as the *proportionality assumption.* This means that if given two subjects with different values for

the covariates, the ratio of the hazard functions of the two observations are not dependent on time. The second assumption states there is a log-linear relationship between the underlying hazard function and the covariates (Hill & Lewicki, 2005).

Proportion Hazard Model with Time-Dependent Covariates

One of the assumptions of the proportion hazard model is that the ratio of the estimated hazard functions remain the same over time, but it is often the case that some of the covariates change with time. The changing nature of the covariates causes the validity of this assumption to be suspect. When a researcher has covariates that change over time it is possible to "explicitly define covariates as functions of time." Suppose we separate students into two groups those that receive financial aid and those who do not receive financial aid. We let z be a grouping variable with codes 1, if a student receives financial aid, and 0, if a student does not receive financial aid. We are then able to fit the following proportional hazard model with time dependent covariates:

$$h(t, z) = h_0(t) \exp\{b_1 z + b_2[z \log(t)]\}. \tag{23}$$

In this model, the conditional hazard function at time t is a function of the covariates z, the baseline hazard function h_0, and z times the logarithm of time. A researcher can test the proportionality assumption with a similar model. If the b_2 parameter is significantly different from zero then the effects of the covariates are dependent on time, and thus the proportionality assumption is not met (Hill & Lewicki, 2005).

Hierarchical Linear Models

Hierarchical linear models (HLM) use data from multiple levels in order to find the effect of individual level factors, and the effect grouping level factors have on the individual level factors of interest. The hierarchical linear model closely resembles the ordinary least squares regression model. The base level, or the individual level, of the hierarchical linear model consists of a dependent variable as a function of level one variables along with an intercept. The model is the following:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + r_{ij} \tag{24}$$

where β_{0j} is the intercept of group j, β_{1j} is the slope of variable X_1 of group j, and r_{ij}

is the residual of the i^{th} individual in group j. As one looks at higher levels of the HLM, the individual level slopes and intercept become dependent variables for the second level variables. These models have the following form:

$$B_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \tag{25}$$

$$B_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}. \tag{26}$$

When combining equations 24, 25, and 26 we have the following combined form:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j + u_{0j} + u_{1j} X_{ij} + r_{ij} \tag{27}$$

where we assume the following:

$$\mathrm{E}(r_{ij}) = 0, \qquad \mathrm{Var}(r_{ij} = \sigma^2),$$

$$\mathrm{E} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}, \qquad \mathrm{Var} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \mathbf{T},$$

$$\mathrm{Cov}(u_{0j}, r_{ij}) + \mathrm{Cov}(u_{1j}, r_{ij}) = 0.$$

We have that $i = 1, \ldots, n_j$ level one units nested with $j = 1, \ldots, J$ level two units. Where β_{0j}, β_{1j} are level one coefficients, $\gamma_{00}, \ldots, \gamma_{11}$, are level two coefficients also known as fixed effects, X_{ij} is a level one predictor, W_j is a level two predictor, r_{ij} is a level one random effect, u_{0j}, u_{ij} are level two random effects, σ^2 is the level one variance, and τ_{00}, τ_{01}, τ_{10} τ_{11} are level two variance-covariance components (Raudenbush & Bryk, 2002).

There are several advantages to using multilevel data analysis. First, it allows a researcher to determine the amount of variability caused by each level of data hierarchy. Second, a researcher is able to model the first level of data analysis in terms of effects at all levels. Third, by using a multilevel model a researcher is able to test the possible interactions between each level of data. Finally, the subjects within the data set are similar because they originate from similar environments. This common origin can cause subjects to not be independent (Raudenbush & Bryk, 2002).

Multilevel Discrete Time Proportional Hazard Models

Multilevel discrete time hazard models are a combination of survival analysis and hierarchical linear modeling. Starting in a state of being enrolled at a university, an individual can move into a state of stopout. We let y, the response variable, take the value of zero if a student is still enrolled at the end of a semester and the value of one if the student has stopout at the end of a semester. Thus the hazard function at time t is

$$h_{ijk(t)} = P\{y_{ijk(t)} = 1 | y_{ijk(t-1)} = 0\}$$

where k indexes individuals, j indexes episode, and i indexes the state (Goldstein, 2003).

We will now use a logit link function and the model has the following form:

$$\text{logit}(h_{ijk(t)}) = \beta_0 + \sum_{h=1}^{p} \alpha_n (z_{i(t)})^h + \sum_{l=1}^{m} \beta_l x_{lijk(t)} + u_{ijk} + v_{ik}$$

where $y_{ijk(t)} \sim Bin(1, H_{ijk(t)})$ and "$z_{i(t)}$ indexes for the modelled interval at discrete time t using a p-order polynomial" to illustrate the baseline hazard function, the covariates are represented by $x_{kij(t)}$; v_{ik} and u_{ijk} represent the random effect for the i^{th} state, and the random effect associated with the j^{th} episode for the k^{th} individual respectively. If we assume a level two model with no within-individual-between episode variation; the model has the following form:

$$\text{logit}(h_{ijk(t)}) = \beta_0 + \sum_{h=1}^{p} \alpha_k (z_{i(t)})^h + \sum_{l=1}^{m} \beta_l x_{lijk(t)} + v_{ik}. \tag{28}$$

We will now generalize equation 28 for a level two model

$$\text{logit}(h_{ij(t)}) = (z\alpha)_{i(t)} + (X\beta)_{ij(t)} + u_{ij} \tag{29}$$

where $y_{ij(t)} \sim \text{Bin}(1, h_{ij(t)})$ and $u_{ij} \sim \text{MVN}(0, \Omega_u)$, this is a model of a proportional hazards model. This model can be used for non-proportional hazards by having an interaction between the Z and X variables.

Study Issues

Sample and Population

This study explored the multilevel factors that impact retention of college students. The population used in this study consisted of first-time freshmen at four-year universities. The data on these individuals were gathered from the NLSY1997 longitudinal study.

Data Analysis Procedures

Preliminary Exploration

Exploratory data analysis was first used to inspect the data. This exploratory data analysis consisted of creating frequency tables and a life table.

Data Analysis

Models

Before the multi-level analysis is discussed, it is beneficial to estimate first-level discrete-time hazard models to be used for comparisons reasons. We first estimated a simple discrete-time hazard model with logistic regression that included only a set of duration dummy variables and no intercepts. The model is represented by the following equation:

$$\begin{aligned} \eta_{ijt} &= \ln\left(\frac{h_{ijt}}{1-h_{ijt}}\right) \\ &= \sum \alpha_t(\text{DURATION}_{ijt}), \end{aligned}$$

where h_{ijt} is the hazard of leaving for person i in school type j at year t, and DURATION_{ijt} is a dummy variable for year t for person i in school type j. The estimated coefficients, α_i, give the shape of the baseline logit-hazard curve (Reardon et al., 2002).

In the next model, we added demographic covariates. Model 2 is represented by the following equation:

$$\eta_{ijt} = \sum \alpha_t(\text{DURATION}_{ijt}) + \beta X_{ij}, \tag{30}$$

where X_{ij}, is a vector of time-invariant covariates for student i in school type j. Model 2 was used to estimate the effects of the demographic covariates on the logit-hazard curve without

taking into account school type (Reardon et al., 2002). This model was used to determine the effect of the demographic covariates on the hazard of not returning to a university.

We must now determine if it was possible to ignore the clustering biases, we must estimate a conditional logit discrete-time model:

$$\eta_{ijt} = \sum_{j \in J} \gamma_j + \sum \alpha_t(\text{DURATION}_{ijt}) + \beta X_{ij}, \tag{31}$$

where γ_j was the cluster-specific intercept for cluster j. The estimated coefficients provided by this model give the estimated average effects of within-cluster difference in individual-level covariates. By comparing 30 to 31 we were able to determine to what extent the observed relationships between individual-level characteristics and the hazard rates were caused by the clusters (Reardon et al., 2002).

We will now look at the level two discrete time hazard models. When using a level two discrete-time hazard model one must pay close attention to several proportionality assumptions not necessary for the level one models.

The most basic level two discrete-time hazard model, which was model 4 for this study, has the following form:

$$\begin{aligned} \eta_{ijt} &= \sum \alpha_t(\text{DURATION}_{ijt}) + \beta X_{ij} \\ \alpha_{j0} &= \gamma_{01} Z_j + u_{j0} \qquad , \\ \alpha_{jt} &= \gamma_{t0} \quad \forall t \in \{1, 5\} \end{aligned}$$

where X_{ij} was time-invariant first-level covariate for student i in school type j, and Z_j was a time-invariant second-level covariate for school type j. There were three proportionality assumptions for this model. The first assumption was the *level-one proportional odds assumption*–which dealt with the "effects of X_{ij} on the log-odds of initiation is the same at all time points." This assumption can be tested with the following model:

$$\begin{aligned} \eta_{ijt} &= \alpha_{j0} + \sum \alpha_{jt}(\text{DURATION}_{ijt}) + \sum (\beta X_{ij} \cdot \text{DURATION}_{ijt}) \\ \alpha_{j0} &= \gamma_{01} Z_j + u_{j0} \qquad , \\ \alpha_{jt} &= \gamma_{t0} \quad \forall t \in \{1, 5\} \end{aligned}$$

equation 32 allowed the first-level covariate X to vary across years. "Testing the level-one proportional odds assumption is accomplished by testing the null hypothesis that the β's are equal"(Reardon et al., 2002, p. 308).

"The second proportionality assumption is the *level-two proportional odds assumption*, the assumption that the effect of academic majors is the same at all time points"'(Reardon et al., 2002, p. 308). The level-two assumption can be tested with the following model:

$$\begin{aligned} \eta_{ijt} &= \alpha_{j0} + \sum \alpha_{jt}(\text{DURATION}_{ijt})\beta X_{ij} \\ \alpha_{j0} &= u_{j0} \qquad\qquad\qquad , \\ \alpha_{jt} &= \gamma_{t0} + \gamma_{t1} Z_j \qquad \forall t \in \{1, 5\} \end{aligned}$$

The null hypothesis used to test this assumption was that the coefficients on Z_j were the same across all years (Reardon et al., 2002).

The *level-two proportional error assumption* was the third proportionality assumption. This assumption stated that the second-level error term for school j was constant at all time points. "Another way of stating this assumption is to say that, after controlling for X and Z, the baseline logit hazard curves in the j school type are parallel to one another. This assumption can be tested with the following model:

$$\begin{aligned} \eta_{ijt} &= \alpha_{j0} + \sum \alpha_{jt}(\text{DURATION}_{ijt})\beta X_{ij} \\ \alpha_{j0} &= \gamma_{j0} Z_j \qquad\qquad\qquad , \\ \alpha_{jt} &= \gamma_{t0} + u_{jt} \qquad \forall t \in \{1, 5\} \end{aligned}$$

The null hypothesis used to test this model is that for each of the j school types, the u_js are equal across the semesters (Reardon et al., 2002).

Models $5 - 7$ are difficult to estimate and they may relax the proportionality assumption to much. The proportionality assumptions are an important part in estimating level two discrete-time hazard models, and because of the relaxing of these assumptions by models $5 - 7$ it was better to only relax these assumptions enough that the impact of the covariates and the school type variable's "context are allowed to vary smoothly across" duration (Reardon et al., 2002). For the above reason Models $5-7$ were not estimated in this

study. The following two models were used to test the level-one and level-two proportionality assumptions for a level two model.

In the next model, we used a continuous duration dummy variable and its square. The duration dummy variable was used to define the shape of the baseline hazard curve and the continuous duration dummy variables were used to test the proportionality assumptions of the model. The following model was used to test the level-one proportionality assumption (Reardon et al., 2002):

$$\begin{aligned}
\eta_{ijt} &= \alpha + \sum \alpha_{jt}(\text{DUR}) + \beta_0 X_{ij} + \beta_1(X_{ij} \cdot \text{DURC}_{ijt}) + \beta_2(X_{ij} \cdot \text{DURC}^2_{ijt}) \\
\alpha_{j0} &= \gamma_{01} Z_j + u_{j0} \qquad\qquad . \\
\alpha_{jt} &= \gamma_{t0} \forall t \in \{1, 5\}.
\end{aligned}$$

"Equation 32 allows the relaxation of the level-one proportionality assumption for the variable X, but constrains the effect of X to vary as a quadratic function of duration" (p. 310). To test the level-one assumption of proportionality we tested the null hypothesis that β_1 and β_2 were both equal to zero (Reardon et al., 2002).

We used the following model to test the level-two proportionality assumption:

$$\begin{aligned}
\eta_{ijt} &= \alpha_{j0} + \alpha(\text{DURC}_{ijt}) + \alpha_{j2}\text{DURC}^2 + \sum \alpha_{jt}(\text{DUR}_{ijt}) + \beta X_{ij} \\
\alpha_{j0} &= \gamma_{01} Z_j + u_{j0} \\
\alpha_{j1} &= \gamma_{11} Z_j \qquad\qquad . \\
\alpha_{j2} &= \gamma_{21} Z_j \\
\alpha_{jt} &= \gamma_{t0} \qquad \forall t \in \{1, 5\}
\end{aligned}$$

In order to test the level-two proportionality assumption, we tested the null hypothesis that γ_{11} and γ_{21} are both equal to zero (Reardon et al., 2002).

With model 10, we relaxed the level-two proportional error assumption. We did this by allowing the level-two error terms for school type j to vary smoothly with duration. Model 10 can be seen as an alternative to model 4 and model 7. Model 10 called for us to constrain the logit-hazard error curve to be a smooth function of duration. There were

only three parameters, in Model 10, to define the shape of the logit-hazard error curve in each school type, rather than the 11 required for Model 7. Model 10 can be written in the following multilevel notation:

$$
\begin{aligned}
\eta_{ijt} &= \alpha_{j0} + \alpha_{j1}(\text{DURATIONC}_{ijt}) + \alpha_{j2}(\text{DURATION}^2_{ijt}) + \sum \alpha_{jt}(\text{DURATION}_{ijt}) + \beta X_{ij} \\
\alpha_{j0} &= \gamma_{01} Z_j + \delta_{j0} \\
\alpha_{j1} &= \delta_{j1} \qquad\qquad . \\
\alpha_{j2} &= \delta_{j2} \\
\alpha_{jt} &= \gamma_{t0} \qquad \forall t \in \{1, 5\}
\end{aligned}
$$

where δs are the level-two random effects (Reardon et al., 2002).

> Note that this model, because it included a vector of level two covariates on the intercept term, made the level two proportional odds assumption–it assumed that the effects of the school type covariate was the same at all time points. We could of course, relax this assumption just as we did in model 9. The key difference between model 4 and model 10 was that 10 included the additional random effects δ_{j1} and δ_{j2}. Testing the level-two proportional error assumption in this model was accomplished by testing the null hypothesis that both δ_{j1} and δ_{j2} in Model 10 had zero variance. If we rejected H_0, the level-two proportional error assumptions was invalid and the logit-hazard curves in different school types were not parallel (Reardon et al., 2002, p. 311).

By using the duration dummy variables in models 8 - 10, we were able to have unconstrained estimation of the baseline logit-hazard curve, with the continuous duration variables we were able to test the proportionality assumptions, "under the constraint that any non-proportional effects vary smoothly (quadratically) with duration" (Reardon et al., 2002, p. 311).

In Model 11, we had the assumption that the shape of the baseline logit-hazard curve in each school type j was quadratic in shape, this allowed the quadratic curve to vary across school type. This assumption was stronger than the assumption in model 10 because of this Model 11 was used to test the level-two proportional error assumption. Model 11 has the following form:

$$
\begin{aligned}
\eta_{ijt} &= \alpha_{j0} + \alpha_{j1}(\text{DURATIONC}_{ijt}) + \alpha_{j2}(\text{DURATION}^2_{ijt}) + \beta X_{ij} \\
\alpha_{j0} &= \gamma_{00} + \gamma_{01} Z_j + \delta_{j0} \\
\alpha_{j1} &= \gamma_{10} + \delta_{j1} \\
\alpha_{j2} &= \gamma_{20} + \delta_{j2}.
\end{aligned}
$$

Using a continuous duration variable, allowed researchers to test the level-two proportional error assumption, but it also had several other advantages over the dummy variables. A continuous duration variable had fewer parameters to estimate and it was more parsimonious. Second, when using sparse amounts of data the continuous duration variable had stronger assumptions. Also when working with sparse amounts of data, estimations were easier because of the constraining of the hazard curve to a simple functional form. It was also important to remember that Model ll "is valid only to the extent that the specification of the shape of the logit-hazard curve was reasonable (Reardon et al., 2002).

> Models 4 - 11 were all variations of what we might call the *random-baseline hazard models;*, that is, the only random effects in the models were on the coefficients that determined the level or shape of the baseline logit-hazard curve (the αs). More specifically, we might call Models 4 - 6, 8, and 9 *random-level baseline hazard models,* since they included no random effects on the terms that specify the shape of the baseline logit-hazard curve. Similarly, we might call Model 7, 10, and 11 *random-shape baseline hazard models,* since they included random effects on the terms defining the shape of the baseline logit-hazard curve (Reardon et al., 2002, p. 313).

Chapter Summary

This chapter included the sampling procedures, an overview of the data, including the research design. The variable selection, and the data analysis that was performed is also included. Details of survival analysis, hierarchical linear modeling, and discrete multilevel hazards analysis was provided. The issues involved when analyzing a complex data set have also been addressed.

4. Summary of Results

This chapter describes the findings from this study, and includes the following sections:

- Characteristics of the Sample
- Results of the Exploratory Data Analysis
- Level One Discrete-Time Hazard Models Estimates and Results
- Level Two Discrete-Time Hazard Model Estimates and Results

Characteristics of the Sample

The sample used in this study was taken from the National Longitudinal Survey of Youth, 1997 (NLSY97). The sample was made up of students who took part in Round 3 through Round 8, which occurred during 1999 through 2004. The data used in this study, centered around the time the participants were undergraduate students.

The NLSY97 data set consisted of $8,984$ individuals between the ages of 12 to 16 as of December 31, 1996. The participants were selected using cluster sampling. The primary sampling units consisted of non-overlapping metropolitan areas, a single county or a group of counties; households were the secondary sampling unit. Thus the NLSY97 contained a cross-sectional sample of $6,748$ individuals representative of the U.S. population as of 1997, and a supplemental sample of $2,236$ individuals from African-American and Hispanic populations.

The sample, used in this study, consisted of N $= 3,072$ students who were enrolled at a four-year university beginning in 1999. Several of the independent variables had extensive data missing and thus had to be removed from the study. These independent variables were removed because convergence could not be reached in the PROC LOGISTIC and the PROC NLMIXED SAS procedures.

Table 6 summarizes the variables removed from the analysis. The highest percentage of missing observations, for government loans, was seen in the year 2004 with 71.62 percent of the data missing. In the years $2000 - 2003$, over 60 percent of the observations

Table 6: List of Independent Variables Removed from Study

Variables	Year	Percentage Missing
Government Loan	1999	97.6
	2000	65.29
	2001	66.89
	2002	66.56
	2003	64.97
	2004	71.02
Work Study	1999	98.88
	2000	100.00
	2001	99.36
	2002	90.76
	2003	92.36
	2004	94.59
Grants and Scholarships	1999	95.7
	2000	40.45
	2001	49.68
	2002	60.83
	2003	57.96
	2004	68.47
Grade Point Average(GPA)	1999	98.81
	2000	20.38
	2001	19.11
	2002	26.75
	2003	29.20
	2004	40.44

were missing these data. When looking at work study, in each year over 90 percent of the observations were missing these data. For grants and scholarships the largest percentage of missing data was found in the year 2004 with 68.47 percent of the data missing. The largest percentage of missing data, for grade point average, can be found in the year 2004, with 40.44 percent of the observations missing these data.

Table 7 summarizes the demographic variables, used in this study. In Table 7 we see that females made up 55.43 percent of the sample and males made up 44.57 percent. Looking at race, we have that white students made up 66.93 percent of the sample and African American, American Indian, Asian, and other made up 21.57 percent, 0.46 percent, 3.22 percent, and 7.83 percent respectively. Due to the larger percentage of white students

Table 7: List the Demographic Variables Used in the Study

Variable	N	Percent
Gender		
Male	1382	44.57
Female	1719	55.43
Race		
White	2060	66.93
African American	664	21.57
American Indian	14	0.46
Asian	99	3.22
Other	241	7.83

in the sample, the race variable was changed into a dichotomous variable. This new variable was called ethnicity with a value of one (1) if the student was white and a value of zero (0) if the student was non-white.

The time-varying independent variables that were to be used in the study were: enrollment in remedial math, remedial English, and if a student lived on campus or off campus. There were only 315 individuals who had these variables in common; because of this small number of individuals these variables were also removed from the analysis. Thus the individual-level variables used in this study were gender, ethnicity, and the school-level variable was school-type. The missing data found in the financial aid variables, grade point average, remedial courses, and living on or off campus caused a reduction in the independent variables.

The information found in Table 8 showed how the data originally looked at the beginning of this study. Table 8 depicts a person-oriented data set. In a person-oriented data set, there is only one record of each individual in the study. In order to analyze the data using a discrete-time hazard model the person-oriented data set must be reformatted into a person-period data set. A person-period data set is a data set where each individual in the study has multiple lines of data, one line of data for each period the individual is observed.

Table 8: Person-Oriented Data Set

ID	Race	Type	Duration
31	4	1	5
70	2	1	4
75	4	1	5
107	1	3	3
121	1	3	2
128	1	1	5
130	4	1	1
135	1	1	6
156	4	1	4
176	5	1	5

The first column in Table 8 is the ID number for each student in the study. The next column indicates the student's race. Column three indicates whether or not the student attended a public or private four-year university. The last column is the amount of time in years a student was enrolled.

Table 9 illustrates how the data was reformatted into the person-period data set. A person-period data set has a line of data for each year the student was part of the study. Looking at Table 9, we see observation 31 was in the study for five-years, but observation 121 was in the study for only two-years. Table 9 has the same information as found in Table 8, with some new variables created to be used in the analysis. Those variables were: Y, the dependent variable; D01 through D06 a set of dummy variables used to indicate the year of enrollment, duration squared, and the censor indicator variable.

Exploratory Data Analysis

Table 10 shows the enrollment patterns of the $3,072$ students used in this study. It also shows that when these students ceased to be enrolled at a four-year university between the years 1999 and 2004, when the collection of data ended. The first column shows the length of time in years a student was enrolled at a four-year university. The three columns that follow show the number of students enrolled at the beginning of each year, the number of students who did not return at the beginning of the next year, and the number of students

Table 9: Person-Period Data Set

ID	Y	D01	D02	D03	D04	D05	D06	ST	D	D^2	Censor
31	0	1	0	0	0	0	0	1	5	25	0
31	0	0	1	0	0	0	0	1	5	25	0
31	0	0	0	1	0	0	0	1	5	25	0
31	0	0	0	0	1	0	0	1	5	25	0
31	1	0	0	0	0	1	0	1	5	25	0
31	1	0	0	0	0	0	0	1	5	25	0
70	0	1	0	0	0	0	0	1	5	25	0
70	0	0	1	0	0	0	0	1	5	25	0
70	0	0	0	1	0	0	0	1	5	25	0
70	0	0	0	0	1	0	0	1	5	25	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
121	0	1	0	0	0	0	1	0	2	4	1
121	0	0	1	0	0	0	0	1	2	4	1

Table 10: Life Table Describing the Number of Years a Student is Enrolled

	Number			Proportion of	
Year	Enrolled at the beginning of the year	Who left during the year	Censored at the end of the year	All students still enrolled at the end of the year	Students at the beginning of the year & left during year
0	3072			1.0000	
1	3072	115	8	0.9626	0.0374
2	2949	111	21	0.9263	0.0376
3	2817	177	28	0.8681	0.0628
4	2612	256	87	0.7831	0.0980
5	2269	358	6	0.6595	0.1578
6	1905	1070	835	0.2891	0.5617

who were censored at the end of the year. Censored students were those students who were still enrolled when the study ended. When the study ended 2,082 students did not continue their education at a four-year university, and 985 students were still enrolled.

The proportion of all students still enrolled at the end of each year is presented in column five of Table 10. The information presented in the fifth column is also known as the *survival probability*. The survival function was the proportion of students that survived through each year of the study. Looking at column five, we see that 96.26 percent of the sample was still enrolled at the end of year one, and we see that 28.91 percent of the sample was still enrolled at the end of the study.

The last column of Table 10 is the proportion of students known to be enrolled at the beginning of the year and did not return to the university at the beginning of the next year. The number of students at the beginning of each year was known as the risk set. The proportion of students who left the university by the end of the year was known as the hazard probability. Looking at Table 10 we have that 3.74 percent of the 3072 students did not return to a university the next year. We also have that 3.76 percent, 6.28 percent, 9.80 percent, 15.78 percent and 56.17 percent of the students did not return to a university in years two, three, four, five, and six respectively.

Person-Level Discrete-Time Hazard Models

The main objective of this study was to demonstrate how a multilevel discrete-time hazard model could be used to determine possible factors that influenced student retention in higher education. This objective was achieved by first constructing a series of discrete-time hazard models using the SAS PROC LOGISTIC procedure. A discrete-time hazard model was used because it is suited to analyze longitudinal data, it can incorporate both time-invariant and time-varying predictors, violations of the model can easily be tested and corrected, and censored observations are part of the analysis.

The first model was a simple discrete-time hazard model. This model was used as the baseline model for the level one models. The next model was a discrete-time hazard model with demographic covariates.

Table 11: The Estimates from the Model 1 Logistic Regression

			Standard	Waid		Point	95% Waid	
Parameter	DF	Est	Error	Chi-Square	Pr > χ^2	Est	Conf.	Limits
D01	1	-3.2470	0.0950	1167.0636	<0.0001	0.039	0.032	0.047
D02	1	-3.2413	0.0968	1122.2974	<0.0001	0.039	0.032	0.047
D03	1	-2.7024	0.0776	1211.3918	<0.0001	0.067	0.058	0.078
D04	1	-2.2195	0.0658	1137.5467	<0.0001	0.107	0.096	0.124
D05	1	-1.6748	0.0576	845.7860	<0.0001	0.187	0.167	0.210

Simple Discrete-Time Hazard Model

The first model to be fitted was a simple discrete-time hazard model. This model was fitted as a logistic regression with no intercept and only a set of duration dummy variables. This model was used to estimate $\alpha_1, \ldots, \alpha_5$ and gave the shape of the baseline logit-hazard curve. This model was used as the baseline model for the level one hazard models.

Table 11 gives the parameter estimates for Model 1. The estimates found in Table 11 were the parameter estimates for the time-indicator variables. These time indicator variables allowed for the estimation of the risk of a student leaving a university each year. In year one, we had $\widehat{\alpha_1} = -3.25$ (s.e. $= 0.095$, $p < 0.0001$). The estimate of α_1 gave an estimate of h_1 to be $\widehat{h_1} = 0.04$. Thus for all students in their first year of enrollment we estimated there was a 3.74 percent risk of leaving a university. For year two, we had $\widehat{\alpha_2} = -3.24$ (s.e. $=$ 0.0376, $p < 0.0001$); therefore, students had a 3.76 percent risk of leaving a university in their second year. Next we looked at the parameter estimate for year three, we had $\widehat{\alpha_3} = -2.7$ (s.e = 0.0776, $p < 0.0001$). This gave a $\widehat{h_3} = 0.06$, which implied there was a 6.28 percent risk of a student not returning to a university. We had a parameter estimate of $\widehat{\alpha_4} = -2.22$, which implied there was a 9.80 percent risk of a student not returning to a university in their fourth year. In a student's fifth year, we had a parameter estimate of $\widehat{\alpha_5} = -1.67$ (s.e. $=$ 0.0576, $p < 0.0001$) which estimated a 15.78 percent risk of a student not returning. From the results of model 1, we determined each year a student was enrolled at a university the risk of the student not returning increased.

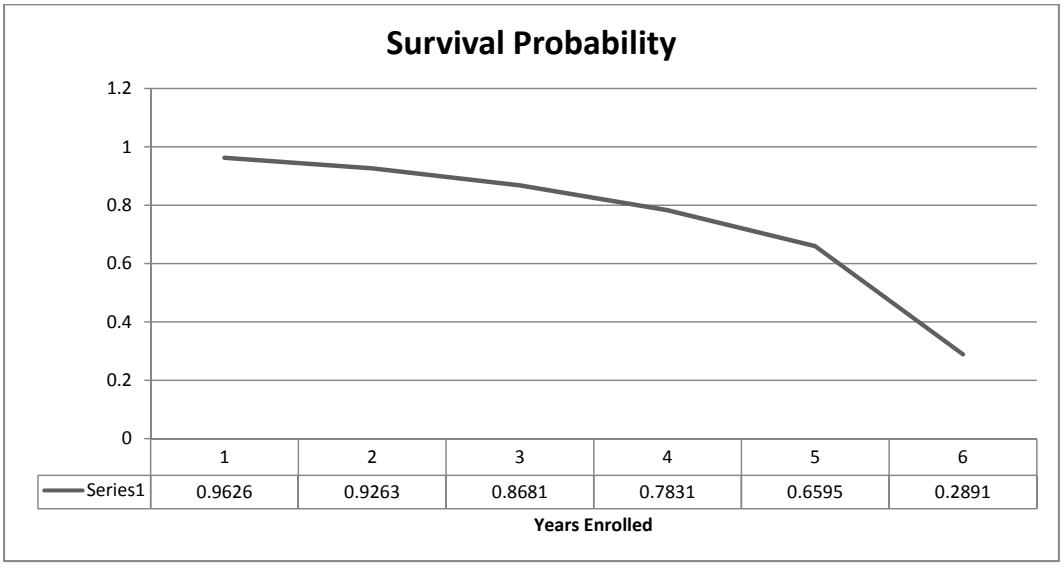

Figure 2: Survival Probability Curve

Figure 2 and Figure 3 are graphical representations of the survival probability curve and the hazard probability curve for Model 1. From the survival probability curve, we see the median duration occurred between year five and year six. We also see that the survival probability had a sharp drop at the end of year five. The hazard curve was an increasing curve which implied for each year a student was enrolled the risk of students not completing the year increased.

Discrete-Time Hazard Model with Demographic Covariates

The second model in this study was a discrete-time hazard model with demographic covariates. Two demographic covariates were added to the model. The first covariate was gender, male or female. The second covariate was ethnicity, white or non-white. Table 12 gives the maximum likelihood estimates for the second model. Similar to the first model, the estimates for the time indicator variables were significant at the $p < 0.0001$ significance level. Ethnicity was found to be significant at the 0.01 significance level, and gender was found to be significant at the 0.1 significance level. We had $\widehat{\alpha_1} = -3.39$, (s.e. $= 0.1014$, $p < 0.0001$) which implied there was a 3.27 percent risk of a student not returning to a

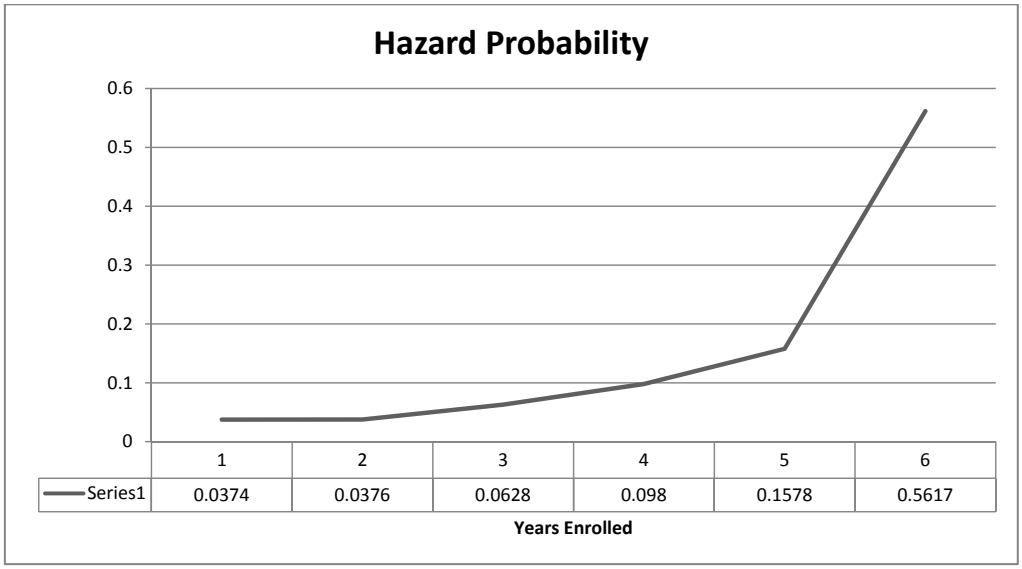

Figure 3: Hazard Probability Curve

university. In the second year, we had $\widehat{\alpha_2} = -3.39$ (s.e. $= 0.1034$, $p < 0.0001$) which gave a 3.25 percent risk of not returning to a university. Year three gave a parameter estimate of $\widehat{\alpha_3} = -2.85$ (s.e. $= 0.0855$, $p < 0.0001$), which implied a 5.47 percent risk of a student not returning to a university. In year four we had $\widehat{\alpha_4} = -2.38$ (s.e. $= 0.0751$, $p < 0.0001$), which gave a risk of 8.53 percent of not returning. We had an estimate of $\widehat{\alpha_5} = -1.83$ (s.e. $= 0.0482$, $p < 0.0001$) which implied there was a 13.92 percent risk of a student not returning to a university.

Table 12: The Estimates from the Model 2 Logistic Regression

			Standard	Waid		Point	95% Waid	
Parameter	DF	Est	Error	Chi-Square	Pr > χ^2	Est	Conf Limits	
D01	1	-3.3868	0.1014	1115.82	<0.0001	0.034	0.028	0.041
D02	1	-3.3903	0.1034	1075.15	<0.0001	0.034	0.028	0.041
D03	1	-2.8482	0.0855	1109.20	<0.0001	0.058	0.049	0.069
D04	1	-2.3729	0.0751	998.73	<0.0001	0.093	0.080	0.108
D05	1	-1.8219	0.0678	722.11	<0.0001	0.162	0.142	0.185
Gender	1	0.0921	0.0482	3.66	0.0559	1.096	0.998	1.205
Ethnicity	1	0.1396	0.0469	8.87	0.0029	1.150	1.049	1.260

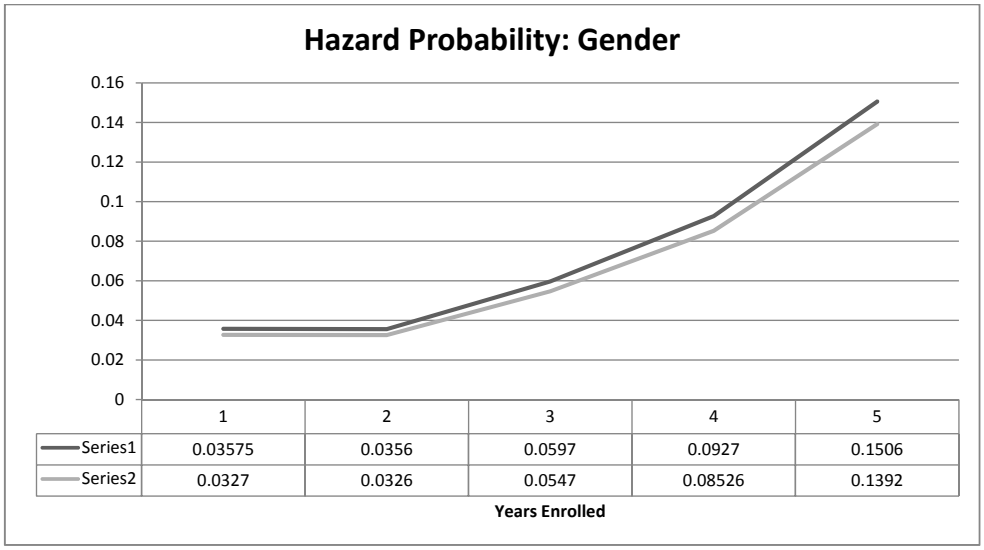

Figure 4: Hazard Probability Curves for Gender

The two demographic variables were found to be significantly different from zero; which implied there was a significant difference in the hazard functions between male and female students, and white and non-white students. When looking at gender, there was significant difference in the hazard function between male and female students. We had $\widehat{\beta_1} = 0.09$ (s.e. $= 0.0482$, $p = 0.0559$), this implied the graph of the fitted logit-hazard function for female students was elevated above the fitted logit-hazard function for male students. Now if we take the antilog of $\widehat{\beta_1}$ we get the estimated odds of a student leaving a four-year university. We found the estimated odds of a female student leaving in any given year was 1.10 times greater than a male student. We had a $\widehat{\beta_2} = 0.14$ (s.e. $= 0.0469$, $p = 0.0029$), which was the parameter estimate for ethnicity. Thus we had ethnicity was significantly different from zero. We had the fitted logit-hazard curve for white students was elevated above the fitted logit-hazard curve of non-white students. We estimated the odds of a white student leaving a four-year university in a given year to be 1.15 times greater than a non-white student.

Figure 4 is the hazard probability curve for gender, the series 1 curve represented female students and the series 2 curve represented male students. The hazard probability

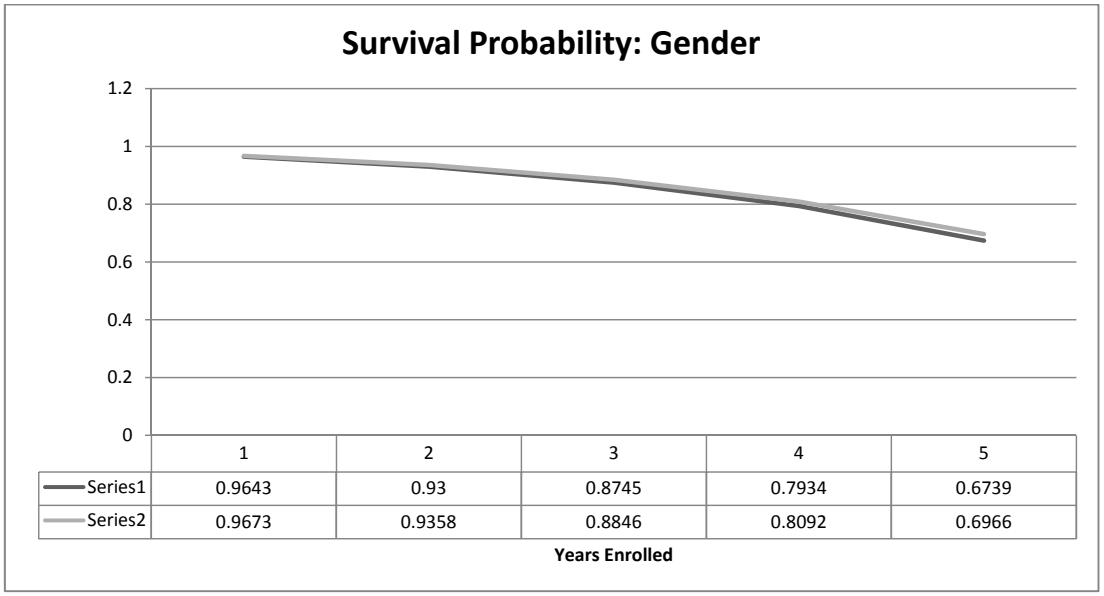

Figure 5: Survival Probability Curves for Gender

curve for male students was used as a baseline curve. The baseline hazard curve was used to demonstrate if there was a difference between hazard curves of male and female students.

The hazard probability curves for male and female students in Figure 4, were both parallel. These two curves were parallel because of the proportional-odds assumption that was part of the logistic model. Because $\widehat{\beta_1}$ was found to be significant we had separation between the hazard curves of male and female students, and because $\widehat{\beta_1}$ was positive we had the female students' hazard curve to be elevated above the male students' hazard curve. This implied that female students had a greater risk of not returning to a university at the end of each year than male students. We also had that the hazard curves were increasing which implied for each year a student was enrolled the risk of not returning at the end of the year increased.

Figure 5 is the fitted survival plot for male and female students. We had little separation between the survival curves in male and female students. We had female students represented by series 1 and male student represented by series 2. The two survival curves were both decreasing which implied for each year a student was enrolled the probability of

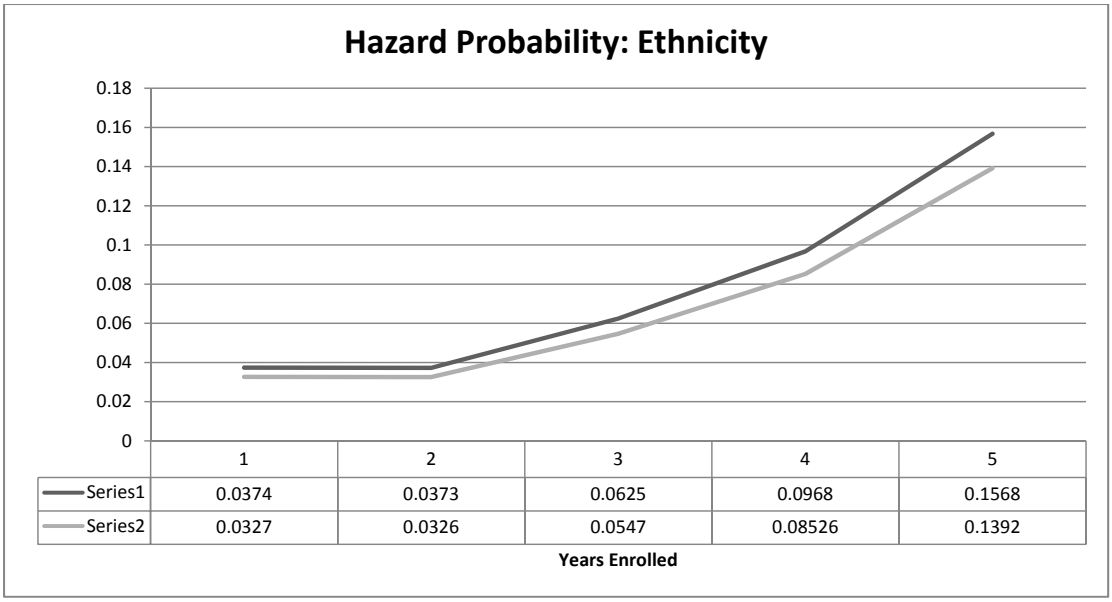

Figure 6: Hazard Probability Curves for Ethnicity

survival decreased. The survival curve for female students was lower than the survival curve for male students, this implied the survival probability for female students was smaller than the survival probability for male students.

Figure 6 was the Model 2 hazard curves for ethnicity. The series 1 curve represented white students and the series 2 curve represented non-white students. Similar to the results for gender, the hazard curves for white and non-white students were parallel. Since the estimate for $\widehat{\beta}_2$ was significantly different from zero, we had separation between the hazard curves of white and non-white students. Also because the estimate of $\widehat{\beta}_2$ was positive we had that the hazard curve for white students was elevated above non-white students. This implied white students were at greater risk of not returning to a university at the end of each year than non-white students. We also saw the hazard curves were increasing which implied, for each year a student was enrolled the risk of not returning increased for both white and non-white students.

Figure 7 is the survival curves of white and non-white students. We had white students represented by the series 1 curve and non-white students represented by the series

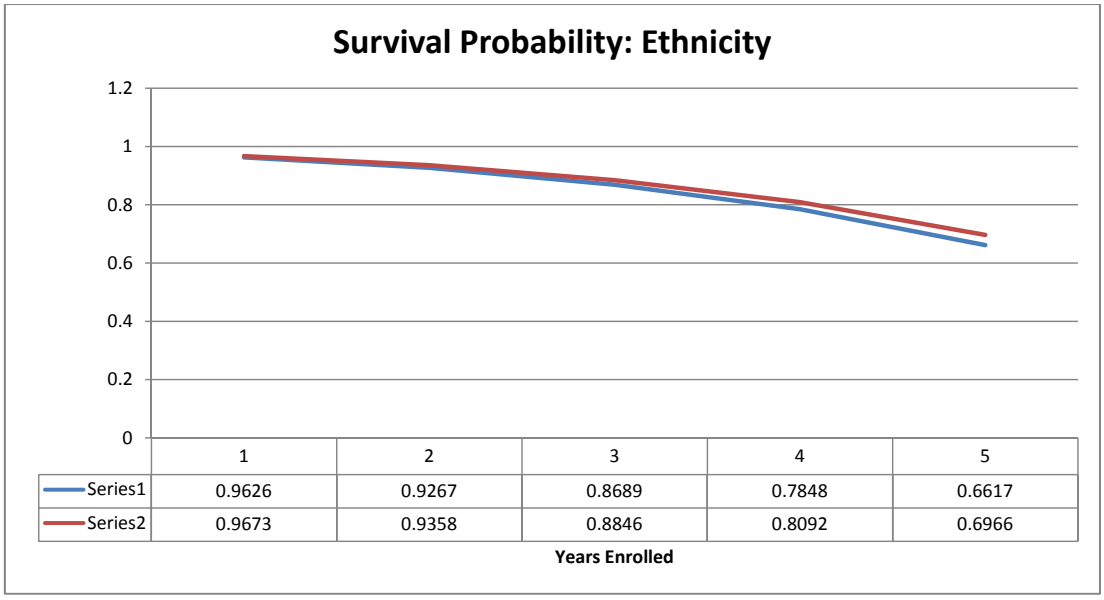

Figure 7: Survival Probability Curves for Ethnicity

2 curve. We saw at the beginning of the study the survival curves were overlapped, but as the years increased the separation between the survival curves of white and non-white students increased. We had that both survival curves were decreasing, this implied as the years increased the probability of surviving decreased.

Brief Summary of Results of the Person-Level Discrete-Time Hazard Models

In all the models, we found the hazard function increased for each year a student was enrolled, that is, the longer a student was enrolled the greater the risk of the student leaving before the end of the year. The time-indicator variables were found to be significant in all the models. In Model 2, a significant difference was found in the hazard functions between male and female students, and white and non-white students.

Two-Level Discrete-Time Hazard Models

A two-level model was used to determine the influence of one school-level variable in our analysis. The school-level variable was school type. School type was a dichotomous variable that indicated if a student was enrolled in a public or private four-year university.

The individual-level model and the neighborhood level model had the following structures.

The individual-level model:

$$y_{ijt} = \beta_{0j} + \beta_{1j}\text{Ethnicity}_i + \beta_2\text{Gender}_i + \beta_3\text{Dur}_{it} + \beta_4\text{Dur}_{it}^2$$

The school-level model

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{School Type}_j + \epsilon_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\text{School Type}_j + \epsilon_{1j} \\
\beta_2 &= \gamma_{20} \\
\beta_3 &= \gamma_{30} \\
\beta_4 &= \gamma_{40}
\end{aligned}$$

where

β_{0j} represented the overall level of retention in school type j which may have varied by being enrolled in a public or private university.

β_{1j} represented the overall effect of ethnicity for school type j which may have varied by being enrolled in a public or private university.

β_2 represented the overall effect of gender.

β_3 represented the overall effect of an interaction with duration.

β_4 represented the overall effect of an interaction with duration squared.

Discrete-Time Hazard Model with Intercept and Demographic Variables

Model 3 was a discrete-time hazard model with an intercept, the time indicator variables, time-invariant covariates: gender and ethnicity. This model was the base model for the level two models.

Model 3 had no second level variables and had an intercept term. Model 3 was similar to Model 2, except it had an intercept. The only parameter estimates that were

Table 13: The Estimates from the Model 3 Logistic Regression

			Standard	Waid		Point	95% Waid	
Parameter	DF	Est	Error	Chi-Square	Pr > χ^2	Est	Conf Limits	
Int	1	0.2320	0.0689	11.3286	0.0008			
D01	1	-3.4900	0.1058	1088.8246	<0.0001	0.030	0.025	0.038
D02	1	-3.4934	0.1077	1052.4188	<0.0001	0.030	0.025	0.038
D03	1	-2.9510	0.0906	1060.2124	<0.0001	0.052	0.044	0.062
D04	1	-2.4751	0.0808	938.1178	<0.0001	0.084	0.072	0.099
D05	1	-1.9237	0.0741	674.4509	<0.0001	0.	0.126	0.169
Gender	1	0.0062	0.0544	0.0128	0.9099	0.851	0.904	1.119
Ethnicity	1	0.0228	0.0579	0.1550	0.6938	1.023	0.913	1.146

significant in Model 3 were the time indicator variables. The time indicator variables were significant at the $p < 0.0001$ significance level.

We had $\widehat{\alpha_1} = -3.49$, (s.e. $= 0.1058$, $p < 0.0001$). Thus for students enrolled in their first year there was a 3.0 percent risk of not returning to a four-year institution. The maximum likelihood estimate for the second year was $\widehat{\alpha_2} = -3.49$, (s.e. $= 0.1058$, $p < 0.0001$), which implied there was a 3.0 percent risk a student left a university in their second year. We had $\widehat{\alpha_3} = -2.95$, (s.e. $= 0.0.906$, $p < 0.0001$) which gave a risk of 4.96 percent that students left in their third year. The maximum likelihood estimate for the fourth year was $\widehat{\alpha_2} = -2.48$, (s.e. $= 0.0808$, $p < 0.0001$), which implied there was a 7.76 percent risk that a student left a university in their fourth year. We had $\widehat{\alpha_3} = -1.92$, (s.e. $= 0.0741$, $p < 0.0001$) which gave a risk of 12.74 percent that students left in their fifth year. Thus we had the risk of a student not returning to a four-year university increased as the number of years they were enrolled increased, that is, the longer a student was enrolled the higher the risk of them not returning.

The two demographic variables were found to have no significance.

Simple Two-Level Discrete-Time Hazard Model

In order for the statistical software package, SAS, to analyze a level two model, the level two model must be written as a level one model.

$$Y_{ijt} = \beta_{j0} + \beta_{1j}\text{Race}_i + \beta_{2j}\text{Gender}_i + \beta_3\text{Duration} + \beta_4\text{Duration}^2$$

This meant that the parameter estimate, the β_i were not estimated directly, but the β_i were found by estimating the γ_i and their error terms.

$$\begin{aligned} Y_{ijt} &= \beta_{j0} + \beta_{1j}\text{Eth}_i + \beta_2\text{Gender}_i + \beta_3\text{Duration}_{ti} + \beta_4\text{Duration}_{ti}^2 \\ &= \gamma_{00} + \gamma_{01}\text{ST}_j + \epsilon_{0j} + (\gamma_{10} + \gamma_{11}\text{ST}_j + \epsilon_{1j})\text{E}_i + \gamma_{20}\text{G}_i + \gamma_{30}\text{D}_{ti} + \gamma_{40}\text{D}_{ti}^2 \\ &= \gamma_{00} + \epsilon_{0j} + (\gamma_{10} + \epsilon_{1j})\text{E}_i + \gamma_{20}\text{G}_i + \gamma_{01}\text{ST}_j + \gamma_{11}\text{ST}_j\text{E}_i + \gamma_{30}\text{D}_{ti} + \gamma_{40}\text{D}_{ti}^2 \end{aligned}$$

where

- $\beta_0' = \gamma_{00} + \epsilon_{0j}$
- $\beta_1' = \gamma_{10} + \epsilon_{1j}$
- $\beta_2 = \gamma_{20}$
- $\beta_3 = \gamma_{30}$
- $\beta_4 = \gamma_{40}$
- $\beta_5' = \gamma_{01}\text{School Type}_j$
- $\beta_6' = \gamma_{11}\text{School Type}_j\text{Ethnicity}_i$

Model 4 was the next level two model to be analyzed. Model 4 was different from Model 3 because the school-level variable was added to the model. The school-level variable indicated whether a student was enrolled in a public or private four-year university. We had the time-indicator variables were significant at the $p < 0.0001$ significance level, and γ_{01} was significant at the $p < 0.0001$ level.

The parameter estimates for Model 4 can be found in Table 14. We had an $\widehat{\alpha_1} = -3.7768$ (s.e. $= 0.33$, $p < 0.001$). This parameter estimate gave a risk of 2.24 percent of a student not returning to the university. In year two, we had $\widehat{\alpha_2} = 3.76$ (s.e. $= 0.1094$, $p < 0.0001$), which gave an estimated hazard function of $\widehat{h_2} = 0.02$. We had $\widehat{\alpha_3} = -3.19$ (s.e.

Table 14: Parameter Estimates for Model 4.

Parameter	Estimate	Standard Error	Pr > |t|
α_{01}	-3.7768	0.1094	< 0.0001
α_2	-3.7589	0.1111	< 0.0001
α_3	-3.1946	0.0943	< 0.0001
α_4	-2.6931	0.0845	< 0.0001
α_5	-2.0903	0.0773	< 0.0001
γ_{00}	1.0217	0.0926	< 0.0001
γ_{10}	-0.0281	0.08434	0.7392
γ_{20}	-0.0009	0.0554	0.8704
γ_{01}	-1.1015	0.0992	< 0.0001
γ_{11}	0.0766	0.1178	0.5154

$= 0.1111$, $p < 0.0001$), which gave an estimated hazard function of $\widehat{h_3} = 0.039$. In year four, we had $\widehat{\alpha_4} = -2.69$ (s.e. $= 0.0845$, $p < 0.0001$). This parameter estimate gave a risk of 6.34 percent of a student not returning to the university. In year five, we had $\widehat{\alpha_5} = -2.09$ (s.e. $=$ 0.0773, $p < 0.0001$). This parameter estimate gave a risk of 11.00 percent of a student not returning to the university. We had $\widehat{\gamma_{00}} = 1.02$ (s.e. $= 0.09260$, $p < 0.0001$), which implied $\widehat{\beta_{00}} = 1.11$. We had $\widehat{\gamma_{01}} = -1.10$ (s.e. $= 0.09917$, $p < 0.0001$), which implied $\widehat{\beta_5'} = -1.00$. The parameter estimate $\widehat{\beta_5'}$ represented the overall effect of school type, with $\widehat{\beta_5'} = -1.00$ we estimated the odds of a student leaving a university to be 0.37 time smaller for student who attended a public university than a student who attended a private university. This implied the risk of not returning to a university for a student who attended a public-four year university was smaller than the risk for a student who attended a private-four year university.

Both gender and ethnicity were found not to be significant, which implied there was no difference in the risk for male and female students, and non-white and white students returning to a university each year.

Tests for Assumptions in the Two-Level Discrete-Time Hazard Model

The level-one proportionality odds assumption was tested using models 5 and 6. The level-two proportionality odds assumption was tested using Model 7. The level-one proportionality odds assumption, for ethnicity, was tested by looking at the interactions

Table 15: Parameter Estimates for Model 5.

Parameter	Estimate	Standard Error	$\Pr > \|t\|$
α_{01}	-7.8585	0.2342	< 0.0001
α_2	-6.7998	0.1890	< 0.0001
α_3	-5.2391	0.1461	< 0.0001
α_4	-3.8794	0.1125	< 0.0001
α_5	-2.5152	0.0857	< 0.0001
γ_{00}	2.1185	0.118	< 0.0001
γ_{10}	7.8687	0.4768	0.7392
γ_{20}	0.0394	0.0611	0.5187
γ_{30}	-1.0770	0.2136	< 0.0001
γ_{40}	-0.0921	0.0247	0.0002
γ_{01}	-1.7425	0.1228	< 0.0001
γ_{11}	1.4268	0.1428	< 0.0001

between ethnicity and the two continuous variables duration and duration squared. That is, we tested the following null hypothesis $H_0 : \beta_3 = \beta_4 = 0$. The results from Model 5 can be found in Table 15. We have the parameter estimate for the interaction between gender and duration was $\widehat{\beta_3} = -1.07$ (s.e. $= 0.2136$, $p < 0.0001$). Thus the interaction between ethnicity and duration was significantly different from zero. The parameter estimate for the interaction between ethnicity and duration squared was $\widehat{\beta_4} = -0.09$ (s.e. $= 0.0247$, $p = 0.0002$). Therefore the interaction between ethnicity and duration squared was significantly different from zero. Thus we rejected the $H_0 : \beta_3 = \beta_4 = 0$, therefore we rejected the proportionality odds assumption for the ethnicity.

The level-one proportionality odds assumption, for gender, was tested by looking at the interactions between gender and the two continuous variables duration and duration squared. That is, we tested the following null hypothesis $H_0 : \beta_3 = \beta_4 = 0$. The results from Model 6 can be found in Table 16. We had the parameter estimate for the interaction between gender and duration was $\widehat{\beta_3} = -0.98$, $p < 0.0001$. Thus the interaction between gender and duration was significantly different from zero. The parameter estimate for the interaction between gender and duration squared was $\widehat{\beta_4} = -0.03$ $p = 0.0258$. Therefore the interaction between gender and duration squared was significantly different from zero. Thus

Table 16: Parameter Estimates for Model 6.

Parameter	Estimate	Standard Error	Pr $> \|t\|$
α_{01}	-6.1995	0.1853	< 0.0001
α_2	-5.4831	0.1539	< 0.0001
α_3	-4.2833	0.1184	< 0.0001
α_4	-3.3174	0.0961	< 0.0001
α_5	-2.3117	0.0804	< 0.0001
γ_{00}	1.3094	0.1012	< 0.0001
γ_{10}	0.0221	0.0912	0.8079
γ_{20}	7.0446	0.4505	< 0.0001
γ_{30}	-1.0001	0.2115	< 0.0001
γ_{40}	-0.0544	0.0244	0.0258
γ_{01}	-0.8103	0.1057	< 0.0001
γ_{11}	0.0493	0.1257	0.6949

we rejected $H_0 : \beta_3 = \beta_4 = 0$, so we reject the proportionality odds assumption for gender. This imples the hazard functions of gender varies over the years.

Table 17 shows the results for the level-two proportional odds assumption. Table 17 has the results of the level two variable, school type. We looked at the interactions between school type and the continuous variables duration and duration squared. The hypothesis test for the level-two proportional odds assumption was similar to the hypothesis test for the level-one proportional odds assumption, $H_0 : \beta_3 = \beta_4 = 0$.

For the level-two proportional odds assumption of school-type we had the parameter estimate for the interaction between duration and school-type was $\widehat{\beta_3} = 1.00$ $p = 0.0175$. We had that $\widehat{\beta_3}$ was significantly different from zero. The parameter estimate for the interaction between school type and duration-squared was $\widehat{\beta_4} = -0.20$, $p < 0.0001$). Thus the interaction between duration-squared and school-type was significantly different from zero. Since $\widehat{\beta_3}$ and $\widehat{\beta_4}$ were significantly different from zero we rejected the null hypothesis that $H_0 : \beta_3 = \beta_4 = 0$. Thus the level-two proportional odds assumption for school type was not met. This implied that there was variation across school-type.

Model 8 was used to test the level-two proportional error assumption. The level-two proportional error assumption is satisfied if both error terms of the parameter estimates

Table 17: Parameter Estimates for Model 7.

Parameter	Estimate	Standard Error	$\Pr > \|t\|$
α_1	-4.9404	0.1310	< 0.0001
α_2	-4.9160	0.1322	< 0.0001
α_3	-4.1800	0.1132	< 0.0001
α_4	-3.4171	0.0991	< 0.0001
α_5	-2.5153	0.0875	< 0.0001
γ_{00}	1.6452	0.1072	< 0.0001
γ_{10}	-0.0369	0.0921	0.6890
γ_{20}	0.0433	0.0579	0.4546
γ_{30}	0.7057	0.2970	0.0175
γ_{40}	-0.2323	0.6316	< 0.0001
γ_{01}	2.1201	0.1235	0.0008
γ_{11}	0.1751	0.1257	0.1564

Table 18: Parameter Estimates for Model 8

Parameter	Estimate	Standard Error	$\Pr > \|t\|$
γ_{00}	2.0299	0.2626	< 0.0001
γ_{10}	0.0456	0.0766	0.5515
γ_{20}	0.0562	0.0495	0.2561
γ_{30}	-1.1230	0.1284	< 0.0001
γ_{40}	0.0715	0.0149	< 0.0001
γ_{01}	-0.2652	0.0883	0.0027
γ_{11}	0.0633	0.1055	0.5482

of the duration and the duration-squared variables are zero. Thus if either of the error terms for $\widehat{\beta_3}$ and $\widehat{\beta_4}$ are not zero we have the logit-hazard curves for school type are not parallel. This will cause us to reject the level-two proportional error assumption.

Table 18 has the results of the Model 8 data analysis. We had the standard error of $\widehat{\beta_3}$ to be 0.1284 and the standard error of $\widehat{\beta_4}$ to be 0.01491. Thus both standard errors were not zero and we rejected the level-two proportional error assumption.

Brief Summary of Results of the Level Two Discrete-Time Hazard Models

Similar to the level one models, we had the risk of not returning to a four-year university increase each year a student was enrolled. In Model 4, we had no significant difference between the hazard functions in male and female students and white and non-white students.

Models 5, 6, 7, and 8 were used to test the assumptions of the model. With Models 5 and 6 we found that the level-one proportional odds assumption was violated, and thus the effect of gender and ethnicity varied across the years enrolled at a four-year university. Model 7 showed the level-two proportional odds assumption for school type was not met. Model 8 showed that the level-two proportional error assumption was also not met. Based on our analysis we had the level-one proportional odds assumption, the level-two proportional odds assumption, and the level-two proportional error assumption had not been satisfied.

Comparison of the Models

The model fit statistics can be found in Table 19. The level one models were compared with the -2 log likelihood statistic and the likelihood ratio χ^2 statistic. For Model 1 we had -2 log likelihood statistic of 9,543.92 and a likelihood ratio χ^2 statistic of 12,115.54 ($df = 5$, $p < 0.0001$). Model 2 gave us a -2 log likelihood statistic of 9,456.49, which was an improvement over model 1, and likelihood ratio χ^2 of 12,036.61 ($df = 7$, $p < 0.0001$).

Four model fit statistics were provided for the level two models, but we will focus on the -2 log likelihood statistic. The model fit statistics for Model 3 were -2 log likelihood of 9,445.16. Model 4 had a -2 log likelihood statistic of 9,102.3 and Model 5 had a -2 log likelihood statistic of 336.3 Thus Model 4 and Model 5 were better fitting models than Model 3. Model 6 had a -2 log likelihood statistic of 6,972.4, which implies it was a better fitting model than Model 3. Model 7 had a -2 log likelihood of 7,633.2 which implied it is a better fitting model than Model 3. Model 8 had a -2 log likelihood of 8,145.2 which implied it was a better model than Model 3. Model 9 had a -2 log likelihood of 11,389, which implied it was not a better model than Model 3.

Summary of Major Results

In the level one models we found the hazard function increased for each year a student was enrolled at a university. The time-indicator variables were found to be significant in all level one models. In Model 2, a significant difference was found in the hazard functions between male and female students, and white and non-white students. In Model 4 we

Table 19: Comparisons of the Models

Level-One Models				
Model	Model Fit -2 Log L	Likelihood Ratio χ^2	DF	Pr $> \chi^2$
1	9543.92	12115.54	5	< 0.0001
2	9456.49	12036.61	7	< 0.0001
Level-Two Models				
Model	-2 Log L	AIC	AICC	BIC
3	9445.16	9461.16	9445.16	9426.16
4	9102.3	9122.3	9122.3	9198.8
5	3336.3	3360.3	3360.3	3452.1
6	6972.4	6996.4	6996.5	7088.2
7	7633.2	7657.2	7657.3	7749.0
8	11389	11403	11403	11457

found a significant difference in the hazard functions of the level two variable school type which implied students who attended a private four-year university had a greater risk of not returning to a university than students who attended a public four-year university. Model 5 and Model 6 showed the level-one proportional odds assumptions in Model 4 was not met, and Model 7 showed the level-two proportional odds assumption was not met in Model 4. Model 8 showed the two-level proportional error assumption was not met in Model 4.

5. Conclusion and Discussion

This chapter discusses the conclusions that resulted due to the data analysis done in this study. This chapter is divided into the following sections: (1) an overview of the purpose and importance of the study; (2) the discussion of the main findings and conclusions; and (3) implications for future research.

Overview

The retention rate of a university has a profound impact on student recruitment, public and private funding, public perception, and the surrounding community. Universities study the trends in student retention for several reasons including: to understand why students leave a university, to improve diversity in higher education, to understand the relationship between financial assistance and retention, and to find ways to improve retention. Universities study the factors that effect retention to better determine ways to help students persist (St. John et al., 2004). University officials are not the only individuals interested in student retention. State legislatures are holding university officials accountable for the retention rates of students at their universities (DesJardins et al., 1999). Public officials are looking at the retention rates of universities in order to determine the allocation of state funds (Bowen, 1980). Parents of college-bound students also have a vested interest in the retention rate of universities, because the increased cost of higher education has been shifted from universities to college students and their families (Hu & John, 2001).

This study attempted to use a multilevel discrete-time hazard model, to determine the effect the different nested levels of higher education had on the retention of students. This study also wanted to illustrate the use of statistical software to estimate a multilevel discrete-time hazard model.

The sample used in this study came from the National Longitudinal Survey of Youth, 1997 (NLSY97). The sample contained students who participated in Round 3 through Round 8, which correspond to the years 1999 through 2004. The data in Round 3 through Round 8 contained information about the time the participants spent as undergraduate students in college. Only students enrolled at a four-year university were considered for

the sample. The NLSY97 followed 8,984 individuals between the ages of 12 to 16 as of December 31, 1996. The sample consisted of N = 3,072 students who were enrolled at a four-year university beginning in the year 1999.

This study was divided into two components. The first component looked at level one discrete-time hazard models. The second component looked at a level two discrete-time hazard model and tested the assumptions of a level two discrete-time hazard model.

This study used a sophisticated analytical model to analyze the data of a large national data set. The analytical method used in this study allowed this researcher to investigate retention in higher education. In this way, a new model was introduced to understand the multilevel nature of retention data. A better understanding of the factors that effect student retention would allow university officials, policy makers, and parents to be better prepared to help students persist to the completion of their college education.

The next section discusses the main research objectives and findings of this study and the conclusions that were drawn from them.

Main Findings and Conclusions

A Multilevel Model Used to Analyze Retention Data

The first goal of this study was to describe and analyze retention data in higher education using a multilevel discrete-time hazard model. The level two variable school type was found to be significant. This result implied there was a difference in the risk between students who went to public and private universities. We determined that students who attended a private university had a greater risk of not returning to school each year. The time indicator variables were found to have a significant impact on the risk of students returning to a university the next year. It was found that every year a student was enrolled, the risk of not returning to the university increased each year. Ethnicity was found not to have a significant impact on the retention rate of students in the multilevel model. Results from the multilevel discrete-time hazard model showed there was no significant difference in hazard functions between male and female students. These results implied there was no difference in the risk of not returning to a university for male and female students. These

results also implied there was no difference in the risk of not returning to a university for white and non-white students each year of the study.

No other multilevel discrete-time hazard studies have been done involving the retention of students in higher education. One of the most important aspects of student retention is looking at the time when a student leaves the university.

The next major objective of this study was to explore the likelihood that a student left a university.

The Likelihood a Student Left a University

The second major research objective of this study was to explore the likelihood that a student left a university each year. This objective was used to determine the impact of timing, when a student left a university. Results from the individual level models showed that the longer a student was enrolled the greater the risk of the student not returning the following year. A similar result was seen in the multilevel model, the longer a student was enrolled the greater the risk of not returning the next year. Thinking about this pattern, we realize that the students who dropped out in their first or second year resulted in a greater negative impact on student retention; however, students that matriculated into the third, fourth, and fifth year of enrollment we saw an increase in students who graduated. The large number of censored observations in year five indicated that students were continuing to persist into their sixth year.

These findings were supported by Tinto (1988). Tinto determined students left a university at different times for different reasons. The reasons that students left during the first year, especially after the first-semester, were much different than the reason students leave later in their college career. The students who did not return early in their college career were unable to adapt, or become integrated into the university's environment. This was very similar to the findings of this study, and further research should be done with multilevel models to determine possible causes.

DesJardins et al. (2006) studied the different types of stages of student enrollment: enrolled, stop out, and the effect the length of time of a stop out had on the completion

of an undergraduate degree. They found the average time preceding a student's first stop out was three and a half years. Ishitani and DesJardins (2002 - 2003) found that students had the greatest risk of dropping out at the end of their first year and third year. The average survival time in this study occurred between the fifth and sixth years of enrollment.

The next major objective of this study deals with the individual level discrete-time hazard model.

Individual Level Factors that Impact Student Retention

The third major research objective of this study was to explore what individual level factors were related to a student leaving a university each year.

Results from the individual level discrete-time hazard model showed that the number of years a student was enrolled had a significant impact in determining the risk of returning or not returning to a university. The longer a student was enrolled the greater the risk of not returning to the university the next year.

The final individual level discrete-time hazard model showed that ethnicity and gender were found to have a significant impact on determining the risk of a student not returning to a university. This study found white students had a greater risk of not returning to a university the next year than non-white students.

Past research has shown different results when comparing white students to other ethnic groups. Murtaugh et al. (1999) found that ethnicity did have an effect on student retention. Several studies have found that African American students were more likely to dropout than white students (DesJardins et al., 1999; DesJardins, Ahlburg, & McCall, 2002b; DesJardins et al., 2006). When comparing white students to Asian American students, past research found that Asian American students were more likely to return to a university than white students (DesJardins et al., 1999; Ishitani & DesJardins, 2002 - 2003).

Past research has a conflicting view on the significance of gender. Studies by Murtaugh et al. (1999) and Ishitani and DesJardins (2002 - 2003) found that gender did not have a significant impact on the retention of students. While several other studies found that gender did make a difference in determining student retention. Hu and John (2001)

concluded that female students were more likely to persist than male students. DesJardins, Ahlburg, and McCall (2002b) found that male students were less likely than female students to graduate in four years. DesJardins, Ahlburg, and McCall (2002a) found that 36.5 percent of students graduated without stopping out, and those students were characterized as white females.

No studies of retention in higher education have used a multilevel discrete-time hazard model. The final major objective of this study deals with the school-level variable.

The Impact of School Type on Student Retention

The final major research objective of this study was to explore the extent to which the type of school a student attended effected the retention rate of students.

This research objective dealt with the school-level variable in this study. The type of school a student attended did have a significant impact on the odds of returning to a university the next year. We determined the odds of leaving the next year were smaller for students who attended a public university than students who attended a private university.

Summary of Major Findings

The following were the major findings of this study:

- The individual level models found that the hazard function increased for each year a student was enrolled at a university. The time-indicator variables were found to be significant in all individual level models.

- In the individual level model, we found a significant difference in the hazard functions between white and non-white students, and male and female students.

- In the level two model we found a significant difference between students who attended a public university and those who attended a private university.

- We found that the level two model did not meet the level-one proportional odds assumptions, the level-two proportional odds assumption, or the level-two proportional error assumption.

Implications for Research Practice

Analysis of NLSY97 Data

This study used a national database to study retention in higher education and the different factors that effected the retention of students. There were several advantages to using a national database. The first advantage was that the information on the students had been gathered at many levels, such as high school and college. The sample and the secondary sample gave a national representation of the population of interest. The creation of such databases made it possible for individuals to conduct a large study. The information available in the NLSY97 database, the methods used to collect the data, and the representative nature of the sample allowed this study to address new issues in student retention and important implications for research practice.

Longitudinal Data Analysis

Longitudinal data are collected from the same population over a prolonged period of time. Longitudinal data allow a researcher to follow patterns of change in the same population over a certain time frame (Creswell, 2002). Collecting longitudinal data allow a researcher to have a better understanding of a student's college career, a way to follow factors that effect student's decision to stay or leave a university, and to "increase statistical power" (Willett & Singer, 1991).

Longitudinal data analysis has several advantages. The first advantage was the economical cost in gathering information on the subjects. The next advantage was the ability to use the information gathered on subjects as the control for the subjects. Another advantage was the between-subject variation was omitted from the error. When the patterns and observations were the same, longitudinal designs provide better estimators than with cross-sectional designs. Longitudinal data analysis separated effects that changed over time from differences between individuals at the beginning of a study. One other advantage of longitudinal data analysis was that it provided information on the changes that occurred in each subject (Hedeker & Gibbons, 2006).

Hierarchical Linear Modeling

A hierarchical linear model is a model that consists of nested data, for example the productivity of workers may be influenced by workplace characteristics. In this example, data were gathered on the workplace and the workers with analysis done on both levels. There was a hierarchy to the data in this example the workers were nested within the workplace (Raudenbush & Bryk, 2002).

There were several advantages to using multilevel data analysis. First, it allowed a researcher to determine the amount of variability caused by each level of data. Second, a researcher was able to model the first level of data analysis in terms of effects at all levels. Third, by using a multilevel model a researcher was able to test the possible interactions between each level of data. Finally, the subjects within the data set were similar because they came from similar environments because of the subjects were not independent, and multilevel data analysis was able to handle the absence of independence in the subjects (Raudenbush & Bryk, 2002).

Hazard Modeling

"The population hazard function describes the risk of an event's occurrence in each time period, the probability that a randomly selected population member will experience the event in the period given that the event has not already occurred" (Willett & Singer, 1991, p. 954).

Discrete-time survival analysis has several advantages. First, discrete-time survival analysis was suited to analyze longitudinal data. Second, discrete-time survival analysis handled time-invariant and time-variant predictors. Third, violations of the model were easily tested and corrected. Finally, censored observations were handled with discrete-time survival analysis (Willett & Singer, 1991).

Multilevel Discrete-Time Hazard Model

In this study we used a multilevel discrete-time hazard model to analyze individual and school-level factors that effected the risk of a student leaving a university. The main purpose of this study was to put forth an example of how to use a multilevel discrete-time

hazard model to analyze retention data in higher education. We showed how this analysis was done with a common statistical software package, and demonstrated how to convert the data from a person-oriented data set to a person-period data set.

By using a multilevel discrete-time hazard model, we were able to use the advantages of longitudinal data analysis, hierarchical linear modeling, and hazard modeling in one model. We were able to look at the change in the data over time, while taking into account the nested nature of the data. We also examined the interactions and variability between and within each level of the data.

Multilevel discrete-time hazard analysis is a powerful analytical tool that will be instrumental in future studies of retention.

Limitations and Next Steps

One of the limitation of this study was the missing observations in the data set. The amount of missing data caused the loss of independent variables that would have been important in helping to determine why students did not return to a university. Not using independent variables such as ACT score, high school rank, college GPA, and financial aid variables limited the variables that were used to explain why students were not returning.

A next step would be to use a larger national data set with fewer missing observations or to use a data set from a single university where necessary data is already collected.

This study measured each period as one year in duration, but most universities separate a year into two semesters. For this study, it was not possible to determine in which semester a student did not return to the university. This caused the study not have a true estimate of how long a student was actually enrolled.

A next step would be to use a data set that had the enrollment periods separated by semesters, instead of years. This would allow for an exact estimation of the timing of when students are leaving a university. This would help university official to be better prepared to help student to persist.

Implication for Future Research on Retention

Different School Level Variables

Understanding the different factors that effect a student's ability to persist at a college or to become integrated into the college environment are important in helping a student complete their undergraduate studies. All of the studies in retention that used a discrete-time hazard model only analyzed the individual level variables. No research, except this study, used a multilevel discrete-time hazard model, because of this future work involving retention data can include more specific models that look at the percentages of each ethnic group at a university and how these percentages effect retention. Another future study could look at the ethnic makeup of each major or each college and how the different percentages of each ethnic group in a major effect student retention.

Expand to a Three-Level Model

In addition to the looking at different school-level variables, future research could expand the level two model to a level three model. This future research could continue studying the nested nature of student retention data. Research could be done by looking at the individual, within a major, within a college or within a university. This would give a better understanding of how the different levels of a university effect a student.

Summary

This dissertation described a study of student retention in higher education and the individual-level and school-level variables that effect a student's risk of returning to a university. This study was directed by four objectives that were used to describe a new method of explaining student retention in higher education. These objectives looked at a multilevel discrete-time hazard model, an individual level hazard model, and factors such as ethnicity, gender, and school type to explain why student were not returning to a university the next year.

The major findings in this study showed (1) that the longer a student was enrolled the risk of not returning the next year increased; (2) looking at the individual level discrete hazard model we found a significant difference in the hazard functions between white

and non white students and male and female students (3) in the multilevel discrete-time hazard model we found a significant difference in the hazard functions between students who attended public and private universities; (4) we determined that the level-one proportional odds assumption, the level-two proportional odds assumption, and the level-two proportional error assumption were not met.

References

Adelman, C. (1999). *Answers to the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment.* Jessup, MD: U.S. Department of Education.

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

Allison, P.D. (1995). *Survival analysis using sas a practical guide.* Cary, NC: SAS Press.

Barber, J.S., Murphy, S.A., Axinn, W. G., & Maples, J. (2000). Discrete-time multilevel hazard analysis. *Sociological Methodology, 30,* 201-235.

Barefoot, B. (2004). Higher education's revolving door: confronting the problem of student drop out in us colleges and universities. *Open Learning, 19*(1), 9-18.

Bayer, A. (1968). The college drop-out: factors affecting senior college completion. *Sociology of Education, 41*(3), 306-316.

Bean, J. (1980). Dropouts and turnover: the syntheses and test of a causal model of student attrition. *Research in Higher Education, 12*(2), 155-187.

Bean, J. (1982). Student attrition, intentions, and confidence: interaction effects in a path model. *Journal of Higher Education, 17*(4), 291-320.

Bean, j. (1983). The application of model of turnover in work organizations to the student attrition model. *Review of Higher Education, 2*(6), 129-148.

BLS. (2003). *National longitudinal survey of youth, 1997.* Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Bowen. H. R. (1980). *The costs of higher education.* San Francisco, CA: Jossey - Bass.

Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education, 63*(2), 143-164.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

Casella, G. (2002). *Statistical inference.* Pacific Grove, CA: Duxbury.

Creswell, J. (2002). *Educational research planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, New Jersey: Merrill.

Daempfle, P. (2003-2004). An analysis of the high attrition rates among first year college science, math, and engineering majors. *Journal of Student Retention: Research, Theory, and Practice, 5*(1), 37-52.

DesJardins, S., Ahlburg, D., & McCall, B. (1999). An event history model of student departure. *Economics of Education Review, 18,* 375-390.

DesJardins, S., Ahlburg, D., & McCall, B. (2002a). Simulating the longitudinal effects of changes in financial aid on student departure from college. *The Journal of Human Resources, 37*(3), 653-679.

DesJardins, S., Ahlburg, D., & McCall, B. (2002b). A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education, 73*(5), 555-581.

DesJardins, S., Ahlburg, D., & McCall, B. (2006). The effects of interrupted enrollment on graduation from college: Racial, income, and ability differences. *Economics of Education Review 25*(6), 575-590.

DesJardins, S., Kim, D., & Rzonca, C. (2002-2003). A nested analysis of factors affecting bachelor's degree completion: *The Journal of College Student Retention, Research, Theory and Practice, 4*(4), 407-435.

DesJarding, S., McCall, B., Ahlburg, D., & Moye, M. J. (2002) Adding a timing light to the "tool box". *Research in Higher Education, 43*(1), 83-114.

Durkheim, E. (1961). *Suicide.* Glencoe: The Free Press.

Elkins, S., Braxton, J. M., & James, G. (2000). Tinto's separation stage and its influence on first semester college student persistence. *Research in Higher Education, 41*(2), 251-268.

Gansemer-Topf, A., & Schuh, J. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education, 47*(6), 613-642.

Glynn, J. Sauer, P., & Miller, t. (2005-2006). Configural invariance of a model of student attrition. *Journal of College Student Retention, 7*(3-4), 263-281.

Goldstein, H. (2003). *Multilevel statistical models.* New York: Oxford Press.

Hedeker, D. & Gibbons, R. D. (2006). *Longitudinal data analysis.* Hoboken, New Jersey: John Wiley and Sons.

Hill, T., & Lewicki, P. (2005). *Statistics methods and applications.* Tulsa, OK: StatSoft, Inc.

Hu, S., & John, E. S. (2001). Student persistence in a public higher education system: understanding racial and ethnic difference. *The Journal of Higher Education, 72* (3), 265-286.

Ishitani, T., & DesJardins, S. (2002-2003). A longitudinal investigation of dropout from college in the united states. *Journal of College Student Retentions, 4*(2), 173-201.

Ishitani, T., & Snident, K. G. (2006). Longitudinal effects of college preparation programs on college retention. *IR Applications, 9,* 1-10.

Lohr, S. L. (1999). *Sampling: Design and analysis.* New York: Duxbury Press.

Ma, X., & Williams, J. D. (1999). Dropping out of advanced mathematics: How much do students and schools contribute to the problem? *Edcucational Evaluation and Policy Analysis, 21*(4), 365-383.

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data.* New York: John Wiley and Sons.

Munro, B. (1981). Dropouts from higher education: path analysis of a national sample. *American Educational Research Journal. 18*(2), 133-141.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(1), 355-371.

Noble, K., Flynn. N., Lee, J., & Hilton, D. (2007-2008). Predicting successful college experience: evidence from a first year retention program. *Journal of College Student Retention, 9*(1), 39-60.

Pascarella, E., & Terenzini. P. (1977). Patterns of student-faculty informal interaction beyond the classroom and voluntary freshman attrition. *The Journal of Higher Education, 48*(5), 540-552.

Price, J. (1977). *The study of turnover.* Ames, Iowa: Iowa State University Press.

Price, J., & Mueller, C. (1981) A causal model of turnover for nurses. *Academy of Management Journal, 24,* 543-565.

Raudenbush, S. F., Brennan, R. T., & Buka, S. L. (2002). Estimating multi-level discrete-time hazard models using cross-sectional data: Neighborhood effects on the onset of adolescent cigarette use. *Multivariate Behavioral Research, 37*(3), 297-330.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18*(2), 155-195.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*(1), 64-85.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange, 2*(3), 38-62.

St. John, E., Shouping, S., Simmons, A., Carter, D., & Weber, J. (2004). What difference does a major make? the influence of college major field on persistence by african american and white students. *Research in Higher Education, 45*(3), 209-232.

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

Tinto, V. (1982). Limits of theory and practice in student attrition. *Journal of Higher Education, 53*(6), 687-700.

Tinto, V. (1988). Stages of student departure: reflection on the longitudinal characteristics of students leaving. *Journal of Higher Education, 59*(4), 433-455.

Tinto, V. (1990). Principles for effective retention. *Journal of Freshmen Year Experience, 2*(1), 35-48.

Van Gennep, A. (1960). *The rites of passage.* Chicago: The University of Chicago Press.

Willet, J. & Singer, J. D. (1991). From where to when: new methods for studying student dropout and teacher attrition. *Review of Educational Research, 61*(4), 407-450.

Willet, J. & Singer, J. D. (1993). Investigation onset, cessation, relapse, and recovery: Why you should, and how you can use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psycology, 61*(6), 952-965.

Willet, J. & Singer, J. D. (1995). It's deja vu allover again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavior Statistics, 30*(1), 41-67.

Appendix: Relevant SAS Code

```
dm'log;clear;output;clear';
PROC FREQ DATA=COMP.COLLCOMP012000COHORT;
TABLES DURATION*CENSOR/NOPERCENT NOROW NOCOL;
RUN;
DATA ONE;
INPUT PERIOD;
DATALINES;
0
;
RUN;
PROC FREQ DATA=COMP.FINALSET2000;
TABLES PERIOD*Y/NOPERCENT NOCOL OUT=SUMMARY OUTPCT;
RUN;
DATA TWO;
SET ONE SUMMARY;
IF Y=1 OR PERIOD = 0;
HAZARD=PCTROW/100;
RETAIN SURVIVOR 1;
IF PERIOD ¿ 0 THEN SURVIVOR=SURVIVOR*(1 - HAZARD);
KEEP PERIOD HAZARD SURVIVOR;
RUN;
PROC PRINT DATA = TWO NOOBS;
RUN;
GOPTIONS RESET=ALL;
SYMBOL COLOR=BLACK I=JOIN VALUE=NONE HEIGHT=2;
AXIS1 LABEL=NONE ORDER=(0 TO .15 BY .05) MINOR=NONE;
AXIS2 LABEL = ('YEARS ENROLLED IN SCHOOL') ORDER=(0 TO 5 BY 1) MI-
NOR=NONE;
AXIS3 LABEL=NONE ORDER=(0 TO 1 BY .25) MINOR=NONE;
PROC GPLOT DATA = TWO UNIFORM;
TITLE2 'HAZARD PROBABILITY';
PLOT HAZARD*PERIOD/VAXIS=AXIS1 HAXIS=AXIS2 NOFRAME; *NOLEGEND;
TITLE2 'SURVIVAL PROBABILITY';
PLOT SURVIVOR*PERIOD/ VAXIS=AXIS3 HAXIS=AXIS2 NOFRAME HREF=5.6 VREF=.5
LHREF=21 LVREF=21;
RUN;
QUIT;

dm'log;clear;output;clear';
/*FITTING THE BASELINE MODEL WITH TIME INDICATORS ONLY*/
PROC LOGISTIC DATA=COMP.FINALSET2000 DESCENDING
NOSIMPLE OUT=COMP.ESTMODEL01;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
```

```
EDUCATION RETENTION DATA';
TITLE2 'MODEL 1 INCLUDES ONLY THE DUMMY VARIABLES FOR ENROLLMENT
AND NO INTERCEPT.';
CLASS Y;
MODEL Y = D01 D02 D03 D04 D05 / NOINT;
RUN;

/*FITTING A MODEL WITH ONLY TIME INVARIANT COVARIABLES AND NO IN-
TERCEPT*/
PROC LOGISTIC DATA=COMP.FINALSET2000 DESCENDING
NOSIMPLE OUT=COMP.ESTMODEL02;
TITLE2 'MODEL 2 INCLUDES THE DUMMY VARIABLES FOR ENROLLMENT AND
DEMOGRAPHIC
COVARIABLES WITH NO INTERCEPT';
CLASS Y;
MODEL Y = D01 D02 D03 D04 D05 GENDER ETHNICITY / NOINT;
RUN;

/*FITTING A MODEL WITH TIME INVARIANT AND TIME VARING COVARIABLES
AND WITH AN INTERCEPT*/
PROC LOGISTIC DATA=COMP.FINALSET2000 DESCENDING
NOSIMPLE OUT=COMP.ESTMODEL04;
TITLE2 'MODEL 3 INCLUDES THE DUMMY VARIABLES FOR ENROLLMENT, THE
TIME INVARIANT
COVARIABLES AND THE TIME VARYING COVARIABLES';

CLASS Y;

MODEL Y = D01 D02 D03 D04 D05 GENDER ETHNICITY;
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
TITLE2 'MODEL 4 THE SIMPLEST TWO-LEVEL DISCRETE-TIME HAZARD MODEL.';
PARMS ALPHA01 = -3.3868 ALPHA02 = -3.3903 ALPHA03 = -2.8482 ALPHA04 =
-2.3729 ALPHA05 = -1.8219
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0931 /*BETA03 = 0.5 BETA04 = 0.5*/
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = (ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+(BETA00 + BETA01*ETHNICITY + BETA02*GENDER + /*BETA03*DURATION +
BETA04*DURSQUARE*/
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY) /*+ U*/;
EXPETA = EXP(ETA);
```

```
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
TITLE2 'MODEL 4A THE SIMPLEST TWO-LEVEL DISCRETE-TIME HAZARD MODEL.';
PARMS ALPHA01 = -3.3868 ALPHA02 = -3.3903 ALPHA03 = -2.8482 ALPHA04 = -
2.3729 ALPHA05 = -1.8219
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0931 BETA03 = 0.5 BETA04 = 0.5
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = (ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+(BETA00 + BETA01*ETHNICITY + BETA02*GENDER + BETA03*DURATION +
BETA04*DURSQUARE
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY)/* + U*/;
EXPETA = EXP(ETA);
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
TITLE2 'MODEL 5 THE TEST OF LEVEL-ONE ETHNICITY ';
PARMS ALPHA01 = -3.3868 ALPHA02 = -3.3903 ALPHA03 = -2.8482 ALPHA04 = -
2.3729 ALPHA05 = -1.8219
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0931 BETA03 = 0.5 BETA04 = 0.5
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = (ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+(BETA00 + BETA01*ETHNICITY + BETA02*GENDER
+ BETA03*DURATION*ETHNICITY + BETA04*DURSQUARE*ETHNICITY
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY) /*+ U*/;
EXPETA = EXP(ETA);
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
```

```
TITLE2 'MODEL 6 THE TEST OF LEVEL-ONE GENDER';
PARMS ALPHA01 = -3.3868 ALPHA02 = -3.3903 ALPHA03 = -2.8482 ALPHA04 = -
2.3729 ALPHA05 = -1.8219
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0931 BETA03 = 0.5 BETA04 = 0.5
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = (ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+(BETA00 + BETA01*ETHNICITY + BETA02*GENDER + BETA03*DURATION*GENDER
+ BETA04*DURSQUARE*GENDER
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY) /*+ U*/;
EXPETA = EXP(ETA);
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
TITLE2 'MODEL 7 THE TEST OF LEVEL-TWO PROPORTIONAL ASSUMPTION ';
PARMS ALPHA01 = -3.3868 ALPHA02 = -3.3903 ALPHA03 = -2.8482 ALPHA04 = -
2.3729 ALPHA05 = -1.8219
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0931 BETA03 = 0.5 BETA04 = 0.5
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = (ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+(BETA00 + BETA01*ETHNICITY + BETA02*GENDER
+ BETA03*DURATION*SCHOOLTYPE + BETA04*DURSQUARE*SCHOOLTYPE
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY) /*+ U*/;
EXPETA = EXP(ETA);
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;

PROC NLMIXED DATA=COMP.FINALSET2000 QMAX = 75;
TITLE1 'CHRISTOPHER W. GUILLORY DISCRETE-TIME MULTILEVEL HAZARD
ANALYSIS OF HIGHER
EDUCATION RETENTION DATA';
TITLE2 'MODEL 8 CONTINUOUS TWO-LEVEL DISCRETE-TIME HAZARD MODEL.';
PARMS /*ALPHA01 = 2.4616 ALPHA02 = 1.4694 ALPHA03 = 1.6836 ALPHA04 = 0.8674
ALPHA05=3.8731*/
BETA00 = 0.2320 BETA01 = 0.1396 BETA02 = 0.0921 BETA03 = 0.5 BETA04 = 0.5
BETAPR05 = 0.5 BETAPR06 = 0.5 /*ETABAR = 1.88*/;
ETA = /*(ALPHA01*D01 + ALPHA02*D02 + ALPHA03*D03 + ALPHA04*D04 + AL-
PHA05*D05)
+*/(BETA00 + BETA01*ETHNICITY + BETA02*GENDER + BETA03*DURATION +
```

```
BETA04*DURSQUARE
+ BETAPR05*SCHOOLTYPE + BETAPR06*SCHOOLTYPE*ETHNICITY)/* + U*/;
EXPETA = EXP(ETA);
P = EXPETA/(1 + EXPETA);
MODEL Y   BINARY(P);
RUN;
```

Vita

Christopher W. Guillory obtained Bachelor of Science degree from the University of Southwestern Louisiana at Lafayette. He then attended Louisiana State University at Baton Rouge, where he received a Master of Science degree in mathematics in 2000, followed by a Master of Applied Statistics degree in 2003. He will be awarded a Doctor of Philosophy degree in educational research at the spring commencement in 2008.

While at Louisiana State University, Christopher was a Huel D. Perkins fellow and an IGERT fellow. Christopher was also appointed as a graduate assistant. His work involved teaching statistical labs to graduate students as well as assisting several graduate students with research projects. From 2005 to 2008, Christopher has worked as a faculty member at Baton Rouge Community College.

Christopher is a member of the American Statistical Association, the Southwest Educational Research Association, the Mid-South Research Association, and the Louisiana Association of Teachers of Mathematics. He plans to continue to remain active in educational research. As a first step, he will work on several projects that emerged from his work in multilevel discrete-time hazard models.