5-2014

Categorical Data Analysis of First-Year Programs and its Effect on Retention of LSU First-Year Students

Morgan Matchett

Categorical Data Analysis of First-Year Programs and its Effect on Retention of LSU First-Year Students

by

MORGAN MATCHETT

Undergraduate Honors Thesis under the direction of

Professor George Cochran

Department of Mathematics

Submitted to the LSU Honors College

In partial fulfillment of the Upper Division Honors Program

May 2014

Louisiana State University

& Agricultural and Mechanical College

Baton Rouge, Louisiana

May 2, 2014

TABLE OF CONTENTS

ABSTRACT

Background: Higher Education institutions are looking for ways to increase retention of students by implementing programs that target first-year students and give them support. The significance of Louisiana State University's first-year program S.T.R.I.P.E.S. on retention in first-year students from fall 2011 is studied while controlling for race, gender, and grade point average.

Methods: SAS is used to compute the Chi-Square Test, the Breslow-Day test, homogeneity, odds ratio, and relative risk. These tests were used to calculate the significance of S.T.R.I.P.E.S. on retention when controlling for gender, race, and grade point average. Each control study (gender, race, and grade point average) was analyzed to determine the impact.

Results: This research showed that when controlling for gender, race, and grade point average, the effect of S.T.R.I.P.E.S. on retention is moderately significant.

Conclusions: Future work will need to be done to check error in the analysis. Areas to study and improve in the future include choosing a larger sample population of students and controlling for different or more variables.

1. INTRODUCTION

Higher Education is a rapidly changing field, fueled by the diverse students that make up each student body on a university campus. The major goal of attending an institution for higher education is to commit to higher learning and submit to achieving a higher degree. Oftentimes, a student's chances of succeeding in college are impacted by outside circumstances, such as socioeconomic and ethnic background. According to the system of American College Testing (ACT, 2013), approximately 22 percent of first-year students do not return for their second year at four-year public universities in the United States.

Universities across the United States are very interested in creating a better system of supporting students so that this retention rate is higher, which would also increase the graduation rate at these institutions as well. The issue of retention is important to university officials at Louisiana State University (LSU), because the success of students in their first year helps to predict the graduation success for students throughout their time at the institution. According to an article published about LSU's graduation rate, which recently increased to 69.1 percent, "LSU had its most successful period ever in enrolling students in 2007 and shepherding them through to graduation by the end of 2013" (Addo). This is a trend that LSU officials wish to continue to increase by growing the retention rate at the university.

Many universities implement measures to ensure that first-year students stay in school and return the following year. At LSU, university officials have programs run by the office of First Year Experience that enhance first-year student engagement before classes begin each fall. These programs are typically for first-year students only and

work to help them adjust to finding their way around an institution's campus, activities, and academics. Their main function is to create an environment for student success.

A program in existence at LSU called S.T.R.I.P.E.S., or Student Tigers Rallying, Interacting, and Promoting Education and Service, is a four-day program that classifies as one of these first-year programs. The S.T.R.I.P.E.S. program occurs during the summer before the start of a first-year student's fall semester. Three sessions are offered each summer for 300 students at each session. S.T.R.I.P.E.S. aims to assist in transitioning high school students into college life at LSU (LSU STRIPES). The number of students who may participate in S.T.R.I.P.E.S. is limited by both cost and available space. The cost of the program is between $290 and $330, fluctuating depending on how many nights a student may need to stay in the provided lodging on LSU's campus (LSU STRIPES Handbook 3). Not all students may be able to afford the cost of attending S.T.R.I.P.E.S., but LSU is able to provide fee waivers to students that qualify according to LSU's Federal Student Aid (LSU STRIPES Handbook 4). Programs such as S.T.R.I.P.E.S. intend to increase a first-year student's feeling of support and knowledge of campus resources. Through this extra support, the hope is to diminish students' fears of isolation by giving them a social setting where students can thrive.

Other factors in a student's life may affect whether they return to school after the first year, such as involvement in sorority or fraternity life and grade point average in the first semester. All of these parts of a student's life have been shown to make a difference in their success in college and likelihood to graduate. Past studies have shown that higher retention is linked to more social interaction within the first few weeks of college (Woosley). Programs like S.T.R.I.P.E.S. work to create this social environment for

students to transition and create a sense of community and belonging on a university campus.

The question is: do these first-year programs and factors really correlate or affect retention? The interaction of many different variables can impact a student's decision to complete a degree, including race, socioeconomic background and finances, and sense of belonging. The use of mathematical statistics to study this problem allows mathematical theory to be applied in a way that is efficient and valid. In particular, categorical data analysis allows for the most efficient way of organizing and analyzing this data. The previously mentioned variables that could possibly affect a student's return to school, such as finances, gender, race, grade point average, and involvement, all can be identified as categories of nominal scale. The study of categorical data analysis allows for these different variables to be analyzed so that the occurrence of certain variables can be determined as dependent or independent on other variables in the data set. Utilizing the contingency tables that make up categorical data, the research can check this independence by using the statistical variance for categorical data (Light and Margolin 534).

2. SUMMARY & ANALYSIS OF DATA

2.1. Methods and Procedures

LSU officials look at the efficiency of their first-year programs, particularly in the Office of Budget and Planning, to see if they make an impact on graduation and retention rates amongst the student population. The Office of Budget and Planning at LSU collects data on all incoming students and follows them through their time at the university. All of these descriptors are tied to an LSU student's personal identification number and tracked

through the LSU system to help with research and assessment projects facilitated by the university.

For this project, data was collected on a subset of first-year students that entered LSU in the fall of 2011. Focusing on the students that attended S.T.R.I.P.E.S. in the summer prior to fall of 2011, the data set reflected 514 randomly selected students, whose identities were stripped from their personal identification numbers to respect the privacy of the students. This data set was given for this research by the Office of Budget and Planning, and was produced as a random set of data per request. The Office of Budget and Planning determined the specific number of students selected from the pool of first-year students. The data also includes a matching data set from the following spring in 2012 to see if the students returned to LSU and whether or not their grade point average changed. The data includes information on gender, race, high school grade point average, curriculum college at LSU, living conditions (whether on-campus or off-campus), Greek fraternity or sorority involvement, S.T.R.I.P.E.S. involvement, fall 2011 grade point average, spring 2012 grade point average, and a value for retention. The retention value measures whether or not a student returned in spring 2012, or the semester after their first year, by yielding a “yes” or “no” response.

After the data was accumulated, general data management took place to prepare it for SAS, which is a statistical software system. This process consisted of giving categorical values to each piece of data. Any missing data was assumed to be completely at random. SAS eliminated any of the data entries that were missing values. For example, if an entry was missing the gender or other variable for a student, this entry was taken out

of the data set by SAS. SAS processes the data and checks if S.T.R.I.P.E.S. has a significant impact on retention.

Categorical data analysis allows variables that are described based on a series of categories to be compared and analyzed to see which variables are most responsible for causing a response (Agresti 2). The main models that are of interest in this research are analysis methods for two-way and three-way contingency tables. A contingency table is a way of organizing data in which all-possible outcomes for each combination of variables is shown (Agresti 16). SAS used the information from the Office of Budget and Planning to create tables of the data, which made it easier to organize. In this research, the focus of the project is on whether S.T.R.I.P.E.S. has an impact on retention when controlling for the three specific variables of gender, race, and grade point average. The motivation behind choosing control variables is to sharpen the analysis of the effect of S.T.R.I.P.E.S. on retention. Gender, race, and grade point average were chosen as the control variables because they are values that describe each student.

2.1.1. Summary of Data

It is useful to examine the data in terms of how many total students are involved in the study that achieved retention and did not achieve retention before processing the data in SAS. It can be observed whether there appears to be a higher likelihood of retention before processing the SAS tests to determine statistical significance. This information is broken down by participation in S.T.R.I.P.E.S. and by gender to understand if there is a higher probability odds of retention when students are involved with the S.T.R.I.P.E.S. program. The purpose of processing statistical analysis of this

data further is to see if these percentages could be recreated on other sets of data and if S.T.R.I.P.E.S. is a good predictor of retention.

Table 1. Total Retained vs. Not Retained Students

Total Sample of Students (*n* = 514)		
Number of Students Retained	423	82%
Number of Students Not Retained	91	18%
Total Students Enrolled in S.T.R.I.P.E.S. (*n* = 57)		
Number of Students Retained	52	91%
Number of Students Not Retained	5	9%
Total Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 457)		
Number of Students Retained	371	81%
Number of Students Not Retained	86	19%

Table 1 shows that in the total sample of students, more students achieve retention from fall 2011 to spring 2012 than students that do not achieve retention. It also shows that there is a higher retention rate for students enrolled in S.T.R.I.P.E.S. than students that are not enrolled.

In Appendix A, further summary of the data can be seen. In Tables 2 and 3, it is evident that a higher number of female students were retained than male students (Appendix A). It is also interesting to note that the number of female students that chose to enroll in S.T.R.I.P.E.S. was more than double the amount of male students enrolled in the set of data. As the purpose of S.T.R.I.P.E.S. is outlined as building a sense of community on a college campus, this trend could indicate a higher desire in women to build community than men. A similar trend can be seen in Tables 4 and 5, which show the summary of White and non-White students in the data (Appendix A). A higher percentage of retention is seen in students that identify as White than non-White. It is

also seen in these same Tables 4 and 5 that almost twice as many White students enrolled in S.T.R.I.P.E.S. than non-White students. It is evident that in this particular set of data, there appears to be a higher number of students being retained when enrolled in S.T.R.I.P.E.S., as seen in Table 1.

2.1.2. Simple Linear Regression

To assess whether the value of S.T.R.I.P.E.S. participation serves as a reliable predictor of retention, SAS computed a simple linear regression model. S.T.R.I.P.E.S. was analyzed to see if its presence in the data is strong enough to predict and affect retention. SAS processes a linear regression model procedure that examines a regressor variable and uses that variable to predict a response.

This linear regression model was utilized to fit the data (where the number of students = 514, as seen in Table 1) and produced the data seen in Table 8 (Appendix B). Table 9 (Appendix B) shows the values for the model and the error computed in the prediction of retention by S.T.R.I.P.E.S., as well as the number of degrees of freedom. The F statistic given in Table 8 (Appendix B) for the model is not significant (F=3.50 and $p = 0.0620$), which means S.T.R.I.P.E.S. is not an overly strong predictor of retention. However, the probability value of 0.0620 in Table 9 shows a moderately significant result of S.T.R.I.P.E.S. predicting retention since the probability value is between 0.05 and 0.10. The Root Mean Squared Error is calculated as 0.38086 for the model, which approximates the standard deviation of the error, or how closely the values fall within the average error (SAS Institute Inc. 2878). Table 9 and Table 10 demonstrate the values received from SAS for the parameter estimates (Appendix B). To determine the fit of the model, the values of R-Square and Adjusted R-Square illustrate how well the prediction

model fits the data. Since both R-Square and Adjusted R-Square are not very close to the value 1 (R-Square = 0.0068 and Adjusted R-Square = 0.0048), this indicates that there is not a strong linear relationship between S.T.R.I.P.E.S. and retention. The correlation value of -0.3327 in Table 10 provided an expected weak result (Appendix B). These results are expected since the retention value is measured as a 0:1 response.

However, since the R-Square and Adjusted R-Square values do not show a strong linear trend for the model, this signifies that the regressor is a weak predictor of retention. Through this evaluation, it is shown that S.T.R.I.P.E.S. alone cannot solely predict retention. Instead, more variables may be needed in conjunction with S.T.R.I.P.E.S. to properly assess its effect on retention. This is why categorical data analysis was used for this research to compare the impact of multiple factors on retention.

To assess the relationship of S.T.R.I.P.E.S. and retention without controlling for any variables, SAS computed a Chi-Square Test (Table 11, Appendix C). The Chi-Square Test processes the effect of S.T.R.I.P.E.S. participation on retention, resulting in a probability value of 0.061. This probability value indicates there is a moderately dependent relationship between S.T.R.I.P.E.S. and retention. The odds ratio for the study of S.T.R.I.P.E.S. versus retention means that the odds of a student returning to LSU when enrolled in the S.T.R.I.P.E.S. program is 2.41 times higher than the odds of a student returning to LSU when not enrolled in the program (Table 12, Appendix C). Summing over all control variables created the analysis in Tables 11 and 12. In the following sections, the data will be further analyzed by determining the significance of the relationship between S.T.R.I.P.E.S. and retention when controlling for different control variables.

2.2. Results and Analyses

In order to analyze the relationship of S.T.R.I.P.E.S. on retention, several tests were conducted for each control variable of gender, race, and grade point average. The Chi-Square Test, the Cochran-Mantel-Haenszel procedure, and tests for homogeneity were all used for the three control variables. SAS used the Chi-Square Test on each control variable to decide whether differences in the data were significant, as well as determine which variables are most significant (McHugh 143). The Chi-Square Test also calculates what values would be expected if there was no effect from the values of gender, race, and grade point average. By analyzing the data in this way, it provides a good measure of whether certain variables or factors have a significant impact on a response (in this case, retention).

The Cochran-Mantel-Haenszel procedure is commonly used in cases of $2x2xK$ tables, as is the case with the three analysis cases in this research (Yanagawa 58). This test allows the analysis of two variables while controlling for a third variable (Agresti 61). In the case of the data from the Office of Budget and Planning, the association between S.T.R.I.P.E.S. and retention can be analyzed while controlling for one of the three variables (race, gender, or grade point average).

SAS also produces a logit model that gives values for the odds ratio, relative risk, and homogeneity. A logit model is another linear function that analyzes the odds of a variable happening via the logit link (Agresti 73). The odds ratio is valuable to an analysis model because it gives an approximation for the interaction between two variables and allows for the study of other variables' relationship on that interaction as well (Bland 1468). Relative risk is another analysis that SAS processes for the data. One

definition of relative risk is “the ratio of the risk of disease occurrence in a population exposed to a given factor to that in an unexposed population” (Beral 159). In the case of the data in this research, the relative risk does not represent disease. Instead, for each separate analysis it computes the ratio of the probabilities of returning students enrolled in S.T.R.I.P.E.S. to that in a group who return but are not enrolled in S.T.R.I.P.E.S., all while controlling for the factors of race, gender, and grade point average. Finally, the test for homogeneity verifies whether the odds ratio is computed the same at different levels of the data (Agresti 63).

2.2.1. Gender, S.T.R.I.P.E.S., and Retention

Tables 13 and 14 compute the significance of the relationship between S.T.R.I.P.E.S. and retention while controlling for gender (Appendix D). Tables 13 and 14 also show the number of female and male students enrolled in the S.T.R.I.P.E.S. program in fall 2011 who returned to LSU in the following spring of 2012 (Appendix D).

The Chi-Square Test processes the effect of S.T.R.I.P.E.S. participation on retention from fall 2011 to spring 2012 (Appendix D). Table 13 shows the results of the Chi-Square Test when controlling for female gender, which computes a probability value of 0.0488. This probability value paired with the results of Table 13 represents the likelihood that these results could be replicated under no association, as well as give information regarding the null hypothesis (Agresti 31). Since the probability value when controlling for females is less than 0.05, it means that the result is significant when controlling for female gender.

In Table 14 (Appendix D), the Chi-Square Test analyzes the effect of S.T.R.I.P.E.S. participation on retention from fall 2011 to spring 2012. The probability

value in Table 14 when controlling for male gender is 0.7825, which is substantially larger than the comparison probability value of 0.05. This result may be caused by the low Chi-Square count in Table 14 (Appendix D). Since the Chi-Square Test fails when controlling for male students, as the probability value is greater than 0.05, S.T.R.I.P.E.S. does not have a significant impact on retention when controlling for male gender.

The Cochran-Mantel-Haenszel Statistics Model was used to check the null hypothesis of conditional independence between S.T.R.I.P.E.S. and retention controlling for gender. This test measures how strong an association is between certain categorical variables (Yu and Gastwirth 2377). After computing the Cochran-Mantel-Haenszel test, the probability value is 0.0833, which is moderately significant. The probability value of 0.0833 suggests a moderately significant association between retention and S.T.R.I.P.E.S., implying increased odds of retention for S.T.R.I.P.E.S. participants. This is true for either gender, both male and female.

According to Merriam-Webster's dictionary, homogeneity is defined as "the state of having identical cumulative distribution values" ("Homogeneity"). The Breslow-Day test analyzes homogeneity. This determines if the odds ratio between two variables is the same at each level of the control variable (Agresti 63). Although the association between S.T.R.I.P.E.S. and retention is moderately significant when controlling for gender, the probability value of the Breslow-Day Test ($p=0.2159$) denotes that the association does not vary across gender. Therefore, the retention rate for students in fall 2011 whose success was determined by S.T.R.I.P.E.S. participation showed no variation across gender.

Another helpful statistic for determining association between categorical variables includes the determining of the odds ratio and relative risk. These values are shown in Table 15 (Appendix D), where the estimates for the common relative risk were determined. Relative risk determines the ratio between the probability of students returning to LSU when enrolled in S.T.R.I.P.E.S. and the probability of students returning to LSU when not enrolled in S.T.R.I.P.E.S., while controlling for gender. In Table 15, the relative risk gives a value of 1.7317. Since this value is higher than 1, retention is more likely to occur in students that participated in S.T.R.I.P.E.S. than students that did not. The odds ratio differs in what it calculates for the data, as it provides an analysis for the likelihood of an event to occur (Agresti 22). In this case where gender is controlled, the odds ratio analyzes the odds for students to return to LSU if they were enrolled in S.T.R.I.P.E.S. compared to the odds for students to return if they had not been enrolled in the program. The odds ratio determined in Table 15 by comparing these likelihoods shows that the odds of a student enrolled in the S.T.R.I.P.E.S. program returning to LSU is 2.015 times higher than the odds for a student not enrolled in the S.T.R.I.P.E.S. program (Appendix D).

2.2.2. Race, S.T.R.I.P.E.S., and Retention

The relationship between S.T.R.I.P.E.S. and retention, controlling for race, was also analyzed. To simplify the process of analyzing, race was split into categories of White and non-White. The White category classified students, both male and female, that identify as white or Caucasian in race. The non-White category included students, both male and female, that identify as Black, Hispanic, Nonresident Alien (where race is unknown), Asian, American Indian, or multiracial. Tables 16 and 17 allow for

comparison of race's relationship to S.T.R.I.P.E.S. enrollment and retention at LSU (Appendix E). These tables show the number of White and non-White students enrolled in the S.T.R.I.P.E.S. program in fall 2011 that returned to LSU in the following spring 2012 (Appendix E).

The Chi Square test is performed for partial tables of White race and non-White race. When the Chi-Square Test is utilized on this data for non-White race (as seen in Table 16), it yields a result of a probability value of 0.4696. Since the probability value when controlling for non-White race is greater than 0.05, the result is not significant. This could be caused by the low Chi-Square count seen in Table 16 (Appendix E). When the Chi-Square Test is applied to the data for students that identify as part of the White race (as seen in Table 17 in Appendix E), it gives a probability value of 0.0536, which is a moderately significant result.

The Cochran-Mantel-Haenszel Statistics Model was used to check the null hypothesis of conditional independence between S.T.R.I.P.E.S. and retention, controlling for race. The probability value of 0.0493 suggests a significant association between retention and S.T.R.I.P.E.S., implying increased odds of retention for S.T.R.I.P.E.S. participants when controlling for race. This is true for either race, White or non-White. The association between both retention and S.T.R.I.P.E.S., although significant, does not vary across race, since the probability value of the Breslow-Day Test is 0.3829.

For the case of race, the odds ratio and relative risk were determined in Table 18 (Appendix E). The relative risk test determined the ratio between the probability of students returning to LSU when enrolled in S.T.R.I.P.E.S. and the probability of students returning to LSU when not enrolled in S.T.R.I.P.E.S., while controlling for race. In

Table 18, the relative risk gives a value of 2.0341. Since this value is greater than 1, this means that students that participate in S.T.R.I.P.E.S. are more likely to achieve retention than students not enrolled in S.T.R.I.P.E.S., while controlling for race. The odds ratio for this case conveys the odds for students to return to LSU if they were enrolled in S.T.R.I.P.E.S. compared to the odds for students to return if they had not been enrolled, controlling for race. The odds ratio determined in Table 18 by comparing these likelihoods shows that the odds of a student enrolled in the S.T.R.I.P.E.S. program returning to LSU is 2.38 times higher than the odds for a student not enrolled in the S.T.R.I.P.E.S. program (Appendix E).

2.2.3. Grade Point Average, S.T.R.I.P.E.S., and Retention

The final relationship that was considered is the relationship between S.T.R.I.P.E.S. and retention when controlling for grade point average. The values of grade point average are typically reported on a 4.0 scale, where each value on the scale corresponds to a given grade. For this research, grade point average was categorized on whether it increased or stayed the same from fall 2011 to spring 2012 or if the grade point average decreased on that same time interval. Tables 19 and 20 allow for the study of S.T.R.I.P.E.S. enrollment and retention at LSU when controlling for grade point average (Appendix F).

When controlling for grade point average, the significance of S.T.R.I.P.E.S. participation on retention is calculated with the Chi-Square Test. The Chi Square test is performed for partial tables of both of these categories of students. For the case when grade point average is increasing or staying the same for students from their first semester to their second, the Chi-Square Test gave a probability value of 0.1086, which is greater

than the needed value of 0.05 (Appendix F). This is an indication that for an increasing or steady grade point average, the result is insignificant. Therefore, retention rates given by S.T.R.I.P.E.S. involvement are conditionally independent of an increasing or steady grade point average. For the case where grade point average is decreasing, the Chi-Square Test gives a probability value of 0.3207, which is also greater than the needed value of 0.05. This denotes that for a decreasing grade point average, the result is also insignificant. Therefore, retention rates given by S.T.R.I.P.E.S. involvement are conditionally independent of a decreasing grade point average.

The Cochran-Mantel-Haenszel Statistics Model checked the null hypothesis of conditional independence between S.T.R.I.P.E.S. and retention, controlling for grade point average. The probability value for the Cochran-Mantel-Haenszel Test was 0.0690, which is moderately significant. The probability value of 0.0690 suggests a moderately significant association between retention and S.T.R.I.P.E.S., implying moderately increased odds of retention for S.T.R.I.P.E.S. participants. The association between both retention and S.T.R.I.P.E.S., while moderately significant, does not vary across grade point average, since the probability value of the Breslow-Day Test is 0.3377.

The next aspect of the S.T.R.I.P.E.S. and retention relationship to test is for the values of the odds ratio and relative risk, which are both given values in Table 21 (Appendix F). The relative risk test determined the ratio between the probability of students returning to LSU when enrolled in S.T.R.I.P.E.S. and the probability of students returning to LSU when not enrolled in S.T.R.I.P.E.S., while controlling for grade point average. In Table 21 (Appendix F), the relative risk gives a value of 3.2073, which is greater than 1. When controlling for grade point average, the data shows a higher

likelihood for retention to occur when students are enrolled in the S.T.R.I.P.E.S. program. The odds ratio for this case reveals the odds for students to return to LSU if they were enrolled in S.T.R.I.P.E.S. compared to the odds for students to return if they had not been enrolled, controlling for grade point average. The odds ratio determined in Table 21 (Appendix F) by comparing these likelihoods shows that the odds of a student enrolled in the S.T.R.I.P.E.S. program returning to LSU is 3.45 times higher than the odds for a student not enrolled in the S.T.R.I.P.E.S. program.

2.3. Discussion of Results

The effect of S.T.R.I.P.E.S. on retention is seen in each control variable of gender, race, and grade point average. For all of the control variables, it is seen that there is a moderately significant association between retention and S.T.R.I.P.E.S. participants. Also, the values of the three control variables are shown to be independent of S.T.R.I.P.E.S. predictive value of retention. Overall, it appears that controlling for those variables and predicting the impact of S.T.R.I.P.E.S. involvement on retention does not give a highly significant result, but does produce a moderate effect. For each control variable, the data illustrated a higher likelihood for retention to occur when students were enrolled in the S.T.R.I.P.E.S. program. It was also seen for each control variable that the odds of a student enrolled in the S.T.R.I.P.E.S. program returning to LSU was higher than the odds for a student not enrolled in the S.T.R.I.P.E.S. program. These values of odds ratio and relative risk agree with the idea that S.T.R.I.P.E.S. does produce a moderately significant effect on retention.

The random set of 514 first-year students is a small portion of the LSU student body. Although Tables 1-5 (Appendix A) revealed a higher rate of retention in the group

of students that participated in S.T.R.I.P.E.S., the subset of the data that represents S.T.R.I.P.E.S. participants is small in comparison to the total number of students in the data set. Therefore, it is possible that the dataset may not be large enough to allow for a significant analysis in SAS.

Several issues have been identified as possible causes for contributing to the moderate level of significance. The first suggestion for contributing to the moderate significance of the data is the low Chi-Square value in Tables 14, 16, 17, 19 and 20 (Appendices C, D, and E, respectively). For these Chi-Square Tests, the number of expected counts for retention report as less than 5, which indicates the value could skew the data. According to McHugh, "the value of the cell expected should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one" (McHugh 144). This assumption of the Chi-Square Test helps to ensure that the counts are high enough to perform the needed calculations. Having reported less than a value of 5 for some of these tables suggests that it is possible the Chi-Square Test did not have enough counts to properly estimate the data.

It is possible that the data could have produced errors when it was sampled. Some of the data present in the sample, such as race, gender, and grade point average, all come from LSU and are reported to the Office of Budget and Planning where the data was retrieved. Some of the elements of the data, though, such as S.T.R.I.P.E.S. participation or Greek involvement, could contain error since it is reported by the individual, subject to change, and not automatically known through the university data system. Also, the Office of Budget and Planning gave a dataset for this research that included a random

group of 514 first-year students from fall 2011. It is possible that this random selection is not a good representation of the LSU student body of first-year students from that year.

Categorical limitations of the data may have caused the data to error as well. The factor of race was split into only two categories, White and non-White. Although this factor was split into only two categories, the non-White race category actually represented six different races total. Also, the method that split grade point average into categories may have caused errors. Each time a student's grade point average stayed the same from one semester to the next, that student was assigned the same value as an increasing grade point average. In some cases where the grade point average was zero the first semester and remained zero the second semester, due to a student not finishing the fall 2011 semester, the grade point average was still recorded as staying the same or increasing. Also, the severity of decrease or increase in grade point average is not evident in this nominal format of analysis. When a student's grade point average decreased from fall 2011 to spring 2012, it is not evident if the grades dropped by one letter grade or more. By categorizing the grade point average based on increasing or decreasing, the student could be severely failing or could have dropped from a high grade to a slightly less but still high grade. These different remarks of the nature of the data could have had effects on the analysis.

Finally, the management of the data through the process of list-wise deletion could have caused issues in the dataset. When the data was first received from the Office of Budget and Planning, some values for student's gender, race, or other variables were missing from the original data set. Since the values were missing, the data for those students was deleted from the overall list by SAS as to not assume any information about

those students. This deletion occurred for the specific SAS computations that the deletions affected. By completing this deletion method, it was ensured that no important information was missing from the students chosen in the data set. By deleting these students, their count was not included in the data and could have swayed the significance of the data.

These are all considerations that could have affected the results of the analysis. It is also important to note that deciding to return after the first semester of college could be hindered by outside factors as well, which could be another form of future work to be considered for research.

3. LIMITATIONS AND CONCLUSIONS

After acknowledging some potential problems in the sampling and testing of the data, this research could continue by attempting to solve some of those issues of potential errors mentioned in the previous discussion. Some suggestions for future work include looking at a larger population of students over multiple years. Rather than tracking a first-year student's retention from the first semester to the second semester, it could be more beneficial to track that student's retention from the start of their time at LSU to graduation, however long that may be. By tracking multiple years of S.T.R.I.P.E.S. data, it may be interesting to compare the statistical significance from year to year of the program. This could show how the significance of the relationship of S.T.R.I.P.E.S. on retention has changed over the years of S.T.R.I.P.E.S. being in place at LSU. In general, it appears that a larger data set is needed to better ensure S.T.R.I.P.E.S. as a predictor of retention. This would eliminate the issues of low Chi-Square count warnings that occurred in some of the control variables for S.T.R.I.P.E.S. participation.

Another suggestion would be to look at other first-year programs that provide support to students. For example, similar statistical analysis could be run on first-year residential community involvement while controlling for similar variables like race, gender, grade point average, among others. After analyzing multiple first year-programs, the results could be compared to see which program has the most significant impact on retention. In this research, only one variable was controlled for at a time. In the future, it may be useful to use logistic regression to control for all three variables at once and see the impact of S.T.R.I.P.E.S. on retention. It may be useful to analyze the effect of S.T.R.I.P.E.S. on retention without controlling for any variables, as was done with this research. If a more significant result is seen when variables are controlled for, then this is a positive sign that controlling for more variables at once may produce a more significant result.

Overall, for the data provided by the Office of Budget and Planning, this analysis provided a useful gauge of which variables to concentrate on when analyzing retention. It is possible that the S.T.R.I.P.E.S. program, or similar programs, could be a better predictor of retention in first-year students when controlling for other variables if some of the potential errors are given attention.

APPENDIX A: SUMMARY OF DATA

Table 2. Retained vs. Not Retained Female Students

Total Sample of Female Students (*n* = 267, 52% total)		
Number of Female Students Retained	226	85%
Number of Female Students Not Retained	41	15%
Total Female Students Enrolled in S.T.R.I.P.E.S. (*n* = 40, 15%)		
Number of Students Retained	38	95%
Number of Students Not Retained	2	5%
Total Female Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 227, 85%)		
Number of Students Retained	188	83%
Number of Students Not Retained	39	17%

Table 3. Retained vs. Not Retained Male Students

Total Sample of Male Students (*n* = 247, 48% total)		
Number of Male Students Retained	197	80%
Number of Male Students Not Retained	50	20%
Total Male Students Enrolled in S.T.R.I.P.E.S. (*n* = 17, 7%)		
Number of Students Retained	14	82%
Number of Students Not Retained	3	18%
Total Male Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 230, 93%)		
Number of Students Retained	183	80%
Number of Students Not Retained	47	20%

Table 4. Retained vs. Not Retained White Students

Total Sample of White Students (*n* =396, 77% total)		
Number of White Students Retained	330	83%
Number of White Students Not Retained	66	17%
Total White Students Enrolled in S.T.R.I.P.E.S. (*n* = 37, 9%)		
Number of Students Retained	35	95%
Number of Students Not Retained	2	5%
Total White Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 359, 91%)		
Number of Students Retained	295	82%
Number of Students Not Retained	64	18%

Table 5. Retained vs. Not Retained Non-White Students

Total Sample of Non-White Students (*n* = 119, 23% total)		
Number of Non-White Students Retained	94	79%
Number of Non-White Students Not Retained	25	21%
Total Non-White Students Enrolled in S.T.R.I.P.E.S. (*n* = 20, 17%)		
Number of Students Retained	17	85%
Number of Students Not Retained	3	15%
Total Non-White Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 99, 83%)		
Number of Students Retained	77	78%
Number of Students Not Retained	22	22%

Table 6. Retained vs. Not Retained Increasing/Same Grade Point Average Students

Total Sample of Increasing/Same GPA Students (*n* = 218, 47% total)		
Number of Increasing/Same GPA Students Retained	199	91%
Number of Increasing/Same GPA Students Not Retained	19	9%
Total Increasing/Same GPA Students Enrolled in S.T.R.I.P.E.S. (*n* = 24, 11%)		
Number of Students Retained	24	100%
Number of Students Not Retained	0	0%
Total Increasing/Same GPA Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 194, 89%)		
Number of Students Retained	175	90%
Number of Students Not Retained	19	10%

Table 7. Retained vs. Not Retained Decreasing Grade Point Average Students

Total Sample of Decreasing GPA Students (*n* = 244, 53% total)		
Number of Decreasing GPA Students Retained	224	92%
Number of Decreasing GPA Students Not Retained	20	8%
Total Decreasing GPA Students Enrolled in S.T.R.I.P.E.S. (*n* = 29, 12%)		
Number of Students Retained	28	97%
Number of Students Not Retained	1	3%
Total Decreasing GPA Students Not Enrolled in S.T.R.I.P.E.S. (*n* = 215, 88%)		
Number of Students Retained	196	91%
Number of Students Not Retained	19	9%

APPENDIX B: SIMPLE LINEAR REGRESSION, S.T.R.I.P.E.S. VERSUS RETENTION

Table 8. Analysis of Variance for S.T.R.I.P.E.S. and Retention.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.50746	0.50746	3.50	0.0620
Error	513	74.41293	0.14505		
Corrected Total	514	74.92039			

Root MSE	0.38086
Dependent Mean	0.82330
Coefficient Variable	46.26014
R-Square	0.0068
Adj R-Square	0.0048

Table 9. Parameter Estimates, Dependent Variable: Retention

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.81223	0.01780	45.64	<.0001
Stripes	**1**	0.10005	0.05349	1.87	0.0620

Table 10. Correlation of Estimates, Dependent Variable: Retention

Correlation of Estimates		
Variable	**Intercept**	**Stripes**
Intercept	1.0000	-0.3327
Stripes	-0.3327	1.0000

APPENDIX C: S.T.R.I.P.E.S. VERSUS RETENTION

Table 11. S.T.R.I.P.E.S. versus Retention

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	86	371	457
Percent		16.73	72.18	88.91
Row %		18.82	81.18	
Col %		94.51	87.71	
Frequency	**Yes**	5	52	57
Percent		0.97	10.12	11.09
Row %		8.7	91.23	
Col %		5.49	12.29	
	Total	92	423	514
		17.70	82.30	100.00

Table 12. Estimates of Common Relative Risk for S.T.R.I.P.E.S. versus Retention

Type of Study	Method of Test	Statistical Value
Odds Ratio	Logit	2.4108
Relative Risk	Logit	2.1453

APPENDIX D: GENDER, S.T.R.I.P.E.S., AND RETENTION

Table 13. Three-Way Contingency table of S.T.R.I.P.E.S. by Retention Controlling for Female Gender (fall 2011)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	39	188	227
Percent		14.61	70.41	85.02
Row %		17.18	82.82	
Col %		95.12	83.19	
Frequency	**Yes**	2	38	40
Percent		.75	14.23	14.98
Row %		5.00	95.00	
Col %		4.88	16.81	
	Total	41	226	267
		15.36	84.64	100.00

Table 14. Three-Way Contingency Table of S.T.R.I.P.E.S. by Retention Controlling for Male Gender (fall 2011)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	47	183	230
Percent		19.03	74.09	93.12
Row %		20.43	79.57	
Col %		94.00	92.89	
Frequency	**Yes**	3	14	17
Percent		1.21	5.67	6.88
Row %		17.65	82.35	
Col %		6.00	7.11	
	Total	50	197	247
		20.24	79.76	100.00

Table 15. Estimates of Common Relative Risk Controlling for Gender

Type of Study	Method of Test	Statistical Value
Odds Ratio	Logit	2.0150
Relative Risk	Logit	1.7317

APPENDIX E: RACE, S.T.R.I.P.E.S., AND RETENTION

Table 16. Three-Way Contingency Table of S.T.R.I.P.E.S. by Retention Controlling Other (Non-White) Race (fall 2011)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	22	77	99
Percent		18.49	64.71	83.19
Row %		22.22	77.78	
Col %		88.00	81.91	
Frequency	**Yes**	3	17	20
Percent		2.52	14.29	16.81
Row %		15.00	85.00	
Col %		12.00	18.09	
	Total	25	94	119
		21.01	78.99	100.00

Table 17. Three-Way Contingency Table of S.T.R.I.P.E.S. by Retention controlling White Race (fall 2011)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	64	295	359
Percent		16.16	74.49	90.66
Row %		17.83	82.17	
Col %		96.97	89.39	
Frequency	**Yes**	2	35	37
Percent		.51	8.84	9.34
Row %		5.41	94.59	
Col %		3.03	10.61	
	Total	66	330	396
		16.67	83.33	100.00

Table 18. Estimates of Common Relative Risk Controlling for Race (White and non-White)

Type of Study	Method of Test	Statistical Value
Odds Ratio	Logit	2.3787
Relative Risk	Logit	2.0341

APPENDIX F: GRADE POINT AVERAGE, S.T.R.I.P.E.S., AND RETENTION

Table 19. Three-Way Contingency Table of S.T.R.I.P.E.S. by Retention controlling for Increasing/Same Grade Point Average (fall 2011 to spring 2012)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	19	175	194
Percent		8.72	80.28	88.99
Row %		9.79	90.21	
Col %		100.00	87.94	
Frequency	**Yes**	0	24	24
Percent		0.00	11.01	11.01
Row %		0.00	100.00	
Col %		0.00	12.06	
	Total	19	199	218
		8.72	91.28	100.00

Table 20. Three-Way Contingency Table of S.T.R.I.P.E.S. by Retention controlling for Decreasing Grade Point Average (fall 2011 to spring 2012)

	S.T.R.I.P.E.S. Involvement (Yes/No)	Retention (0=No, 1=Yes)		
		0	**1**	**Total**
Frequency	**No**	19	196	215
Percent		7.79	80.33	88.11
Row %		8.84	91.16	
Col %		95.00	87.50	
Frequency	**Yes**	1	28	29
Percent		0.41	11.48	11.89
Row %		3.45	96.55	
Col %		5.00	12.50	
	Total	20	224	244
		8.20	91.80	100.00

Table 21. Estimates of Common Relative Risk Controlling Grade Point Average

Type of Study	Method of Test	Statistical Value
Odds Ratio	Logit	3.4454
Relative Risk	Logit	3.2073

REFERENCES

ACT (2013). National collegiate retention and persistence to degree rates. Retrieved March 2, 2014 from http://act.org/research/policymakers/pdf/retain_2013.pdf

Addo, Koran. (2014, January), "LSU's graduation rate at all-time high… again," *The Advocate*. Retrieved from http://theadvocate.com/news/8100071-123/lsus-graduation-rate-reaches-all-time

Agresti, Alan. (1996), "An Introduction to Categorical Data Analysis," John Wiley & Sons, Inc.

Beral, Valerie. (1979), "On The Estimation of Relative Risk From Vital Statistical Data,"
Journal of Epidemiology and Community Health (1979-), Vol. 33, No. 2 (Jun., 1979), pp. 159-162.

Bland, J. Martin, and Altman, Douglas G. (2000), "The Odds Ratio," *BMJ: British Medical Journal*, Vol. 320, No. 7247 (May 27, 2000), p. 1468.

"Homogeneity." *Merriam-Webster.com*. Merriam-Webster, n.d. Web. 9 Mar. 2014. <http://www.merriam-webster.com/dictionary/homogeneity>.

Light, Richard J. and Margolin, Barry H. (1971), "An Analysis Of Variance For Categorical Data," *Journal of the American Statistical Association,* Vol. 66, No. 335, pp. 534-544.

LSU STRIPES. LSU Division of Student Life and Enrollment, 2014, n.p. n.d.. Web. 9 Mar. 2014. <http://students.lsu.edu/fye/programs/stripes>

LSU STRIPES Handbook. LSU Division of Student Life and Enrollment, First Year Experience, 2014, n.p. n.d.. Web. 9 Mar. 2014. <http://students.lsu.edu/sites/default/files/Stripes%20Handbook_4.pdf>

McHugh, Mary. (2013), "The Chi-square test of independence," *Biochemia Medica* 2013,
Pp. 143-149.

SAS Institute Inc., *SAS/STAT ® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

Woosley, Sherry A. "How Important Are The First Few Weeks Of College? The Long Term Effects Of Initial College Experiences." *College Student Journal* 37.2 (2003): 201. *Academic Search Complete*. Web. 3 Mar. 2014.

Yanagawa, T., Fujii, Y., and Mastuoka, J. "Generalized Mantel-Haenszel Procedures for

2 x J Tables." *Environmental Health Perspectives*, Vol. 102, Supplement 8: Biostatistics in the Study of Human Cancer (Nov., 1994), pp. 57-60.

Yu, Binbing and Gastwirth, Joseph L. (2008), "A Method Of Assessing The Sensitivity Of The Cochran-Mantel-Haenszel Test To An Unobserved Confounder," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, Vol. 366, No. 1874, Mathematical and Statistical Methods for Diagnoses and Therapies, pp. 2377-2388.

ACKNOWLEDGMENTS

First, I would like to thank my thesis advisor, Professor George Cochran. Without his patience, guidance, and care, this research and thesis would not have been possible. His willingness to listen and work with me to create a topic that not only interested me but also benefitted my future career was extremely helpful. Secondly, I would like to thank the rest of my thesis committee, including Professor Brian Marx and Instructor Guy Forrest. Thank you for the insightful questions and honest encouragement throughout this process. This project would also not have been possible without the help of Professor Marx's graduate student, Leaha Davis, who took a personal interest in my research and helped with the conception of the project. Also, this project would not have been possible without the data from the LSU Office of Budget and Planning and the help and guidance of Michel Schnexnayder within that office.

I would also like to thank the LSU Honors College and Marybeth Smith for their guidance and support throughout the past four years, and especially during my time spent researching. Their genuine care for my educational wellbeing is much appreciated. I appreciate the challenge that the LA-STEM Research Scholars Program instilled within me to always search for answers and strive to make a better world. Their approach to research is what inspired me to continue with my studies in mathematics and within the LSU Honors College. Finally, I thank my family and friends for keeping me motivated to accomplish my goals until the end. Throughout the hard times of researching and writing, I received nothing but enthusiastic encouragement, without which this thesis would not have been possible.