UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

APPLYING HIERARCHICAL NON-LINEAR MODELING AND CLUSTER ANALYSIS TO INVESTIGATE THE INFLUENCE OF HIGH SCHOOL CHARACTERISTICS IN FIRST-YEAR COLLEGE RETENTION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JACOB PLEITZ
Norman, Oklahoma
2014

APPLYING HIERARCHICAL NON-LINEAR MODELING
TO INVESTIGATE THE INFLUENCE OF HIGH SCHOOL
CHARACTERISTICS IN PREDICTING FIRST-YEAR COLLEGE RETENTION

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Robert Terry, Chair

Dr. Nicole Judice Campbell

Dr. Jorge Mendoza

Dr. Hairong Song

Dr. Barbara Greene

To my Parents. Boomer.

ACKNOWLEDGMENTS

The following dissertation would not have been possible without the invaluable contributions of a number of people. Throughout my graduate tenure, I have been blessed to work with some of the finest researchers and analytical minds in the industry, and the following represents the culmination of years of patience, training, and effort on their parts.

To my major professor and academic advisor Dr. Robert Terry. Friend, advisor, confidant, and counselor, Dr. Terry has worn many shoes while walking with me, and for that I will forever be grateful. More than simply providing academic counseling, Dr. Terry has taught me the reigns of professionalism in the academic environment, overseen my personal and intellectual development, and been a constant guide and friend throughout the rigors of grad school.

To Dean Nicole Judice Campbell, one of the kindest and smartest colleagues I have ever been blessed to work with. The current research would not have been possible without Dean Campbell's passion for higher education, and I am blessed every day to have worked with such a wonderful person.

To my committee, Dr. Hairong Song, Dr. Jorge Mendoza and Dr. Barbara Green, whose guidance and supervision throughout this process has been invaluable. To my many friends I have made while at OU: Caleb, Collin, Daniel, Tyler, Cal, Coach, Luisa and Alyssa – the best friends a person could ever have.

And finally, to my family: Josh, Jordan, Krystal, Oliver, Maggy, the three amigos (Desi, Charli and Laci) and my pride and joy, Esther Lillian.

TABLE OF CONTENTS

ACKNOWLEDGMENTS....iv
TABLE OF CONTENTS....v
LIST OF TABLES....vii
LIST OF FIGURES....viii
ABSTRACT....ix
CHAPTER I: INTRODUCTION....1
Overview of the Current Dissertation....2
CHAPTER II: METHODOLOGY REVIEW....7
Overview....7
Clustering Analysis....7
Describing Clusters....9
Ward's Method....11
Hierarchical Agglomerative Methods....12
Representing The Results of Ward's Method....14
Shortcomings of Ward's Method....15
K-means Clustering....15
Combining the Two Methods....17
Multi-Level Modeling....18
Summary....20
CHAPTER III: VARIABLES IN THE CURRENT ANALYSIS....21
Overview....21
Student-Level Variables....21
High School GPA and Standardized Test Scores....21
Self-Efficacy....24
Academic Engagement....26
Financial Concerns....29
Institutional Commitment....32
School-Level Variables....34
High School Class Size / Student-Teacher Ratio....34
District Spending Per Student/ Free Lunch Eligibility....36
School Sector....38
Summary....40
CHAPTER IV: SAMPLE AND METHODOLOGY....41
Overview....41
Participants....41
Creating the Data Set....41
Sample Descriptive Statistics....43
Procedure....47
Phase 1....48
Phase 2....49
Phase 3....50
Phase 4....50

Summary....50
CHAPTER V: RESULTS....52
Overview:....52
Factor Analytic Results....52
Internal Reliability Results....53
Clustering Analysis Results....55
Ward's Method....55
K-Means Clustering Method....60
Differences in Means Between the Groups....64
Non-Mixed Logistic Regression Models....65
Student Level Variables....65
School Variables....67
Multi-Level Models....68
Examining the Intra-Class Correlation....68
Model 1....69
Model 2....71
Model 3....72
Model 4....73
Summary of Fit Statistics....75
Summary....76
CHAPTER VI: DISCUSSION....78
Overview....78
Student-Level Variables....78
Self-Efficacy....79
Academic Engagement....79
Financial Concerns....81
School-Level Variables....83
Free Lunch and Class Size....83
The Role of Academic Engagement in Retention....84
Implications for Policy....85
Implications for Students (Learned vs. Earned)....86
Implications for Administrators and Educators....87
Conclusions....89
REFERENCES....92
APPENDIX 1: Summary of Variables....106
APPENDIX 2: Additional Canonical Variate Graphs....108
Appendix 3: Results and Description of Model 5....110
Appendix 4: Missing and Final Data File Comparisons....112

LIST OF TABLES

Table 1: Descriptive Statistics for Adjustment Variables 43
Table 2: Descriptive Statistics for High School Behaviors....44
Table 3: Descriptive Statistics for Academic Engagement Variables....45
Table 4: Descriptive Statistics for School and Student Variables....46
Table 5: Descriptive Statistics for Importance Questions 46
Table 6: Descriptive Statistics for Experiences and Expectations Variables....47
Table 7: Factor Loadings After Rotation 53
Table 8: Internal Reliability and Correlations for Financial Concerns....54
Table 9: Internal Reliability and Correlations for Academic Engagement....54
Table 10: Internal Reliability and Correlations for Self-Efficacy....54
Table 11: Internal Reliability and Correlations for Institutional Commitment .55
Table 12: Eigenvalues of Covariance Matrix 57
Table 13: Descriptive Statistics for Ward's Method Clustering 57
Table 14: City Type Breakdown for Clusters 58
Table 15: Canonical Coefficients....59
Table 16: Statistics for K-Means Clustering....61
Table 17: Descriptive Statistics for K Means Clustering 61
Table 18: Classification Of Observations 63
Table 19: Summary of Results From Model 4....74
Table 20: Summary of Fit Statistics....75
Table 21: Variables in the Current Analysis 106
Table 22: Correlations Amongst School-Level Variables.... 107
Table 23: Summary of Results from Model 5.... 110
Table 24: Descriptive Statistics from Full File.... 112
Table 25: Descriptive Statistics from Final File 113

LIST OF FIGURES

Figure 1: Cluster Analysis Plot For Ward's Method 56
Figure 2: Cluster Analysis of High Schools Using Ward's Method 60
Figure 3: Scatterplot of K-Means Clustering........ 62
Figure 4: Canonical Variate Axes........ 64
Figure 5: Canonical Variate Axis 2 by Graduation Rate........ 108
Figure 6: Canonical Variate Axis 2 by Free Lunch........ 108
Figure 7: Canonical Variate Axis 2 by Class Size........ 109

ABSTRACT

Student retention in higher education is an incredibly important social and psychological phenomenon. The impact of student retention reaches across multiple domains, influencing individual students, state and local economies, and even national prestige and viability on a global scale. Although a great deal of research has been conducted examining the influence of both student and school characteristics on student retention, less is known about how the two variables interact with each other. The present dissertation is designed to examine this phenomenon by applying multi-level modeling to estimate how variables at the school level (such as graduating class size and district spending) interact with variables at the student level (such as high school GPA and financial concerns) to predict first-year retention in college. Psychosocial and academic data were collected from over 6,500 students across 950 schools and applied to construct a series of multi-level models to estimate retention. Clustering analysis was then applied to see if the schools could be grouped according to “type” rather than used individually. Results indicated that multi-level models could be used to predict student retention in higher education, and that the most influential predictors of retention were academic and financial in nature. Implications and future research are discussed.

CHAPTER I: INTRODUCTION

Researchers have long been interested in the causes for student attrition (Bean, 1980; Tinto, 1975; 1993; Braxton, 1999; Braxton, Hirschy & McClendon, 2004; Munt & Merydith, 2011-2012; Pascarella & Terenzini, 1980; 2005). Although there have been great strides made in understanding the process of student attrition, results show that the overall six-year graduation rate in colleges still remains only 56% (Shapiro et al, 2013)[1]. According to a 2011 ACT report, approximately 22% of students at public, Ph.D. granting institutions fail to return for their sophomore year (ACT, 2011).

The high cost of student attrition is evidenced at multiple levels. At the student level, not having a college degree may result in multiple negative life outcomes. Results have shown that, compared to their peers, students without a college degree are more likely to suffer from lower economic earning potential (U. S. Census Bureau, 2010), be more likely to end up in prison (Sum, Khatiwada, McLaughlin & Palma, 2009), and be more likely to live in poverty (National Center for Educational Statistics, 2011). Additionally, these individuals are less likely to have children who will attend college, effectually creating a cycle of potential negative impact throughout the family.

At the institutional level, the high costs of attrition are felt as well. Increased attrition rates have many negative impacts on a school, including fewer economic resources through lost tuition and fee revenue, as well as declining national prestige

---

[1] Retrieved February, 24, 2014 from http://www.huffingtonpost.com/2013/12/16/college-students-finish-degrees-study_n_4455026.html

and rankings. These hardships are becoming increasingly felt as state and local governments are forced to cut funding for higher education. According to a 2010 report by the American Institutes for Research (AIR), student attrition cost the states an estimated $1.4 billion in student grants, with 13 state posting a loss of $200 million in state funds for first year dropouts (AIR, 2010).

Finally from a national level, the increasing costs of attrition are beginning to be felt as the United States begins to fall behind other countries in terms of math and science production. According to a 2008 study by the National Science Board, the United States has fallen behind other countries in terms of the percentage of the population receiving a higher education. These effects are ultimately felt in the decreased scientific capital of the nation, along with lower earning potential of the citizens—a factor that ultimately harms the national economy.

Overview of the Current Dissertation

Given the large costs of student attrition, it is clear that keeping students enrolled in higher education is an incredibly important goal. Because of this, the current dissertation seeks to apply clustering analysis and multi-level modeling to investigate the how academic and non-academic factors predict persistence to the second year in college. Along with examining the influence of student- and school-level variables in predicting retention, the current dissertation also seeks to build upon the previous work of sociological researchers (such as Durkheim, 1897) by examining how the dropout process may be moderated by factors such as integration and a sense of belonging.

The evaluation of student-level characteristics on predicting retention is nothing new. Researchers have long been interested in how high school GPA and standardized test scores predict performance at the next level. The influence of non-academic factors has become increasingly popular within recent years. Factors including emotional maturity (Sparkman, Maulding & Roberts, 2012), self-efficacy, expectations, and institutional commitment (Tinto, 1975) have become increasingly useful in predicting withdrawal, particularly for students who are not at risk for academic failure.

As with student-level characteristics, the evaluation of student and high school characteristics on academic performance is also nothing new. Specifically, the influence of high school characteristics has lead to the ground breaking legislative movements including Brown vs. Topeka (1954) where the Warren court voted unanimously that the separate but equal clause of Plessy vs. Ferguson (1896) was unconstitutional on the grounds that "separate educational facilities are inherently unequal…" (Brown, 1954).

Although Brown v. Topeka marked a groundbreaking turn in educational equality across the United States, there continue to be differences in high schools and students entering higher education. Differences in high school class sizes, student-teacher ratios, average standardized test scores, district spending per student and a variety of other variables continue to create differences in high schools across the nation. This discrepancy is further marked by the ever-present differences in student variables, including high school GPA, academic engagement, institutional commitment, and others.

Given the variability in student and institutional characteristics, it is no surprise that higher education researchers have become interested in how these variables influence academic performance. Commissioned in the 1960's, the work of Coleman et al., (hereafter referred to as the Coleman Report) was one of the largest undertakings of educational research in history. Aimed at researching educational inequalities in the United States, the report analyzed data from over 650,000 students to investigate how school-level variables (including funding and spending per student) interacted with student-level variables such as background and SES to influence performance (Coleman, 1966).

Although Brown v. Topeka and the Coleman report have provided strong foundations for educational research and national incentives, there continues to be a variety of questions still remaining about how school and student-level variables influence academic performance and behavior. Within this framework, the current dissertation has two primary goals. The first goal is to examine how student and school-level characteristics interact to influence a student's predicted probability of second-year retention. By examining the influence of these variables, the current dissertation seeks to provide a more in-depth, and multi-level model of student retention.

The second goal is to examine whether high schools may be clustered according to "type" to ease in model construction and interpretation. Although it may be possible to construct multi-level models to better predict retention, the difficulty in estimating unique intercepts or slopes for over 950 schools quickly causes convergence and processing difficulties. Because of this, if it is possible to cluster the schools according to a specific “type”, then these estimates can potentially be reduced to only intercepts for the clusters (effectively eliminating a majority of the processing time, as well as potentially eliminating any worry associated with discrimination against a specific school).

Research in higher education has come a long way since the days of Brown and Plessy. And thankfully the atrocities of mandatory segregation have been removed from the educational system. Yet the scars still remain. As will be discussed in the current dissertation, there still exists a vast inequality in the educational opportunities of high school students. Differences in school spending, academic challenges, and student-teacher ratios continue to impact high school students academic performances in college, with the results often times significantly increasing or decreasing the probability of persistence in higher education.

The remaining chapters proceed as follows. First, the methods for clustering and multi-level modeling are discussed. Generally speaking, cluster analysis allows for the grouping of second-level variables (in this case high schools) according to certain characteristics. As previously stated, this will allow (ideally) for a more clean understanding of how certain high school “types” are classified, as well as provide for a cleaner method of analysis.

Next, the psychosocial and academic variables are presented. Because the psychosocial variables have to be created through factor analysis, the steps taken to create the variables are detailed, and the methods for determining internal reliability are discussed. Additionally, the background and previous research findings associated with the variables are discussed. Specific variables include both traditional academic variables (high school GPA, standardized test scores) as well as more aggregate school variables (private vs. public sector).

After detailing the background information, the remaining chapters present the specific methodology used for assessment and then the discussion of the findings. The results section presents findings in a naturally sequential order, and begins with the data cleaning and variable creation, before moving into the clustering analysis, and then concluding with findings from the different models. Models are presented in sequential order as well, beginning with the individual level models before moving onto the final mixed-models.

The dissertation then concludes by discussing implications for findings, identifying research shortcomings, and then making suggestions for future research. Each chapter is introduced by a brief overview giving the reader a preview of things to come, and is then book-ended with a conclusion providing a brief summary of the chapter findings. Tables are presented within the text when necessary, however certain findings are detailed in the appendix when more appropriately referenced in such a way. Interested readers are encouraged to seek out the supplied references to gain a greater understanding of any specific material.

CHAPTER II: METHODOLOGY REVIEW

Overview

This chapter provides the background and overview of the methods applied in the current research. Specific methods include clustering analysis and multi-level non-linear modeling. Within the cluster analyses, two separate methods were used. The first method was Ward's method and the second method was K-means. After performing the analyses, the results were compared for agreement on item location. After the clustering analyses, multi-level modeling was performed to estimate the influence of using either the clusters or individual school level variables as level-two predictors, and the student level predictors as level-one predictors.

Clustering Analysis

Although it is possible to fit a unique equation for each of the 800 plus high schools within the current sample, it is both highly impractical, as well as statistically unstable. Specifically, the unique equation derived from each school would be extremely time consuming and confusing for admissions committees, while low sample sizes from several schools contribute to high instability of estimates. Fortunately, there exist a number of statistical methods that may be applied to group schools according to specific types. One family of procedures that can be useful in grouping items is clustering analysis.

Specifically, clustering analysis represents a broad group of procedures that can be used to classify objects according to similar quantitative or qualitative properties (Massart & Kaufman, 1983). Beginning with the psychological work of Zubin (1938) and Cattell (1943), clustering methods have enjoyed a wide range of uses across psychology and other disciplines. One of these disciplines is the field of analytical chemistry, where researchers were able to demonstrate the power of both linear and nonlinear clustering methods to classify chemical compounds according to certain properties (Massart & Kaufman, 1983). This work, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, is a seminal work within the clustering field, and is often cited as a paramount resource for those interested in understanding clustering analysis.

Generally speaking, all clustering methods follow three specific steps. First, the items to be classified are defined or represented according to certain characteristics. As previously noted, these characteristics may either be qualitative or quantitative in nature. From a statistical perspective, this often involves representing the data in a traditional matrix of M rows by N columns (or M items identified by N variables). Accordingly then, the object of clustering would be to classify the M items according to their characteristics as defined by values on the N variables. Note however that the opposite approach may also be taken, where the M variables may be classified according to their values on the N items.

After the items are defined according to their certain properties, the next step is to define a measure of similarity representing the closeness of the items. Because the goal of clustering analysis is to classify items according to their closeness, this measure of similarity is extremely important in defining which items will be grouped. Although multiple measures of similarity may be used, the most common methods include correlation matrices, Euclidean distances, and Minkowski Distances.

Finally, after the objects are defined, and the measure of similarity is specified, the final step is to classify them according to one of a particular number of algorithms. These algorithms represent the multiple forms of clustering techniques available, as well as help researchers understand the complications in defining clustering analysis as a single method of analysis. Within the current dissertation, these methods include two of the better-known techniques, k-means clustering (Sarle, 1982), and Ward's method (Ward, 1963).

Describing Clusters

Although there are many different procedures for clustering, all methods share five common concepts: density, variance, dimension, shape and separation. Density is defined as group of data points that congregate around a given point. Because the goal of clustering analysis is to create well-defined groups of items, the ideal situation is to identify groups of items that are close in proximity to one another, but far apart from other items. When this is the case, the clusters should form in tight, identifiable groups, where intra-group distance is minimized, and inter-group distance is maximized. One of the methods for measuring cluster density is Dunn's index (Dunn, 1974), where the ratio of maximal intra-cluster distance to minimal inter-cluster distance is calculated. The equation for Dunn's Index is presented in Equation 1 below.

$$Dunn = \min_{1 \le i \le n} \left\{ \min_{1 \le j \le n, i \ne j} \left\{ \frac{d(i,j)}{\max_{1 \le k \le n} d'(k)} \right\} \right\} \qquad \textbf{Equation 1}$$

Variance is defined as the amount of dispersion from the center within a given cluster. Because clusters are technically hypothetical groups of objects created by the researcher, they do not possess a true "center". However, multiple definitions may be used to identify this point, with the most common being referred to as the centroid, where the centroid represents the mean vector of the items within the cluster. Minimizing the variance within groups is often a primary goal within the different methods of clustering analysis, including both Ward's Method and K-means Analysis. Note also that variance is also tied closely to the idea of separation defined later. More specifically, well-separated clusters often demonstrate low intra-cluster variance (a term often referred to as compactness).

Dimension is defined as the fewest number of coordinates required to identify a given point within a particular space. Because many clustering analyses represent data in terms of M*N matrices (again M items by N variables), the M objects are represented by an N number of variable vectors. The number of vectors then represents the dimensionality of the cluster, because each item requires N coordinates to place it within the particular vector space. Because the number of vectors often exceeds three, representing the data graphically is often impossible using standard graphing techniques. As such, many techniques for dimensionality reduction, including principal component analysis, factor analysis, and even cluster analysis itself are often employed.

Within clustering analysis, shape is defined as the arrangement of the points within a given space. Because of this, describing the shape of a cluster is similar to the describing the shape of any group of objects. And clusters may be defined as spherical, oval-shaped, linear, non-linear, or any other number of possibilities. However, because centroid methods are often used for measuring distance, andmany datasets are high dimensional in nature, many methods prefer to create spherical clusters.

Finally, separation is defined as the amount of overlap or distance between the clusters. Because many fit indices, including Dunn's index (Equation 1) take into account the distance between clusters, cluster separation is often used as an element in defining good model fit. When defining separation, a number of different methods may be used. These methods may include the pairwise distances between cluster centroids, the pairwise minimal distance between objects in different clusters, or any number of other measures (Liu, Li, Xiong, Gao & Wu, 2010). The use of separation in defining well-fit clusters is explained further in the following section on fit indices for cluster analysis.

Ward's Method

The current dissertation applies two specific forms of clustering analysis. The first form of analysis is referred to as Ward's method. Outlined in Ward (1963), Ward's method (or Ward's minimum variance method) is a clustering analysis method designed to minimize the variance between clusters according to an objective function specified by the investigator. Although this objective function may take any function specified by the researcher, sum of squares is often defined as the variance measure to be minimized, due to its wide use within the literature. Equation 2 presents the formula for the sum of squares reduction. ESS is defined as the error sum of squares, and X_{ijk} is defined as variable k in observation j in cluster i.

$$ESS = \sum_i \sum_j \sum_k \left| X_{ijk} - \overline{X}_{i.k} \right|^2$$ **Equation 2**

Writing out Equation 2 in words, it can be said that the error sum of squares is equal to the sum of the squared distance between the observations for variable k in cluster i between the individual scores and the mean score. Note that this definition is similar to the one applied by Fisher in defining the sum of squares within for the traditional ANOVA formula.

Continuing with the ANOVA analogy, Ward's method also employs a similar formula to calculate total sum of squares (TSS). This formula is presented in Equation 3 below, where once again X_{ijk} is defined as variable k in observation j in cluster i. In this situation however, the variance is calculated as the squared distance between the individual observations and the overall grand variable mean.

$$TSS = \sum_i \sum_j \sum_k \left| X_{ijk} - \overline{X}_{..k} \right|^2$$ **Equation 3**

Finally, Ward's method then interprets the clustering efficiency as the reduction in total proportion of variance obtained by clustering the two groups. This method is again similar to the proportion of variance reduced within an ANOVA and can be written as the amount of variance explained by the cluster (r^2) or:

$$r^2 = \frac{TSS-ESS}{TSS}$$ **Equation 4**

Hierarchical Agglomerative Methods

Within the Clustering Analysis literature, Ward's method is referred to as a specific form of hierarchical agglomerative method (El-Hamdouchi & Willett, 1987). The method is referred to as hierarchical because cases can be subsumed within other clusters. Additionally, the method is referred to as agglomerative because it seeks to join clusters in a bottom-up fashion, by starting with all items in unique clusters and then joining them based on a defined similarity. To join observations, Ward's method clusters items by beginning with n clusters, or every item in its own cluster. Next, the two items with the nearest similarity are joined, creating two clusters—one of size n-1 and one of size two.

Importantly, Ward's method defines similarity as the increase in sum of squares when two clusters are joined. Returning to the ANOVA comparison employed in the previous section, Ward's method defines the distance between two clusters as the change in sum of squares (Δ) when the two clusters are joined. This change is often referred to as the merging cost of combining the clusters.

For example, when joining clusters Y and Z, Ward's method calculates the distance, Δ(Y,Z) as:

$$\Delta(Y,Z) = \sum_{i \in Y \cup Z} \left\| \vec{x}_i - \vec{m}_{Y \cup Z} \right\|^2 - \sum_{i \in Y} \left\| \vec{x}_i - \vec{m}_Y \right\|^2 - \sum_{i \in Z} \left\| \vec{x}_i - \vec{m}_Z \right\|^2 \qquad \textbf{Equation 5}$$

where $\vec{x}_i$ represents the individual score vector for person i, $\vec{m}_{y \cup z}$ is the mean vector for the union of the Y and Z clusters, and $\vec{m}_z$ is the mean vector for cluster z. Because Ward's method begins with all data points being in their own cluster, the original Sum of Squares will originate at zero and then increase as clusters are joined.

The relative efficiency of Ward's method relative to other methods in recapturing the correct structure was investigated in Kuiper and Fisher (1975) using Rand's index to calculate proportion of correct clusters. Rand's index is defined as

$$\text{Rand} = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}} \qquad \textbf{Equation 6}$$

where N_{00} represents the number of pairs of items that were correctly placed in different clusters, N_{01} represents the number of pairs of items were not placed in the same cluster but should have been, N_{10} represents the number of pairs of items that should not have been placed in the same cluster but were, and N_{11} represents the number of pair of items that were correctly placed in the same cluster[2].

The results of Kuiper and Fisher showed that Ward's method worked especially well for clusters with equal sample sizes, particularly when the number of clusters increased. These results highlight one of the advantages of Ward's method, with other advantages including its ability to recreate structure when cluster sizes contain a fairly similar number of items, and the clusters form a spherical shape. Furthermore, because Ward's method is similar to ANOVA in methodology, it makes many assumptions similar to ANOVA. These assumptions include a multivariate normal mixture of items, equal spherical covariance matrices for all clusters, and equal sampling probabilities for all clusters. This familiarity with assumptions often makes Ward's method a preferred method for researchers.

Representing The Results of Ward's Method

Because the results of hierarchical methods produce clusters that originate as single items, the final product can often be viewed as a tree diagram (or dendrogram) demonstrating how and when the items were joined. A common method for demonstrating clustering analysis results, dendrograms possess many

---

[2] Because it is then a proportion of correct classification of pairs, Rand's index effectively reads similar to the concordance rate within logistic regression analyses.

positive qualities. First, dendrograms show the point at which items were linked together (showing the order in which the items were clustered). Second, many dendrograms will produce axes demonstrating the difference (or cost of merger) in joining two clusters. Third, dendrograms are very easy to understand, and may be used to represent the clustering process in a more user-friendly, graphical manner.

Shortcomings of Ward's Method

Although Ward's method enjoys great popularity within the research literature, it also demonstrates certain potential shortcomings. First, because Ward's method does not specify the number of clusters to be created, it is important to remember that it will continue to join n cases until the n-1 step, when all cases are joined into a single large cluster. As such, it is up to the researcher to determine the appropriate number of clusters post-hoc, often by examining the dendrogram or the merging cost of combining two clusters. Fortunately, there do exist resources suggesting the appropriate number of clusters, including Milligan and Cooper (1985). Secondly, because Ward's method primarily looks for spherical clusters, it may not be as effective as other methods in recovering structure when the actual clusters are not spherical in shape. More specifically, because Ward's method is reliant upon mean vectors to calculate the merging costs, it is susceptible to outliers (Milligan, 1980).

K-means Clustering

Along with Ward's method, a second form of clustering analysis that is employed within the current dissertation is k-means clustering. First referred to as k-means in MacQueen (1967), the goal of this form of analysis is to group n observations into k clusters where the clusters are defined by fixed centroids identified prior to beginning the analysis. After the items are grouped according to the closest centroid, the position of each centroid is once again re-calculated and the items are again re-positioned. This process continues until the positioning of the centroids no longer changes between iterations.

Because k-means clustering is a bottom-up method of clustering (unlike Ward's method, which is considered a top-down method), it is referred to as a partitioning method. Specifically, k-means seeks to create clusters that maximize inter-cluster distance, while minimizing intra-cluster distance. If we define the cluster centroid as c_j and the individual data point as x_i^j then we can say that k-means clustering seeks to satisfy the objective function D where D is defined as:

$$D = \underset{k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{i=1}^{n} \left\| X_i^{(j)} - C_j \right\|^2 \qquad \textbf{Equation 7}$$

To accomplish the task of minimizing the distances between items and the cluster means, k-means clustering employs one of a number of iterative algorithms. The most popular of these is referred to as Lloyd's algorithm (named after Stuart P. Lloyd), or alternatively as Voronoi iteration or k-means algorithm (MacKay, 2003). According to the k-means algorithm, the clustering of observations to groups alternates between an assignment step and an update step.

In the assignment step, items are placed into clusters according to the mean location that best minimizes the intra-cluster variance between cluster mean and item location. Because the intra-item variance is minimized, this can be thought of as minimizing the distance between the items and the center of the cluster (Equation

7)[3]. Once items are assigned to clusters, the centroids are re-defined again to minimize the intra-item variance.

Because of its methodology, the k-means algorithm guarantees that certain properties must hold true. First, it is guaranteed that there will always be at least one item per cluster. Second, it is guaranteed that there will always be k number of clusters, where k is the pre-set number of centroids defined prior to the analysis. Third, it is guaranteed that the clusters will not overlap. Fourth, it is guaranteed that items within the clusters will be closer to their own cluster than to any other cluster.

Although k-means clustering is a widely popular method, it does contain certain drawbacks. Primarily, because the number of clusters must be defined a-priori, it is important that the researcher correctly specify the number of clusters within the data. Secondly, because of the distance locations used in defining the clusters, it is important that the centroids themselves be located in the correct positions at the starting point. Third, it is almost guaranteed that the k-means clusters will converge to a local minimum unless multiple starting points and means are used (Peña, Lozano & Larranaga, 1999).

Combining the Two Methods

Because K means relies on the number of clusters specified a-priori, it is often helpful to run Ward's method first to obtain an approximation of the number of existing clusters. Specifically, researchers are encouraged to examine the dendrogram and merging costs of the cluster analysis produced by Ward's method,

---

[3] Note: The current distance metric is being defined as standard Euclidean metric. However, other metrics are available for classifying distance including Manhattan Metric where distance is defined as ($d_1 = (p,q) = \|p-1\|_1 = \sum_{i=1}^{n} |pi - qi|$).

and then use this number of centroids in the k-means method. As with other statistical analyses, by combining the results of both methods, researchers are able potentially to verify and replicate their findings. Furthermore, in the event that the two analyses yield strongly different results (by potentially grouping items in a significantly different manner), then a third method of clustering or a re-analysis of the data may be necessary.

Multi-Level Modeling

The use of hierarchical modeling has previously been applied towards understanding the influence of institutional and other higher-level organizations on student performance (Rocconi, 2013; Raudenbush & Bryk, 2012; Tabachnick & Fiddell, 2012). The influence of school level characteristics, including SES and sector (catholic vs. public) have been used to demonstrate the methods of multi-level modeling in Singer (1998), who was able to demonstrate the influence of socio-economic status (SES) on math achievement at both the school and student level.

The primary advantage of multi-level modeling is its ability to model effects at both higher and lower levels of data. These levels often present themselves in the form of naturally occurring "nesting", such as students nested within schools, or cities nested within states. Because these naturally occurring hierarchies tend to have items that are likely to be correlated, it is not appropriate to treat them through standard OLS regression techniques (which assume independence of errors), random sampling and random assignment.

Multi-level modeling works by creating equations at each level of interest (Singer, 1998). In the current dissertation, this involves creating an equation at the second (school) level, and at the first (student) level. The following equations present the basic theory for creating equations at each level[4]. Beginning with the school level if we are interested in the influence of sector (public/private) on the probability of retention, we can write the mean predicted probability as a combination of a grand mean predictor (Υ_{00}), the unique influence of the sector on predicted probability (Υ_{01}), a unique error associated with each individual school (μ_{0j}) and a unique error associated with each individual student (r_{ij}).

$$\Upsilon_{ij}=\Upsilon_{00}+\Upsilon_{01}(\text{Sector})+\mu_{0j}+r_{ij} \qquad \textbf{Equation 8}$$

Similar modeling can then occur at the level below (the student level) as well. For example, if we were interested in modeling the effects of high school GPA on first year academic performance for student i, in school j, then we could model this as:

$$\Upsilon_{ij}=B_{0j}+B_{1j}(\text{HSGPA})_{ij}+r_{ij} \qquad \textbf{Equation 9}$$

where $B_{0j} = \Upsilon_{00} + \mu_{0j}$ and $B_{1j}=\Upsilon_{10} + \mu_{1j}$. Combining the two equations allows for modeling the influence of both levels of variables, as well as unique errors at both the school (u_{0j}) and student (r_{ij}) level.

Additionally, by modeling for both intercepts and slopes, it is then possible to measure the variance and covariance between the two. Ideally, the goal of multi-level modeling is to explain this variance in intercepts (τ_{00}) and the variance in

[4] Although the current analysis involves predicting a dichotomous (here/not here) variable, the basic principal still applies.

slopes (τ_{11}), as well as examine how the two covary (τ_{01}). These variances are often represented in a tau matrix, where

$$T=\begin{bmatrix}\tau 00 \ \tau 01\\ \tau 10 \ \tau 11\end{bmatrix}$$

Summary

The current chapter has outlined the background theory for the procedures applied in the current dissertation. The chapter opened with a brief description of clustering analysis theory before giving a more detailed explanation of two specific clustering methods: Ward's Method and k-means clustering. Next, the chapter moved into a brief discussion of the concepts and formulas applied in multi-level modeling. Taken together, these procedures form the primary analytical techniques applied in the current dissertation.

CHAPTER III: VARIABLES IN THE CURRENT ANALYSIS

Overview

The current dissertation applies cluster analysis and hierarchical logistic regression to estimate the impact of school and student level variables on a person's probability of persisting to the second year of college. Variables were collected through a combination of university academic records and questionnaires administered the summer before students entered the university. This chapter provides greater detail on the influence of the collected variables on a student's probability of retention as reported in the previous literature. The chapter begins with the traditional academic variables of standardized test scores and high school GPA and then continues on to discuss the psychosocial variables.

Student-Level Variables

High School GPA and Standardized Test Scores

Perhaps no other variable has been more researched within the student retention field than high school GPA. When determining which applicants to admit, college admissions committees nearly always consider a student's high school GPA as a primary way of evaluating how the student will perform in college. And while high school GPA is certainly not a perfect predictor of future performance, previous results do indicate that a student's grade point average in college remains strongly tied to his/her grades in high school (Geiser & Studley, 2004; Kobrin et al, 2008; Sawyer, 2010). High school grades are predictive of college graduation rates as well (Zwick & Sklar, 2005).

The relationship between high school GPA and college grades has been examined across a number of different arenas, including the University of California school system, where Geiser and Santelices examined the influence of high school GPA on both first year GPA, as well as four year grades (Geiser & Santelices, 2007). In their 2007 article, the authors were able to show that high school GPA consistently remained the strongest single predictor of college performance across over 80,000 students within the UC system.

In addition to high school GPA, a student's standardized test scores have also been shown to correlate with a number of academic behaviors. Although research has shown that high school GPA appears to be more strongly correlated with first-year college performance than standardized test scores (Kobrin et al., 2008), there is also evidence that the influence of grade inflation in high schools may be diminishing this effect (Geisinger, 2009; Woodruff & Ziomek, 2004; Camara at al, 2003; Godfrey, 2011). Because of this, standardized test scores may become more influential in predicting student academic performance, particularly for more highly selective institutions (Mattern et al., 2008).

There have been multiple reports demonstrating the significant relationship between standardized test scores and college persistence (Reason, 2003; Conner, Daugherty & Gilmore. 2012-2013). Compared to their peers, students with higher standardized test scores are more likely to obtain a degree and move from community to four-year colleges (Porchea, Allen, Robbins & Phelps, 2010), are more likely to earn higher grades, and are more likely to master the curriculum presented in higher education (Espensade & Chung, 2011).

Taken together, the results from previous research investigating the high school GPA and standardized test scores seems to indicate that the two do significantly predict performance in college. Although results do not present a consistent pattern of prediction (there are certain situations where high school GPA performs better, and certain situations where standardized test scores perform better), given the preponderance of studies demonstrating a strong and significant relationship between previous and future academic performance, it is clear that students who excel academically while in high school are more likely to excel while in college.

However, experience also demonstrates that there are more contributing factors to collegiate achievement than simply previous academic success. Unfortunately, it is all too often the case that a highly capable student simply does not perform well in college. And although a good number of dropouts are entering college with lower standardized test grades and high school GPA's, there still exist a number of perfectly qualified students who, for whatever reason, still opt to leave college early. Given the invaluable benefits associated with earning a college degree (a theme that will be repeated throughout the dissertation), it is surprising that so many students would voluntarily leave.

This phenomenon has lead many researchers (including Tinto) to formulate college dropout not as a single action but rather a process involving a number of moving parts. Within these parts are a number of psychosocial variables, characteristics (including social and institutional commitment) that Tinto hypothesized might better explain withdrawal, particularly voluntary withdrawal where the student does not fail out. The following presents a list of psychosocial variables used to predict retention in the current dissertation. Some variables (such as self-efficacy discussed in the following paragraph) are well known psychological phenomena, established in the literature to influence a variety of social behaviors. Other variables (such as academic engagement) are only recently becoming more popular. Regardless, each has been shown to influence the student persistence decision in some form or another.

Self-Efficacy

In his 1977 paper, Bandura defines self-efficacy as one's belief in one's ability to succeed in specific situations (Bandura, 1977; 1994). From an educational perspective, self-efficacy may be specified as one's ability to succeed (both academically and socially) while in college. The influence of self-efficacy in a learning environment has been examined in multiple studies (Jernigan, 2004; Bandura & Zimmerman, 1992), including how self-efficacy influences fields such as future career options (Lent, Brown, & Larkin, 1986) and number of college credits earned (Zajacova, Lynch, & Espenshade, 2005).

According to Bandura, having a high sense of self-efficacy influences performance by generating "intrinsic interest and deep engrossment in activities" (Bandura, 1976). As such, students who see themselves as being highly capable at successfully performing the required academic and social activities necessary for success in college may be more likely to display and cultivate a deeper interest within these areas. Note that this notion of fostering active engagement in the social and academic arenas of the institution is similar to Tinto's idea of social and academic integration (Tinto 1975; 1982; 1993).

Self-efficacy has been linked to a number of positive outcomes, including institutional satisfaction and goal progress (Feldt, 2012), persistence and academic achievement (Usher & Pajares, 2008; Choi, 2005), and higher course and professor evaluations (Riconscente & Seli, 2012). In addition, higher self-efficacy has been associated with greater academic performance in first-generation students (Vuong, Brown-Selty & Tracz, 2010), increased motivation (Schunk, 1991), and greater college adjustment (Ramos-Sanchez & Nichols, 2007).

The concept of academic self-efficacy was discussed heavily in Solberg, O'Brien, Villarreal, Kennel and Davis (1993). According to the authors, academic self-efficacy is defined as a "student's degree of confidence in performing various college related tasks to produce a desired outcome, such as passing an examination". Because of this, students with high academic self-efficacy may be more likely to approach challenges with a positive outlook, rather than a degree of academic fear and anxiety.

The influence of self-efficacy has been shown to influence persistence rates across a number of different institution types. Brewer and Yucedag-Ozcan (2012-2013) were able to demonstrate the influence of self-efficacy in improving persistence rates in a large online course setting. The authors used an online orientation course where students were able to discuss and plan out their methods for improving time management, engaging in new learning styles, and succeeding in college. Results showed that students participating in the online orientation course not only improved their self-efficacy, but in doing so, improved their persistence rates as well.

Further benefits of self-efficacy were evidenced in Coffman and Gilligan (2002-2003) who were able to connect higher levels of self-efficacy to lower levels of stress as well as to overall life-satisfaction. Because of this, it seems that self-efficacy may play an important role in helping students adjust to the stressors of their newly found college life. Specifically, because these students are having to overcome multiple obstacles in their social and academic life arenas, their perceived ability to overcome these obstacles may provide a buffer and source of comfort in their adjustment.

Academic Engagement

Academic engagement is defined as the amount of conscious effort a student exhibits towards mastering the academic requirements for success while in high school. Academic engagement has been shown to positively influence student behavior across a number of previous studies (Herrmann, 2013; Floyd, Harrington, & Santiago, 2009). According to Herrmann (2013), academic engagement encourages positive student behavior because it is closely tied to the concept of cooperative learning, a form of learning where students share similar goals with their peers and the ultimate results of the group depend on the goals of the individuals (Johnson & Johnson, 1989; Onwuegbuzie & DaRos-Voseles, 2001; Johnson & Johnson, 2009). The authors also relate academic (or student) engagement to a form of active learning, a concept that includes a student's motivation and strategies for learning.

According to Biggs and Tang (2011), students who approach learning with a motivation to master and understand the material (as opposed to simply memorizing the material) are engaging in the deep approach to learning (Entwistle & McCune, 2004). Deep approaches to learning have been associated with a number of activities, including in-class participation (Rocca, 2010; Weaver & Qi, 2005). Student engagement has also been shown to demonstrate a number of positive effects, including grades and persistence (Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008). Using a large database of over 6000 students and across 18 institutions, the Kuh and colleagues were able to show where elevated levels of academic engagement improved both college GPA, as well as likelihood of second-year persistence.

In arguing for the importance of fit in developing engagement, self-determination researchers hypothesize that the ideal situation for academic engagement is when the school encourages "competence, autonomy, and relatedness" (Deci & Ryan, 2000; Wang & Eccles, 2013). More specifically, they argue that healthy engagement is derived from the quality of a student's interactions with the learning activities and academic tasks (Wang & Eccles, 2013; Eccles, 2004). Within this framework, engagement is broken up into three separate areas, behavioral, emotional, and cognitive engagement.

Researchers define behavioral engagement as the actions and practices a student directs at school activities and learning (Connell, 1990). Examples of behavioral engagement may include doing homework in a timely manner, studying for tests outside of class, and completing the assigned reading. In this way, behavioral engagement is somewhat analogous to positive habit cultivation, or development of a form of academic conscientiousness.

Although it is important to cultivate healthy and conscientious academic behaviors, it is also important to develop that students develop a healthy attitude towards learning as well. Researchers define this attitude towards school and learning as a student's emotional engagement (Voelkl, 1997). The influence of emotional engagement has been studied across a wide variety of situations, with results showing that student's attitudes towards school influence their performance in mathematics (Dettmers et al., 2011), as well as their approach to developing strategies for learning (Pekrun, et al., 2011).

Finally, researchers define cognitive engagement as the amount of mental investment a student puts forth towards learning, as well as the willingness a student exerts to master difficult and new concepts (Como & Mansinach, 1983). The construct of cognitive engagement derives from the theory of processing levels (Anderson & Reder, 1979), and elaborated processing (Craik & Lockhart, 1972). The influence of cognitive engagement has been researched in a number of previous studies, including Sedaghat, Abedin, Hejazi & Hassanabadi (2011) who examined the relationship between cognitive engagement and academic achievement, Suarez-Oroz, Pimentel & Martin (2009) who examined the relationship between academic engagement and achievement in immigrant students, and Pintrich & De Groot (1990), who examined the relationship between self-regulation, motivation, and cognitive engagement.

Taken together, the constructs of behavioral, emotional, and cognitive engagement comprise the three-pillar approach of academic engagement. As can be seen, students must take into account many different factors when becoming academically engaged, further supporting the argument that student retention process is a complicated and dynamic process. Along with becoming academically engaged, students must also balance a number of other decisions when deciding to remain in college. Another one of these important decisions is whether the costs of attending college outweigh the benefits. This construct represents a brief introduction into financial concerns, the next psychosocial variable discussed.

Financial Concerns

Financial concerns are defined as a student's level of worry about being able to pay for and afford college. Because paying for college is often seen as an overwhelming obstacle for incoming students, a great deal of prior literature has been dedicated to understanding better the relationship between money and college behaviors (St. John, Paulsen & Carter, 2005; Braunstein, McGrath & Pescatrice, 2000-2001). Moreover, due to the rising price of college over the years (estimates of approximately 4.4% rise in tuition for in-state tuition[5]), more researchers are beginning to investigate both the cause of the increase (Archibald & Feldman, 2008) as well as the influence of such costs on student welfare (Hornak, Farrell, & Jackson, 2010).

---

[5] Source: College Board. *Trends in college pricing. Trends in Higher Education Series.* College Board. Retrieved Jan 16, 2014, from http://www.collegeboard.com/prod_downloads/press/cost06/trends_college_pricing_06.pdf

Previous research has shown that students with higher financial difficulties are more likely to exhibit a number of symptoms associated with higher probabilities of academic withdrawal. Compared to their peers, students with higher levels of financial needs are more likely to believe that college is unnecessary and too expensive (Tierney & Venegas 2010; 2007), with students becoming less likely to persist as they incur greater financial debt (Coffer & Somers, 1998). Because of this, financial considerations are often cited as one of the primary reasons for withdrawal.

Previous research seems to confirm these statements. Compared to their non-working peers, students having to work to afford college may exhibit a number of potentially negative behaviors, including higher levels of anxiety (Mounsey, Vandehey, & Diekhoff, 2013), fewer hours spent socializing (Lang 2012), more hours spent having to work outside of class (Pascarella & Terenzini, 2005), and more hours spent away from campus (Choy & Carroll, 2003). Students with higher financial concerns have also been shown to engage in behaviors such as taking fewer credit hours, as well as reporting higher levels of psychological stress due to financial worry (Robb, Moody & Abdel-Ghany, 2011-2012).

The influence of financial concerns has not only been restricted to behavior in the classroom. In their college choice-nexus theory, Paulsen & St. John (1997; 2002) hypothesize that students take financial costs into consideration when choosing a college. Specifically, students will weigh the overall costs of tuition against the benefits of obtaining a degree from the particular institution. Because of this, many students from lower class backgrounds may be discouraged from even attempting to apply to upper-level colleges due to the cost associated with attendance.

Because financial concerns are such an important part of influencing persistence decisions, a great deal of research has been dedicated towards understanding how different types of financial aid may influence a student's persistence decision. Studies have examined the influence of programs such as merit-based scholarship (Schuh, 1999-2000; Stampen & Cabrera, 1988), college loans (Hochstein & Butler, 1983), and work-study programs (Desjardins, Ahlburg, & McCall, 1999).

In a 2002-2003 paper by Ishitani and DesJardins, the influence of financial aid timing and allotment on student attrition was modeled using hazard modeling. The authors not only examined the influence of financial aid timing on attrition, but also the influence of such external factors as parental educational attainment and student educational aspirations. As with previous findings, the results continued to show that students who received financial aid were less likely to drop out.

Although the present findings do suggest that financial aid strongly influences a student's retention decision, the current dissertation focuses more specifically on financial concerns, rather than financial aid. The distinction between the two is very important. Specifically, while financial aid measures the amount of money a student has to pay for college, financial concerns measures the amount of worry a student has about being able to pay for college. Because of this, the influence of financial concerns may be more directed at the cost-benefit analysis of paying for college, rather than simply the amount of aid received.

Institutional Commitment

Another factor that may strongly influence the decision to persist is the concept of institutional commitment, specifically measuring *the amount of dedication a student feels to a given institution*. Originally proposed by Tinto (1975), institutional commitment refers to a student's level of dedication to the particular school he/she is currently attending. If goal commitment refers to a student's level of dedication towards graduation in general, institutional commitment refers to a student's dedication to graduating from a particular institution. This level of commitment may be demonstrated in a number of ways, including participating in college fairs and attending welcome week orientations (Goenner, Harris & Pauls, 2013), engaging in transition experiences (Birnie-Lefcovitch, 2000), attending sporting events, or joining social clubs.

Results have shown that students with higher institutional commitment are less likely to withdraw from college than their peers (Wang & Kennedy-Phillips. 2013; Davidson, Beck & Milligan, 2009). This research was supported by the findings of Campbell and Mislevy (2012-2013), who showed that, compared to their continuously enrolled peers, stop-out students (those students who discontinued enrollment at some point during their college careers) were more likely to possess negative attitudes towards their institutions.

The influence of institutional commitment was examined in Braxton, Sullivan and Johnson's (1997) meta-analysis testing the propositions of Tinto's interactionist theory. According to the authors, a student's initial institutional commitment is determined by his/her incoming attributes. This initial level of institutional commitment then interacts with social integration to further influence the student's subsequent levels of commitment. Subsequent levels of institutional commitment then ultimately influence the student's decision whether to persist at the institution, transfer, or withdraw from higher education.

The importance of institutional commitment may be evidenced in the overwhelming number of students that fail to graduate from their initial institution. According to Adelman (2004), approximately one out of five students who begin college at a four-year institution will not earn a degree from that institution. Examining the causes of institutional commitment, Hermanowicz (2006-2007) asked leavers at a more selective college why they were choosing to voluntarily leave. Conducting interviews over the phone, Hermanowicz was able to uncover a few recurring themes regarding college attrition. First, he notes that the causes of attrition are usually multi-faceted, and that they typically cannot be resolved by a single solution. Examples include students citing financial difficulties as well as not getting along with peers.

Interestingly, while financial considerations were certainly prevalent amongst the reasons for leaving, many students cited cultural and philosophical differences with the institution's prevailing attitudes. Students cited anecdotal evidence ranging from perceived socialist teachings emphasized in classrooms and dorms, to frustration with teachers' abilities to speak English, to locations of the dorms on campus. Other reasons for leaving included transitions issues related to coming from small towns, homesickness, dissatisfaction with athletic programs and problems with roommates. Taken together, the results of Hermanowicz's research seem to re-affirm the importance of institutional commitment in determining whether a student will remain at a given university.

School-Level Variables

High School Class Size / Student-Teacher Ratio

Along with student-level variables, there also exist school-level variables that have been shown to influence higher education outcomes (Pleitz, Terry, Campbell & Fife, 2012). Although the potentially politically sensitive nature of examining the influence of school level variables on higher education academic performance makes research findings slightly more difficult to obtain than those at the student level, results have shown certain school characteristics to reliably predict performance across a number of observations.

In a 2005 literature review on the influence of school size, Slate and Jones (2005) argued that the size of a high school actually had a curvilinear influence on performance. Specifically, while Slate and Jones acknowledge the potential positives of having larger high schools (increased student diversity, greater curricular offerings), they also argue that larger school sizes do not necessarily equate to greater student achievement. Indeed, Harnisch (1987) found a correlation of only .13 between school size and achievement, suggesting that the two are largely unrelated. This result was again suggested with the findings of Cotton (1996) who reviewed over 31 papers examining the relationship between school size and achievement, only to come to inconclusive results suggested advantages to both large and small schools.

Although naturally related, the influence of student-teacher ratio on academic performance has been less researched than the influence of class size. Indeed, while there are a number of research articles investigating how a public or private school may influence a student's grades at the next level, the study of student-teacher ratio remains fairly limited. Fortunately, longitudinal research examining high school teacher/student ratios over time does seem to suggest a few trends.

First, it is clear that the average student-teacher ratio in America is on a pronounced decline (Ehrenberg, Brewer, Gamoran, & Willms, 2001). Statistics show that the average student-teacher ratio has declined from 19:1 to 14:1 between 1969 and 1997, suggesting that students are more likely to receive individual time with teachers than in previous years (Campbell, Hombo, & Mazzeo, 2000). Interestingly, although the class sizes do appear to be shrinking, longitudinal evidence from the National Assessment of Educational Progress (NAEP) suggests that these declines are not yielding any significant improvement in student knowledge.

The influence of student-teacher ratio on classroom performance was further examined in a 1999 meta-analysis by Hanushek. In the report the author analyzed findings from over 270 schools, investigating whether smaller class sizes truly improved student performance. Results were once again inconclusive, with only 15% of findings showing statistically significant improvement due to smaller classrooms. Given that 13% of reports showed statistically negative results of performance in smaller class sizes, the results seem to suggest that either the effect is highly nuanced, or that additional research is needed (Hanushek, 1999).

District Spending Per Student/ Free Lunch Eligibility

With the passing of the No Child Left Behind[6] (NCLB) act of 2001, increasing attention is being paid to district spending per student. The newest authorization of the Elementary and Secondary Education act of 1965, NCLB requires schools that receive federal funding to demonstrate adequate increases in academic performance. As can be imagined, the requirements for demonstrated improvement to receive federal funding have been met with a number of controversies. While legislators argue that the required standards set an initiative for improvement, detractors believe that cutting funding to already struggling schools may only harm them to a greater degree.

Lost in this debate is effective research on the actual influence of district spending per student. In a 2002 report on school finance, the Government Accountability Office (GAO) found that higher spending per student tended to be directly related to higher staff wages, with districts spending more per student also paying their teachers a higher salary (The U.S. Government of Accountability Office [GOA] report no. GOA-03-234, 2002). Wrapped up in this dilemma is the correlation between higher teacher salaries and greater experience (having worked at the school longer), or greater training (receiving a Master's degree in education rather than simply a bachelor's degree).

[6] PL 107-110, the *No Child Left Behind Act of 2001*, enacted January 8, 2002

Because of these confounds, it is difficult to disentangle the influence of actual spending per student from other potential (and unseen) positive benefits associated with such spending. One of the primary arguments against the influence of district spending and performance comes from Hanushek (discussed previously in the student-teacher ratio section) who again showed inconsistent findings across a number of reports (Hanushek, 1986). According to Hanushek, the inconsistent findings indicated no strong or systematic relationship between school expenditures and student performance (p. 1162), a statement that many have taken to mean that there is no relationship between money and academic achievement.

One vocal opponent of this statement comes from the Albert Shanker Institute, where Baker (2012) argues that research proves three general tenants: a) aggregate measures of spending per student show that greater spending is associated with positive outcomes, b) academic and school resources funded by such spending positively influence student outcomes, and c) greater equality in spending across districts creates more positive student outcomes (Baker, 2012). In the report, Baker cites the works of Greenwald, Hedges and Laine (1996) who re-analyzed the previous findings of Hanushek (1986) and concluded that, amongst the statistically significant findings, results showed a majority of positive relationships between spending and achievement (Greenwald, Hedges, & Laine, 1996)[7].

While district spending per student may be a greater sign of school affluence, student eligibility for free lunch is a greater sign of district poverty. And,

[7] Tying into the argument of class size, Wenglinsky (1997) proposed that greater expenditures in district spending may cause greater academic performance by reducing class size.

just as greater district affluence seems to correlate with greater academic performance (albeit the actual effect is somewhat inconclusive), so too does district poverty seem to correlate with lesser academic performance. These results were demonstrated in a 2001 technical report by the Washington School Research Center, where authors found that discrepancies in income explained a greater amount of variance in performance than other variables, including ethnicity (Abbot & Joireman, 2001).

The role of poverty in academic performance was addressed in Burney and Beilke (2008), who once again found that poverty explained a substantial amount of variance in achievement. According to the authors, low income students suffer from a large number of potential detriments to success, including having parents that are less likely to have attended college (Lee & Burkham, 2002), being more likely to come from single parent families (Caldas & Bankston, 1999), and being more likely to attend high schools with less rigorous curriculum and fewer advanced placement courses (Martin, Karabel, & Vasquez, 2005). These findings also explain why students from lower income families have been shown to be significantly less likely to graduate than their more economically advantaged peers (College Board, 2005).

School Sector

As with previous variables, research examining the influence of school sector on academic performance has generated inconclusive results. Part of the problem with measuring the influence of public vs. private sector naturally occurs due to differences in the types of students attending each school. For example, although NAEP scores typically show higher performances for private school students (Braun, Jenkins & Grigg, 2006) previous research has shown that a majority of private school benefits may simply be attributed to public schools serving a greater number of economically disadvantaged children (Lubienski & Lubienski, 2006), rather than from deficiencies in the school itself. These results echo the continuing theme of economics potentially underlying a majority of differences in achievement (as discussed in the previous sections).

Additional complications arise from fundamental differences in the structure of private vs. public schools. For example, results have shown that, compared to public schools, private schools demonstrate a more equivalent student-teacher ratio (US Department of Education, 1999), have smaller class sizes and typically smaller enrollment (Alt & Peter, 2003). Although private schools do not receive state funding (and therefore technically receive no money from the district), the advantages of smaller class sizes have been previously discussed as an attribute of more affluent school districts, potentially equating private schools with the more wealthy public schools.

The debate over public and private schools became center focus of a 2007 issue of Time Magazine asking “Are private schools really better?” The article cites a study by the Center on Education Policy claiming that, after removing the influence of socioeconomic effects, the advantage of private schools becomes negligible, however the research suggests that the advantage of Jesuit or Catholic schools remains, even after conditioning for SES (Wenglinsky, 2007). This advantage of attending religious schools was replicated by Horowitz and Spector

(2005), although the authors argued that the effect dissipates throughout a student's high school tenure, nearly disappearing by junior and senior year.

Taken together, the comparisons between public or private high schools seems to present a mixed bag of results. Although results do seem to indicate that attending a religious school may present students with a positive advantage in college performance, given the wide variety of options in both public and private schools, it is likely that the influence is contingent upon a number of variables, rather than simply a school being private or public.

Summary

The current chapter has provided a brief background description of the variables applied in the current dissertation. Variables at the student-level include academic credentials such as high school GPA and standardized test scores, as well as psychosocial variables including institutional commitment, financial concerns, academic engagement, and self-efficacy. Variables at the school-level included high school class size, sector, student-teacher ratio, free lunch eligibility, and city size. The creation of these variables is discussed in the following chapter.

CHAPTER IV: SAMPLE AND METHODOLOGY

Overview

The current chapter describes the sampling methods and quantitative techniques applied to measure the relationship between the high school characteristics of incoming college freshmen and their impact on the probability of persistence to the second year in higher education. The following areas are discussed: participants, methods of data collection, and procedure of data analysis.

Participants

Data were collected from 4,407 incoming, first-time, full-time students at a large Midwestern University. Because all participants were first-time and full-time students, it is highly probable that the majority were between the ages of 18-20 years old. The average ACT score for the data set was a 25.88 (SD = 4.07) and the average high school GPA for the data set was a 3.59 (SD = .33). Approximately 85% of students were from public schools with the remaining 15% from private schools. All home-schooled students were removed from the analysis.

Creating the Data Set

The dataset was created in a series of steps. In the first step, data from two separate files were imported from excel into SAS using the PROC IMPORT feature. Data from the first file contained high school variable information, including student-teacher ratio, graduation rate, high school graduating class size, district spending per student, and high school sector (public/private). Data from the second file included student level information, including standardized test scores, high school GPA, and psychosocial variables including academic engagement and financial concerns.

Creating the high school variables of sector and city type involved using internet research. High schools and cities were entered into the Google search engine to determine the size of the city where the high school was located, as well as the type of city where the school was located. Information for graduation rates, student-teacher ratios, free lunch eligibility, and district spending per student were compiled through multiple educational websites. These sites included the domains www.publicschoolreview.com, and www.privateschoolreview.com, as well as state department of education websites.

Creating the psychosocial variables involved importing data collected from the 2012 incoming freshman class cohort at the University of Oklahoma. The data were collected through the New Student Survey (NSS), a questionnaire of approximately 108 items administered to incoming students during the summer before their freshman year. The questionnaire contains items measuring a student's attitudes and behaviors cultivated while in high school, as well as beliefs about what the college experience will be like.

After importing the two datasets, the files were merged by high school code and student ID creating a file that contained both student and school level information. The approximate n size for the sample was 2,898. Tables 1 through 5 present the descriptive statistics for the sample. Questions on the NSS were coded on the Likert scales presented above the tables, while demographic statistics were taken from city websites, Wikipedia sources, or government statistics.

Sample Descriptive Statistics

Table 1 presents the descriptive statistics for the two questions designed to measure the degree of difficulty a student anticipates adjusting to university life. The two specific areas of adjustment addressed include having enough money and having to combine a job with studies while in college. The questions were measured on a Likert scale from 1 to 5, and were coded in the following manner: Using the scale provided, please rate each of the following in terms of how difficult you think the adjustment may be during your first year at OU. 1) Very Easy 2) Easy 3) Neutral 4) Difficult 5) Very Difficult.

Table 1: Descriptive Statistics for Adjustment Variables

Question	N	Mean (SD)	Min	Max
Having enough money	2898	3.15 (1.09)	1.00	5.00
Combining a job with my studies	2891	3.17 (0.94)	1.00	5.00
Doing well academically	2856	2.54 (0.85)	1.00	5.00

Table 2 presents the descriptive statistics for variables associated with high school behaviors. These behaviors primarily address negative habits potentially cultivated while in high school, particularly with an emphasis on those habits that may result in lower academic engagement while in college. The variables were reported using a Likert scale ranging from 1 to 4, and were coded in the following manner: Using the scale provided, please indicate how often you did each of the following while in high school: 1) Very often 2) Frequently 3) Seldom 4) Almost Never.

Table 2: Descriptive Statistics for High School Behaviors

Question	N	Mean (SD)	Min	Max
Went to class without doing assigned reading	2893	3.37 (0.97)	1.00	5.00
Went to class without doing homework or assignments	2894	3.92 (0.84)	1.00	5.00
Waited until the last minute to do my assignments	2865	2.91 (0.97)	1.00	5.00
Waited until the last minute to study for exams	2884	3.02 (1.05)	1.00	5.00
Felt bored in class	2893	2.42 (0.87)	1.00	5.00
Felt overwhelmed by all I had to do	2890	3.07 (1.02)	1.00	5.00
Went late to class	2892	4.15 (0.89)	1.00	5.00
Skipped class	2897	4.61 (0.69)	1.00	5.00

Table 3 presents descriptive statistics for questions asking about a student's attitudes and level of agreement with certain items. This items focus both on high school behaviors, as well as incoming attitudes regarding expectations about what college life will be like. The items are scored on a 1 to 5 Likert scale and are presented as follows: Please indicate the extent to which you agree with disagree with each of the following items using the scale provided: 1) Strongly Agree 2) Agree 3) Neutral 4) Disagree 5) Strongly Disagree.

Table 3: Descriptive Statistics for Academic Engagement Variables

Question	N	Mean (SD)	Min	Max
While in high school, I was challenged to do my best academic work	2893	2.17 (0.96)	1.00	5.00
I rarely studied outside of class when in high school	2892	3.26 (1.14)	1.00	5.00
I need to work to afford to go to school	2881	3.07 (1.25)	1.00	5.00
On occasion, I have had doubts about my ability to succeed in life	2882	3.59 (1.17)	1.00	5.00
I am confident in my ability to succeed at OU	2892	1.64 (0.66)	1.00	5.00
I remain calm when facing difficult academic challenges	2893	2.42 (0.86)	1.00	5.00
I am confident I made the right choice when choosing to attend OU	2896	1.40 (0.61)	1.00	5.00
I feel like I worked harder than most students while in high school	2894	2.26 (0.99)	1.00	5.00
I am confused and undecided as to my future educational goals	2893	3.63 (1.03)	1.00	5.00
I have confidence in my academic abilities	2884	1.83 (0.64)	1.00	5.00
I expect to work hard at studying in college	2886	1.44 (0.54)	1.00	5.00
It is important to me to graduate from OU as opposed to another college or university	2887	1.79 (0.89)	1.00	5.00

Table 4 presents the descriptive statistics for the school and student-level academic characteristics. In addition to academic characteristics, Table 4 also presents demographic statistics for the schools, including the average city size, the breakdown of sectors, and the average student to teacher ratio.

Table 4: Descriptive Statistics for School and Student Variables

Item	N	Mean (SD)	Min	Max
City Size	3873	259114.32 (411244.85)	32.00	3858000.00
District Spending Per Student	3856	9680.00 (3202.20)	2335.00	75075.00
% Available for Free Lunch	3809	23.55 (19.18)	0.00	100.00
Graduation %	3842	91.36 (7.78)	26.00	100.00
Student-Teacher Ratio	3827	15.93 (5.85)	4.00	313.00
High School GPA	3882	3.59 (0.33)	1.70	4.00
ACT Score	3875	25.88 (4.07)	13.00	36.00
Sector				
Public	3303	85.30%		
Private	569	14.70%		
Family Members Attended OU				
Yes	962	33.16%		
No	1939	66.84%		

Table 5 presents descriptive characteristics for variables associated with the items that were important in a student's decision to attend OU. These decisions include financial, social, and academic considerations. The items were scored on a 1 to 4 Likert scale and were presented as follows: Please indicate how important each of the following was in your decision to attend OU, using the scale provided. 1) Extremely Important 2) Important 3) Relatively Important 4) Totally Unimportant.

Table 5: Descriptive Statistics for Importance Questions

Question	N	Mean (SD)	Min	Max
Financial Aid Received	2889	2.15 (0.97)	1.00	4.00
Cost of OU	2896	2.01 (0.81)	1.00	4.00
Was not accepted at my first choice	2878	3.58 (0.79)	1.00	4.00
Could not afford my first choice	2875	3.48 (0.89)	1.00	4.00

Table 6 presents assorted descriptive statistics. These questions ranged from questions asking about academic workload and GPA expectations to work ethic in high school and percentage of friends attending college. Questions about expectations for workload and GPA were coded in a Likert Scale ranging from 1 (significantly more difficult/better) to 5 (significantly easier/worse). The question about academic work experiences was coded from a 1 (I rarely had to work hard to receive good grades) to 5 (I had to work very hard all of the time to receive good grades). In addition to these two questions, the questionnaire also asked students about the amount of studying they did while in high school, as well as the amount of studying they expect to do while at OU. These variables were coded from 1 (0 hours per week) to 10 (more than 40 hours per week).

Table 6: Descriptive Statistics for Experiences and Expectations Variables

Question	N	Mean (SD)	Min	Max
Relative to high school, I expect the college-level academic work to be:	2894	1.61 (0.66)	1.00	5.00
Relative to high school, I expect my GPA in college to be:	2892	2.52 (0.88)	1.00	5.00
Which of the following best describes your academic work experiences while in high school:	2890	2.72 (1.12)	1.00	5.00
While in high school, the amount of time I spent studying outside of class was:	2893	3.35 (1.47)	1.00	10.00
While at OU, the amount of time I expect to spend studying outside of class is:	2896	5.32 (1.57)	1.00	10.00

Procedure

Data analysis occurred in four separate phases. First, the data were cleaned of outliers and merged into a master file that contained all the necessary variables.

Next, factor analysis was used to determine the structure and validity of the latent variables inherent within the New Student Survey questionnaire, and composite scores of specific factor items were used to create proxy-variables representing the latent variables. Finally, four separate models were used to predict probability of first-year retention within the GPA groups. The following section contains more detailed descriptions of the specific phases.

Phase 1

After the New Student Survey data and student demographic information were retrieved, the two files were merged by student ID and then labeled by cohort. After merging the files, student ID's were removed and each student was given a unique study ID. Following de-identification, any question either not related to student retention was eliminated, thereby reducing the number of questions from approximately 100 to 29. Once the variables were reduced, any variable names that had changed throughout the previous cohorts were re-labeled to a uniform name to ease in coding.

After re-labeling, all variables were examined for potential outliers and input errors (such as having a score entered that is not possible). Finally, to ease interpretability, certain variables were reverse coded such that higher scores indicated the higher prevalence of a particular attribute. For example, if a question asked a student to rank, from 1 to 7, how concerned they are about having enough financial resources, and 7 represented very worried, then the question was not reverse coded. However, if a question asked a student how worried they were about having to maintain a job and go to school at the same time, and a 7 now represented not very worried, then the question would be reverse coded. Once all data were re-coded, the cohorts were collapsed into a single aggregate file and then re-divided into new groups based on freshmen year GPA. Appendix 4 presents descriptive statistics and tables regarding the missing data and final data file.

Phase 2

After collecting data, the items were factor analyzed to determine the underlying factor structure of the items. Maximum likelihood was used as the estimation method and a Varimax rotation was applied to maximize interpretability under orthogonal rotation. A minimum eigenvalue of 1.00 was used to determine cut-off points for factors and squared multiple correlations were used to estimate prior communalities. A scree plot was also examined to determine where the eigenvector "bent".

After examining the factor analytic results, variables that did not significantly load onto any factor were removed. In addition, the expectation variables (I expect to work hard at studying for college, and Relative to high school, I expect my GPA/college-level academic work to be:) were not included within the composite variables because it was not clear how to interpret their findings. For example: it was not clear how having unrealistically high or low expectations would interact with the other variables to explain the shared variance. Likewise, weak internal reliability suggested that the latent structure for this composite variable was not statistically sound.

Phase 3

Phase 3 involved applying the two types of cluster analysis methods to the high school level data to determine the underlying structure. Two separate methods of clustering analyses were performed. In the first method, a Ward's method clustering analysis was performed. In the second method, a k-means clustering analysis was performed. Finally, the results of both methods were analyzed to validate and compare the results under both analyses.

Phase 4

Once the cluster analyses were performed and analyzed, the next phase was to construct a series of models to predict student persistence. Four separate models were created, with either PROC LOGISTIC or PROC GLIMMIX being used to conduct the analyses. In the first model (Model 1), only student-level variables were used to predict retention. In the second model (Model 2), student-level variables were used at level-1 and unconditional random effects (random intercepts) were used at level-2. In the third model (Model 3), the best model using student variables was used at level-1, and the aggregate high school variables at were used at level-2. In the fourth model (Model 4), the best student-level variable was used at level-1, and the clusters were used at level-2. The results are interpreted and discussed in the following chapter.

Summary

The current chapter has given a summary of the sample descriptive statistics and methodology employed to estimate the influence of multi-level characteristics on a student's predicted probability of first-year retention. Steps taken to create the psychosocial variables were detailed and descriptive statistics were presented. After presenting the descriptive statistics, the four planned phases of analysis were described. The following chapter presents the results from the four phases.

CHAPTER V: RESULTS

Overview:

The current chapter presents the results of the separate analyses conducted to answer the research questions presented in Chapter 3. These questions included investigating the latent structure of the New Student Survey (NSS) questionnaire, examining the hierarchical clustering nature of the students within schools organization, and then constructing models to estimate student retention using student, school, and both levels.

Factor Analytic Results

Prior to performing the factor analytic procedure, a Kaiser-Meyer Olekin test for sampling adequacy was conducted. The results of the KMO test indicated an overall MSA of .808, suggesting that factor analysis may be appropriate. After examining the scree plot, as well as the eigenvalue cutoffs, it was determined that a four-variable structure best explained the variance. These variables represented the factors discussed in Chapter 2, namely factors for financial concerns, academic engagement, self-efficacy, and institutional commitment.

Table 7 below presents the rotated factor pattern for the variables. For a list of variable names and descriptions, see Table 23 in the appendix. In addition, please see Chapter 2 for the coding specifics of each variable. In the table below, Factor 1 represents academic engagement, Factor 2 represents self-efficacy, Factor 3 represents institutional commitment, and Factor 4 represents financial concerns. All factor loadings represent the rotated factor pattern, and any variable that loaded below a .30 was omitted from the analysis.

Table 7: Factor Loadings After Rotation

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Communality
HsStudy	.44	-	-	-	.22
NewQuestion	.46	-	-	-	.22
WorkedHarder	.43	-	-	-	.24
Challenged	.39	-	-	-	.16
Rarely	.62	-	-	-	.39
WentTo	.61	-	-	-	.39
Homework	.62	-	-	-	.41
Late	.36	-	-	-	.15
Bored	.47	-	-	-	.22
Skip	.35	-	-	-	.15
WaitedAssignments	.71	-	-	-	.52
WaitedExam	.69	-	-	-	.49
Doubts	-	.47	-	-	.27
ConfidentSucceed	-	.70	-	-	.51
ConfidenceAbilities	-	.70	-	-	.50
DoWell	-	.48	-	-	.25
Important	-	-	.46	-	.27
ConfidentOU	-	-	.52	-	.36
OU Choice	-	-	.59	-	.36
Transfer	-	-	.46	-	.25
NotAccept	-	-	.57	-	.32
NotAfford	-	-	.64	-	.44
Need	-	-	-	.81	.67
AidRec	-	-	-	.54	.32
Money	-	-	-	.68	.49
Costs	-	-	-	.33	.17
Resource		-	-	.73	.55

Internal Reliability Results

Following confirmation of the factor structure, internal reliabilities of the composite variables were examined. Tables 8 through 11 display the intra-factor correlations amongst the variables as well as the internal reliabilities. All internal reliabilities were measured using Cronbach's alpha statistics.

Table 8: Internal Reliability and Correlations for Financial Concerns

	Variable				
	Q1	Q2	Q3	Q4	Q5
Q1	1.00	-	-	-	-
Q2	0.41	1.00	-	-	-
Q3	0.55	0.33	1.00	-	-
Q4	0.22	0.43	0.17	1.00	-
Q5	0.62	0.34	0.52	0.16	1.00

*Note: Q1=Need, Q2= AidRec, Q3=Money, Q4=Costs, Q5=Resource; Cronbach's Alpha =.76

Table 9: Internal Reliability and Correlations for Academic Engagement

	Variable											
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Q1	1.00	-	-	-	-	-	-	-	-	-	-	-
Q2	0.35	1.00	-	-	-	-	-	-	-	-	-	-
Q3	0.27	0.23	1.00	-	-	-	-	-	-	-	-	-
Q4	0.29	0.35	0.32	1.00	-	-	-	-	-	-	-	-
Q5	0.48	0.51	0.34	0.40	1.00	-	-	-	-	-	-	-
Q6	0.22	0.20	0.27	0.14	0.23	1.00	-	-	-	-	-	-
Q7	0.29	0.19	0.34	0.18	0.31	0.53	1.00	-	-	-	-	-
Q8	0.06	0.08	0.09	0.09	0.14	0.23	0.33	1.00	-	-	-	-
Q9	0.20	0.27	0.14	0.29	0.33	0.27	0.22	0.20	1.00	-	-	-
Q10	0.09	0.12	0.13	0.11	0.17	0.21	0.31	0.52	0.19	1.00	-	-
Q11	0.20	0.21	0.27	0.17	0.34	0.49	0.50	0.26	0.33	0.22	1.00	-
Q12	0.24	0.23	0.22	0.18	0.40	0.44	0.42	0.26	0.32	0.23	0.69	1.00

*Note: Q1= HsStudy, Q2= NewQuestion, Q3=WorkedHarder, Q4= Challenged, Q5=Rarely, Q6= WentTo, Q7=Homework, Q8= Late, Q9= Bored, Q10=Skip, Q11=WaitedAssignments, Q12=WaitedExam; Cronbach's Alpha = .82

Table 10: Internal Reliability and Correlations for Self-Efficacy

	Variable			
	Q1	Q2	Q3	Q4
Q1	1.000	-	--	--
Q2	0.416	1.000	--	--
Q3	0.340	0.510	1.000	--
Q4	0.267	0.332	0.378	1.00

Note: Q1=Doubts, Q2=ConfidentSucceed, Q3=ConfidenceAbilities, Q4=DoWell; Cronbach's Alpha = .705

Table 11: Internal Reliability and Correlations for Institutional Commitment

	Variable					
	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.000	--	--	--	--	--
Q2	0.520	1.000	--	--	--	--
Q3	0.289	0.324	1.000	--	--	--
Q4	0.451	0.416	0.219	1.000	--	--
Q5	0.137	0.202	0.396	0.269	1.000	--
Q6	0.213	0.274	0.246	0.225	0.505	1.000

Note: Q1=Important, Q2=ConfidentOU, Q3=OuChoice, Q4=Transfer, Q5=NotAccept, Q6=NotAfford; Cronbach's Alpha = 0.742

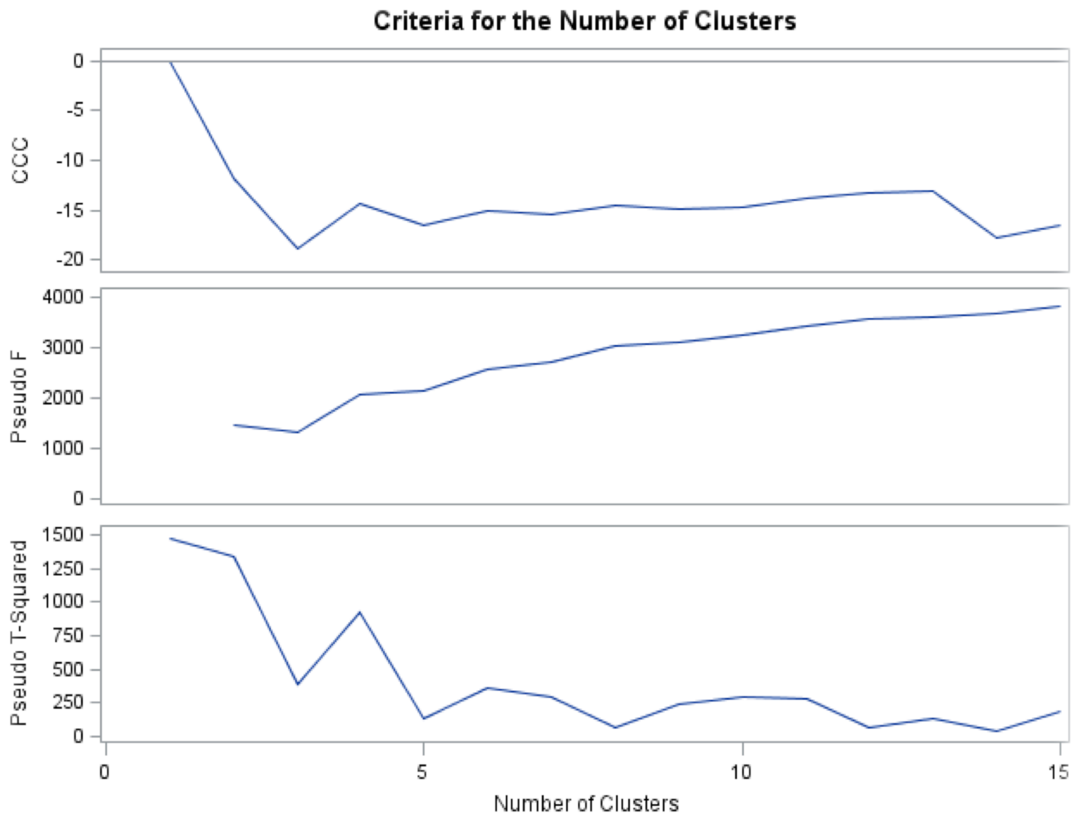
Clustering Analysis Results

After confirming the internal reliability and factor structure of the psychosocial variables, the next step was to examine the clustering analysis of the high school variables. The high schools were clustered according to the following variables: sector, class size, graduation rate, city size, district spending per student, percentage of students qualifying for free lunch, and student-teacher ratio. Because the data for city size was so spread out (particularly due to outliers such as a New York City and Los Angeles), the cities were classified as being in one of six city-types. Cities larger than 500,000 people were labeled as "very large", cities between 250,000 and 500,000 people were labeled as "large", cities between 100,000 and 250,000 people were labeled as "large/medium", cities between 50,000 and 100,000 people were labeled as "small/medium", cities between 25,000 and 50,000 people were labeled as "small", and cities with fewer than 25,000 people were labeled as "very small".

Ward's Method

The first method of analysis involved using PROC Cluster with Ward's method. Because the ideal number of clusters is not known, the cubic clustering criterion (Sarle, 1983) and the pseudo F and t^2 statistics will be examined. Figure 1 below presents the graphs for the CCC (Cubic Cluster Criterion[8]) as well as the pseudo F (Calinski & Harabasz, 1974) and pseudo t^2 (Duda & Hart, 1973) statistics.

Figure 1: Cluster Analysis Plot For Ward's Method



Examining Figure 1, it appears that the appropriate number of clusters may be between 3 and 4. Table 12 below presents the eigenvalues for the cluster analysis using Ward's method.

---

[8] For more on the Cubic Clustering Criterion, see Sarle (1983).

Table 12: Eigenvalues of Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	10432624.9	10378108.0	0.9948	0.9948
2	54516.8	54076.7	0.0056	0.9999
3	440.1	371.7	0.0000	1.0000
4	68.4	54.9	0.0000	1.0000
5	13.5	10.3	0.0000	1.0000
6	3.2	3.2	0.0000	1.0000

After examining the eigenvalues from the cluster procedure, the next step was to graphically examine the results. It was determined that a four-cluster solution best explained the variance; the clusters were then plotted on 2 axes using canonical variate analysis to maximize the interpretability. Figure 2 presents the clustering analysis results. The axes labeled Can1 and Can2 represent the canonical variate axes. Table 13 presents the means for the variables across each cluster.

Table 13: Descriptive Statistics for Ward's Method Clustering

	Cluster			
Variable	1	2	3	4
N	300	414	129	47
Grad Rate	89.48	90.73	92.72	88.63
	(8.85)	(10.00)	(8.27)	(13.97)
Student: Teacher	15.50	15.79	15.35	13.40
	(3.88)	(4.68)	(4.38)	(3.78)
Free Lunch	35.19	22.18	18.96	19.76
	(20.17)	(20.22)	(18.73)	(20.70)
District Spending	8025.79	10639.14	14403.40	20209.98
	(815.39)	(982.00)	(113.90)	(2805.27)
Grad Class Size	204.99	330.16	304.45	253.02
	(201.60)	(253.11)	(207.80)	(204.56)
City Type	1.22	2.32	1.71	1.65
	(1.75)	(1.88)	(1.74)	(1.98)
% Private	8%	21%	17%	23%
% Retained	82%	86%	84%	86%

Recalling from earlier, city types were coded such that any city with fewer than 25,000 people was coded as citytype = 0, cities between 25 and 50,000 was considered citytype=1, cities between 50,000 and 100,000 were coded as citytype = 2, cities between 100,000 and 250,000 were coded as citytype = 3, cities between 250,000 and 500,000 people were coded as citytype = 4, and cities greater than 500,000 people were coded as citytype = 5. Using this coding reference, it can be seen that the cluster with the largest cities was Cluster 2, and the cluster with the smallest cities was Cluster 1.

Table 14 presents the breakdown of city type by cluster. The actual number is presented first, followed by the percentage of the cluster breakdown by city type. For example, it can be seen that approximately 28.74% of cities in Cluster 1 are considered “very Small”,

Table 14: City Type Breakdown for Clusters

	Cluster			
City Type	1	2	3	4
Very Small	119 (28.74 %)	51 (39.53 %)	178 (59.33 %)	22 (46.81 %)
Small	35 (8.45 %)	17 (13.18 %)	25 (8.33 %)	7 (14.89 %)
Small / Medium	64 (15.46 %)	14 (10.85 %)	28 (9.33 %)	4 (8.51 %)
Medium / Large	65 (15.70 %)	26 (20.16 %)	23 (7.67 %)	0 (0 %)
Large	50 (12.08 %)	7 (5.43 %)	14 (4.67 %)	7 (14.89 %)
Very Large	81 (19.57 %)	14 (10.85 %)	32 (10.67 %)	7 (14.89 %)

Examining Table 14, it can be seen that Clusters 1 contains a higher percentage of larger cities than the other 3 clusters. Additionally, the largest percentage of very small cities occurs in Cluster 3. The largest percentage of middle to large cities occurs in clusters 1 and 2. Examining these city types can potentially be helpful in explaining canonical variate axis 2.

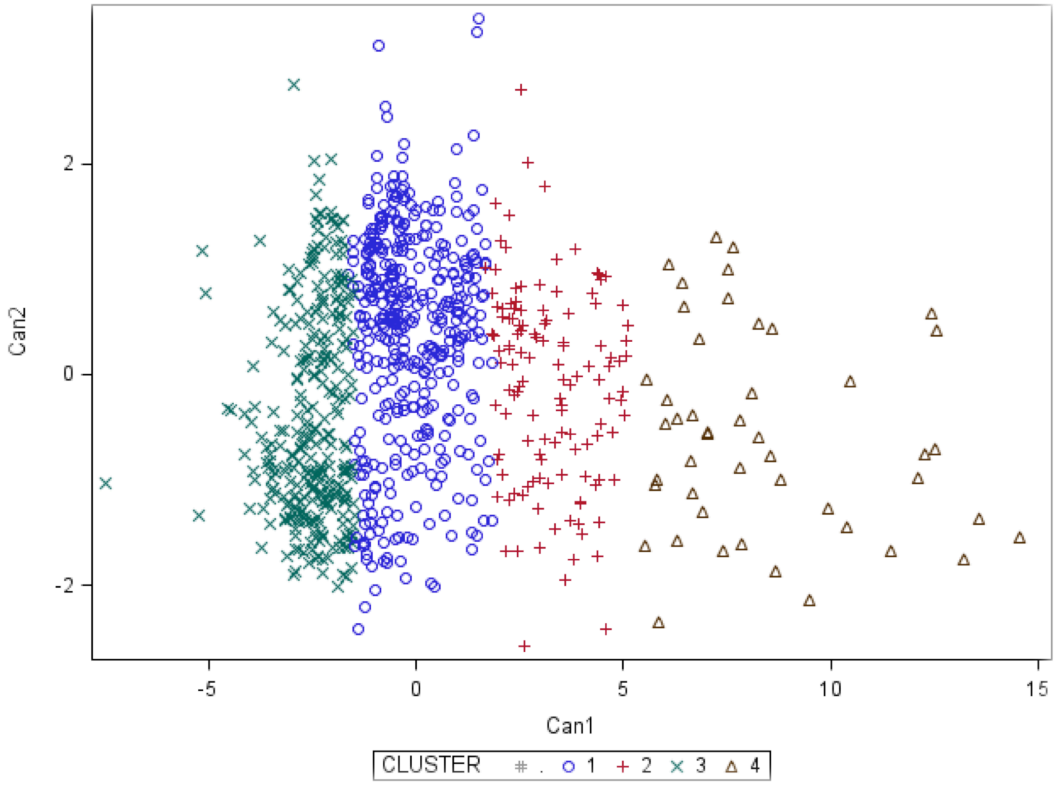
Another way of potentially helping to explain the canonical variate axes is to examine the coefficients associated with each variate. Results from the canonical analysis showed that the first variate was most highly associated with district spending, while the second canonical variate was most associated with high school class size and city type. This makes sense however, because larger cities typically have larger high schools. Table 15 presents the standardized canonical coefficients for the two canonical variates.

Table 15: Canonical Coefficients

Variable	Can1	Can2
Grad Rate	-0.066	-0.246
Student Teacher	-0.045	-0.142
Free Lunch	-0.045	-0.468
District Spending	2.878	-0.248
Class Size	0.040	0.606
City Type	-0.014	0.485
Sector	0.051	0.296

Results from Table 15 show that district spending is clearly the most influential coefficient for variate axis 1. Results also show that canonical variate axis 2 is slightly more evenly dispersed with the highest coefficients being associated with class size and city type. Further interpreting the second canonical variate axis, results show that it also represents contrasts as well, with negative coefficients on graduation rate, and lower student teacher ratios, as well as less district spending and having fewer students on free lunch. It is also positively associated with sector, in this case private schools (private being coded as 1) however to a less degree than with the other variables.

Figure 2: Cluster Analysis of High Schools Using Ward's Method



K-Means Clustering Method

The next method of clustering analysis involved using k-means clustering. The analysis was performed using the PROC Fastclus option in SAS. Based on the results of the previous clustering analysis, the number of clusters was set to four. The maximum number of iterations was set to 1,000, however the algorithm converged after seven. Table 16 below presents the final location statistics for the clusters, including frequencies and maximum distances from the seed observations. As can be seen, the first cluster appears to be further away from the group than the remaining three clusters.

Table 16: Statistics for K-Means Clustering

Cluster	Freq	RMS Std Dev	Maximum Distance From Seed	Nearest Cluster	Distance between Cluster Centroids
1	35	1361.2	14987.5	3	7302.2
2	329	324.7	5752.8	4	2573.2
3	164	584.7	3661.9	4	4100.0
4	407	368.6	2046.1	2	2573.2

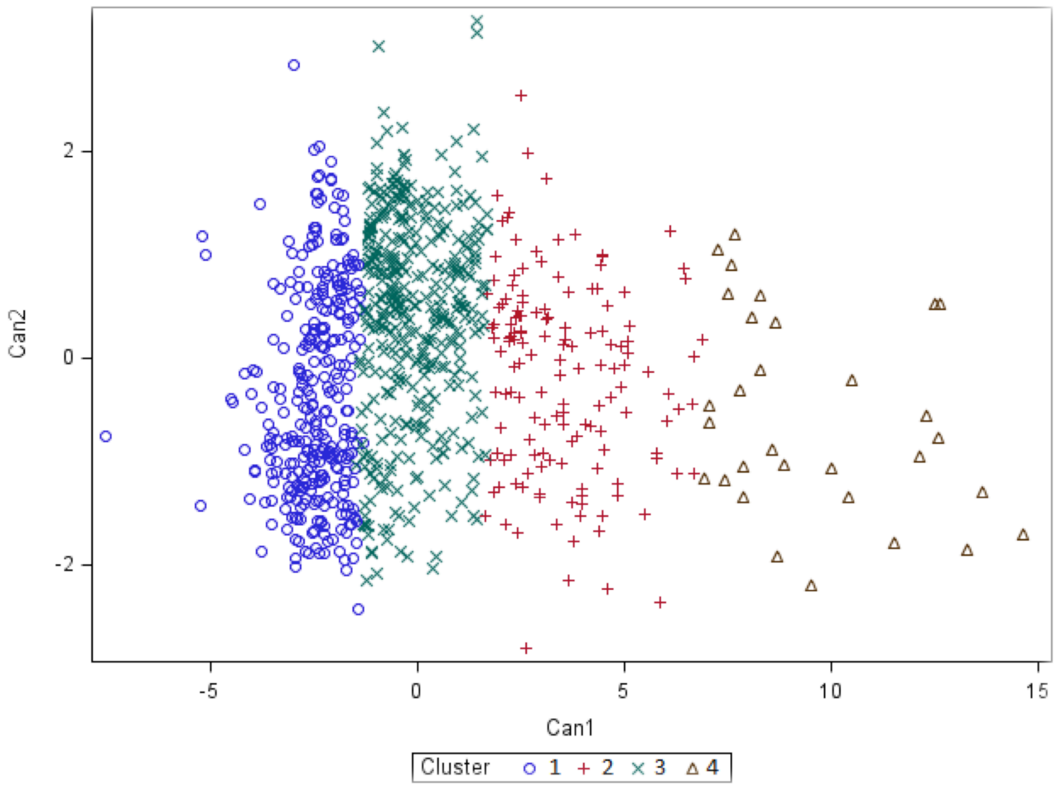
The pseudo-F statistic (2007.26) and approximate expected over-all R-squared value (0.93358) both indicated that the clustering analysis did a good job of partitioning the variance. The descriptive statistics for the K-Means clustering groups is depicted in Table 17. In the table below, the mean for each variable is presented first, followed by the standard deviation in parenthesis. For a table of correlations between the school-level variables, see Table 22 in the appendix.

Table 17: Descriptive Statistics for K Means Clustering

	Cluster			
Variable	1	2	3	4
N	327	407	162	35
Grad Rate	89.82 (8.98)	90.75 (10.12)	92.17 (8.81)	89.11 (13.44)
Student: Teacher	15.48 (4.10)	15.75 (4.65)	17.06 (23.84)	14.08 (4.13)
Free Lunch	34.58 (21.30)	21.43 (19.72)	19.94 (20.32)	18.16 (19.17)
District Spending	8089.83 (834.59)	10660.01 (941.28)	14759.79 (1532.71)	22061.94 (3591.87)
Grad Class Size	207.94 (202.77)	332.21 (253.52)	292.42 (207.97)	311.77 (259.13)
City Type	1.25 (1.74)	2.40 (1.89)	1.62 (1.75)	1.45 (1.96)
% Private	9%	22%	17%	20%

Figure 3 presents the graphical representation of the clustering analysis results using the k-means method. Returning to the previous discussion of the canonical variate axes, it appears once again that axis 1 is capturing the district spending per student variable, while axis 2 is capturing a variety of other characteristics. Importantly as well, it appears as though the clustering structure between the two methods is being replicated fairly well, suggesting agreement between the two.

Figure 3: Scatterplot of K-Means Clustering



As previously noted, the clustering analysis appears to present a very similar graphical depiction as that depicted from the Ward's method clustering. Table 18 presents the cross tabulation for the classification of variables using both Ward's method (Cluster 1) and k- means method (Cluster 2). Examining the classification system, results showed that approximately 97.08% of items remained in the same cluster after using both Ward's method and k-means clustering.

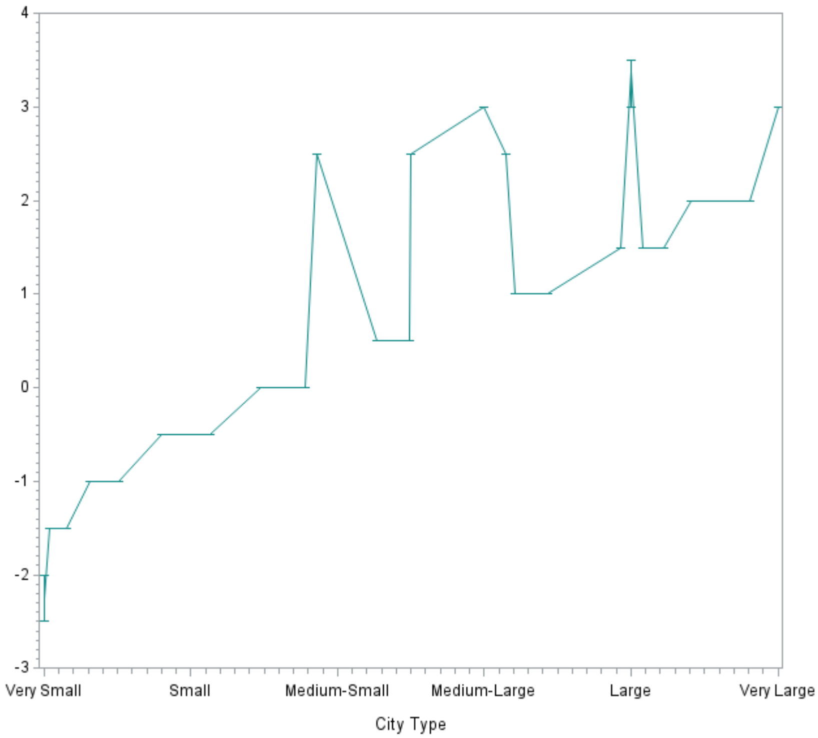
Table 18: Classification Of Observations

	Cluster 2 (K-Means)				
Cluster 1 (Ward's)	1	2	3	4	Total
1	300	0	0	0	300
2	16	394	4	0	414
3	0	0	129	0	129
4	0	0	16	31	47

Examining the canonical variate axes on Figures 2 and 3, it appears that the first canonical variate axis (can1) is capturing the dimension of district spending per student. This is most likely a proxy for socio-economic status as well, as these schools have the fewest percentage of students qualifying for free or reduced lunch. Examining the second canonical variate axis, it appears as though the axis is capturing the dimension of city size. Although not immediately apparent, this dimension becomes more interpretable when analyzing the differences in schools after sorting on the second canonical axis.

To examine the descriptive statistics potentially explaining the axis, the canonical axis values were rounded to the nearest .5, and then means for descriptive statistics were analyzed. Figure 4 depicts the graphical orientation of the high schools across city size. Because district spending per student was such a strong influence in creating the clusters, a 5th model, using district spending at the school-level, and the optimal model of student variables at the student-level was created for exploratory purposes. The results of this model are presented in Appendix 3. For more figures depicting the descriptive statistics by canonical variate axis 2, see Appendix 2.

Figure 4: Canonical Variate Axes



Having analyzed the clustering of high school types, the next step involved performing the hierarchical clustering logistic regression analysis. To perform this analysis, the clustering classification for the high schools were output and merged with the data file to give each high school a cluster location. This cluster then served as a level two variable in the hierarchical logistic regression analysis. In effect, this allowed for clustering by "high school type".

Differences in Means Between the Groups

One final step before moving into the models was as to examine the mean differences in the here / not here group for the student-level (level-1) variables. Of the 3,886 students in the sample, approximately 631 (16.24%) were not retained. This left a retention rate of approximately 83.76%. The means for the here / not here groups are presented in Table 19.

Table 19: Descriptive Statistics for Level-1 Variables

Variable	Group	N	Mean	Sd	Min	Max
ACT Score	Not Here	626	24.61	3.73	13.00	36.00
	Here	3249	26.12	4.08	13.00	36.00
High School GPA	Not Here	629	3.47	0.32	2.16	4.00
	Here	3253	3.61	0.32	1.70	4.00
Academic Engagement	Not Here	463	3.33	0.58	1.00	4.91
	Here	2326	3.46	0.58	1.25	5.16
Financial Concerns	Not Here	475	3.11	0.77	1.00	4.60
	Here	2381	2.82	0.76	1.00	4.60
Alumni Ties	Not Here	482	27%	-	-	-
	Here	2419	34%	-	-	-

As can be seen, the means for the financial engagement variable are higher for the non-returning students, while the means for the academic engagement variable are slightly higher for the returning students. Additionally, as can be seen, the ACT scores and high school GPA's for the returning students are slightly higher as well. Finally, notice that those students returning had a greater percentage of immediate family members who had attended OU.

Non-Mixed Logistic Regression Models

Student Level Variables

Prior to constructing the multi-level models, two separate logistic regression models were built. The first logistic regression model used only the student level variables. This model was designed to give an idea about which of these variables might best predict student persistence.

Table 20 presents the results for the logistic regression model using only the student variables. Results showed that variables for academic engagement (χ^2 (1, 2635) = 18.9383, $p < .01$), financial concerns (χ^2 (1, 2635) = 54.0008, $p < .01$), ACT (χ^2 (1, 2635) = 33.3773, $p < .01$), high school GPA (χ^2 (1, 2635) = 37.9217, $p < .01$), and alumni ties (χ^2 (1, 2635) = 4.0351, $p < .05$) all significantly contributed towards improving model fit. Odds ratio estimates indicated that the largest effect could be attributed to financial concerns (1.510/1.00), with ACT score (1.468/1.00), high school GPA (1.419/1.00), academic engagement (1.298/1.00), and then alumni ties (1.277/1.00) finishing out the list.

Table 20: Results from the Logistic Regression Model Using Student Variables

Parameter	Df	Estimate	Std Err	Wald $\chi 2$	Pr > $\chi 2$
Intercept	1	2.0052	0.1057	359.76	< .0001
Academic Engagement	1	0.2538	0.0583	18.93	< .0001
Financial Concerns	1	-0.4123	0.0561	54.00	< .0001
Institutional Commitment	1	0.0545	0.0577	2.72	NS
Self-Efficacy	1	0.0943	0.0571	2.72	NS
ACT Score	1	0.3841	0.0665	33.37	< .0001
High School GPA	1	0.3500	0.0568	37.92	< .0001
Alumni Ties	1	-0.2425	0.1207	4.03	< .05

Observation of the fit statistics indicated that the model fit the data fairly well. Concordance rates of 70% and an ROC score of 0.7 indicated that the model predicted cases at a rate significantly better than chance (ROC=0.5), but not perfectly (ROC=1.0). Analysis of the results indicated that a one standard deviation increase in academic engagement increased the probability of retention from 86% to 89% while an increase in financial concerns decreased the probability of retention to 80%.

Standard deviation increases in ACT score and high school GPA increased the probability of retention to 90% and 89% respectively. Taken together, the results indicated that the student most likely to be retained was a student who had alumni ties to the institution, was high in academic engagement, low in financial concerns, and was entering with good high school GPA and standardized test scores.

School Variables

The logistic regression model using the school level variables included variables for the number of students eligible for free lunch, the high school class size (rounded to the nearest 20), the sector (public or private), the senior class graduation rate, the student teacher ratio, the district spending per student, and the city type. Results from the logistic regression model using only the school variables indicated that variables for free lunch, $\chi 2$ (1, 3780) = 17.75, $p < .0001$ and high school class size, $\chi 2$ (1, 3780) = 6.67, $p < .0001$ each significantly improved model fit.

Concordance rates of 60% and a Somer's D score of 0.218 each indicated that the model fit the data moderately well. Approximately 60% of the observations were under the c curve. Odds ratio estimates indicated that a one standard deviation decrease in the number of students eligible for free lunch improved the likelihood that a student would persist to a ratio of 1.014/1.00 while a standard deviation increase in high school class size increased the odds of retention 1.001/1.00. Comparison of the fit statistics indicated that the model using only the school level variables did not predict student persistence as accurately as the model using the student level variables.

A summary of results from logistic regression model using only the school level variables is presented in Table 21.

Table 21: Results from Logistic Regression Student Model

Parameter	Df	Estimate	Std Err	Wald $\chi 2$	Pr > $\chi 2$
Intercept	1	1.5276	0.7777	3.85	< .05
Free Lunch	1	-0.0137	0.0032	17.75	< .0001
Class Size	1	0.0006	0.0002	6.67	< .01
Sector	1	0.1068	0.2232	0.22	NS
Graduate Rate	1	0.0024	0.0067	0.13	NS
Student / Teacher	1	0.0.007	0.0170	0.00	NS
District Spending	1	-0.0001	0.0001	0.50	NS
City Type	1	0.0455	0.0273	2.78	< .10

Because variables for city type, percentage eligible for free lunch, and high school class size each significantly improved model fit, these variables will be used when the aggregate high school variables are modeled at level-2.

Multi-Level Models

Examining the Intra-Class Correlation

Before moving into Model 1, the Intra-class Correlation (ICC) was examined to determine the amount of estimated variance in the slopes at the second level. This involved fitting random intercepts at Level-2, while using no predictors at Level-1. The model can be written as:

$$\Pr Y_{ij} = \frac{e^{\Upsilon_{00}+\mu_{0j}+r_{ij}}}{1+e^{\Upsilon_{00}+\mu_{0j}+r_{ij}}} \qquad \textbf{Equation 10}$$

where $r_{ij} \approx N(0, \sigma^2)$, Y_{00} represents the grand mean, and μ_{oj} represents the difference between the individual school mean and the grand mean (the unique school effect). Because the current data set contained over 950 schools, the individual results for the schools will be not be reported. Aggregate results indicated that the estimated variance in the intercepts = .3997 (standard error = .1016) accounted for approximately 10% of the total variance.

$$ICC = \frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + \pi^2 / 3} = \frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + 3.29} = \frac{.3997}{.3997 + 3.29} \approx .10 \qquad \textbf{Equation 11}$$

Results from the ICC indicate that the amount of variance that can be attributed to the second level clusters is relatively small compared to the amount of variance that exists at the student level.

Model 1

Having examined the ICC and examined the descriptive statistics, the next step was to build a series of multi-level models to estimate student persistence. Model 1 involved using the best predictors from the logistic regression model at to model random intercepts at level-1, with no predictors at level-2. Results from Table 17 (the logistic regression model using student variables) indicated that ACT, high school GPA, alumni ties, academic engagement, and financial concerns each significantly predicted retention. Because of this, these variables were used at Level-1.

The two parts of the model can be written as:

$$Y_{ij} = B_{00} + B_{1j}(\text{HsGPA})_{ij} + B_{2j}(\text{Financial\_Concerns})_{ij} + B_{3j}(\text{ACT})_{ij} + B_{04}(\text{HsGPA})_{ij} + B_{02}(\text{Members})_{ij} + r_{ij}$$

where

$$B_{00} = \Upsilon_{00} + u_{0j}$$

Υ_{00} is the grand mean, and μ_{0j} is the unique error associated with each prediction at level-2. Note here that there are no level 2 predictors, and as such, the current model is only using the level-1 variables to try and reduce the variance in the level-2.

Results demonstrated that all variables at level-1 significantly improved model fit. Specifically, results showed that students higher in academic engagement,

$t(1950) = 4.11$, $p < .0001$, lower in financial concerns, $t(1950) = -6.37$, $p < .0001$, with greater incoming ACT scores $t(1950) = 5.27$, $p < .0001$, and higher incoming high school GPA's, $t(1950) = 6.42$, $p < .0001$ were more likely to be retained. Results also showed that those students with no alumni ties to OU were less likely to be retained $t(1950) = -1.95$, $p < .05$. A summary of results from Model 1 is presented in Table 22.

Table 22: Summary of Results From Model 1

Effect	Estimate	Standard Error	DF	t Value	pr > |t|
Intercept	1.9535	0.1144	765	17.08	< .0001
Acad Engagement	0.2365	0.0575	1950	4.11	< .0001
Fin Concerns	-0.3610	0.0567	1950	-6.37	< .0001
ACT Score	0.3496	0.0663	1950	5.27	< .0001
High School GPA	0.3690	0.0565	1950	6.42	< .0001
Alumni Ties	-0.2407	0.1232	1950	-1.95	< .05

Examination of the T-matrix indicated that the estimated variance in the intercepts = 0.2258 with a standard error of 0.1179. This suggests that including the level-1 variables did shrink the variance in intercepts from the model with no predictors (where the variance = .3997). Tests of the covariance parameter for intercepts based on the residual pseudo-likelihood rejected the null hypothesis that there were no random effects $\chi^2(1) = 6.35$, $p < .001$. Examination of the fit statistics indicated that the pseudo AIC = 13456.67 and the pseudo BIC = 13498.01. These fit statistics set the standard for comparison between models (where smaller statistics indicate better fitting models).

One final note—when performing multi-level modeling, it is often helpful to center the variables around the mean. This can help in interpreting the results, as the coefficients will show the influence of the variable as being a certain amount above or below the mean score. In the current case, many of the student level variables were standardized to a mean of 0 and a standard deviation of 1. In doing so, this has effectively mean-centered the variables in a way that is traditionally performed using multi-level modeling.

Model 2

Having constructed the random intercepts model, the next step was to use the best fitting variables from the level-1 model, and then fit a unique intercept for each school variable at level 2. This model can be written out in two separate parts as:

$$Y_{ij}=B_{00}+B_{1j}(HsGPA)_{ij}+B_{2j}(Financial\_Concerns)_{ij}+B_{3j}(ACT)_{ij}+B_{04}(Academic\_Engagement)_{ij}+B_{02}(Members)_{ij}+r_{ij}$$

where

$$B_{00}=\Upsilon_{00}+\alpha_{0j}+u_{0j}$$

Υ_{00} is the grand mean, α_{0j} is the unique effect of each individual school on a student's score, and μ_{0j} is the unique error associated with each prediction at level-2. For example, if a student came from a particular school, and the mean probability of retention was 83%, then Model 2 effectively allows for the fixed effect of having attended a given school on either increasing or decreasing this probability.

After fitting a unique effect for each school, the results showed that all variance in the intercepts was effectively reduced to zero. Results showed that variables for academic engagement, $t(1956) = 2.19$, $p < .05$, financial concerns, $t(1956) = -3.27$, $p < .001$, high school GPA, $t(1956) = 7.82$, $p < .001$ and ACT score, $t(1956) = 3.38$, $p < .001$ each significantly contributed to model fit. For comparison purposes, the AIC = 2932.60 and the BIC = 6506.31. Tests of the random effects (based on likelihood) indicated that the null hypothesis of no random effects could not be rejected, χ^2 (1) = 0.00, p = 1.00.

Model 3

Model 3 involved creating a model using the aggregate school level variables from Table 18 at level-2, while retaining the student level variables from Table 17 at level-1. The specific level-2 variables included the percentage of students qualifying for free lunch, the high school class size (rounded to the nearest 20) and the type of city where the school is located. All variables at Level-2 were treated as fixed effects.

Writing out the model in an equation, the probability of a student persisting can be written as:

$$Y_{ij}=B_{00}+B_{1j}(HsGPA)_{ij}+B_{2j}(Financial\_Concerns)_{ij}+B_{3j}(ACT)_{ij}+B_{4}(Academic\_Engagement)_{ij}+B_{02}(Members)_{ij}+r_{ij}$$

where

$$B_{00}=\Upsilon_{00}+\gamma_{01}(Free\_Lunch)_{j}+\gamma_{02}(Class\_Size)_{j}+\gamma_{03}(City\_type)_{j}+\mu_{0j}$$

Results from Model 3 are presented in Table 23. Results indicated that, at level-2, variables for free lunch, t (737) = -3.63, p < .001, high school class size, t (737) = 2.90, p < .001, and city type, t (737) = 3.67, p < .001 each significantly improved model fit. Additionally, results from the level-1 variables indicated that academic engagement, t (1934) = 3.17, p < .001 financial concerns, t (1934) = -5.32, p < .001, ACT score, t (1934) = 3.81, p < .001, high school GPA, t (1934) = 7.83, p < .0001, and alumni ties, t (1934) = -1.89, p = .057 each significantly improved model fit.

Table 23: Summary of Results From Model 3

Effect	Estimate	Standard Error	Df	t value	pr |t|
Intercept	1.7353	0.1769	737	9.85	< .001
Free Lunch	-0.0111	0.0030	737	-3.63	< .0001
Class Size	0.0006	0.0002	737	2.90	< .001
City Type	0.1195	0.0326	737	3.67	< .0001
High School GPA	0.4661	0.0959	1934	7.83	< .0001
ACT Score	0.2586	0.0679	1934	3.81	< .0001
Acad Engagement	0.1862	0.0587	1934	3.17	< .001
Financial Concerns	-0.3102	0.0583	1934	-5.32	< .0001
Alumni Ties	-0.2339	0.1231	1934	-1.90	0.057

Examination of the covariance parameter estimates indicated that the intercept variance = 0.1131 (SE = 0.087), suggesting that Model 3 effectively reduced the variance in intercepts more than Model 2. Examination of the fit statistics indicated that the Pseudo-AIC = 13429.00, the Pseudo-BIC = 13497.90 and the generalized Chi-SQ / DF = 0.93. Additionally, tests of the variance in intercepts (based on the residual pseudo-likelihood) indicated that the null hypothesis of no variation in intercepts could not be rejected, χ^2 (1) = 2.58, ns.

Model 4

Model 4 involved using the clusters at level-2 and the best fitting student variables at level-1. Although the results from Model 2 and 3 show that the variance in intercepts can be reduced significantly by including unique intercepts and estimators for the schools at level-2, given that there are so many schools (over 800), including a unique effect for each school can be tedious. Because of this, it can be helpful to see if there is a unique effect associated with each cluster. The variables at level-1 continue to be those used in the previous analysis.

Writing out the equation, it is clear that including the clusters as the level 2 predictors has dramatically decreased the complexity of the model. Continuing with the level-1 variables from the previous analyses, the first level model can be written as:

$$Y_{ij}=B_{00}+B_{1j}(HsGPA)_{ij}+B_{2j}(Financial\_Concerns)_{ij}+B_{3j}(ACT)_{ij}+B_{4j}(Academic\_Engagement)_{ij}+B_{05}(Members)_{ij}+r_{ij}$$

however, the second level model is dramatically reduced to

$$B_{00}=\Upsilon_{00}+\gamma_{01}(Cluster)_{j}+\mu_{0j}$$

Table 23 presents a summary of the results from Model 4. Results from Model 4 indicated that the influence of no single cluster was significantly different from zero. Although the influence of academic engagement, $t\ (1950) = 3.70$, $p < .001$, financial concerns, $t\ (1950) = -5.76$, $p < .0001$, ACT scores, $t\ (1950) = 4.81$, $p < .0001$, high school GPA, $t\ (1950) = 6.96$, $p < .0001$, and alumni ties, $t\ (1950) = 02.38$, $p < .05$ each significantly improved model fit, no individual cluster was significantly more or less likely to retain students at the individual level. Interestingly, results did show that the variance in the clusters did significantly differ at the aggregate level, $f\ (3,762) =3.97$, $p < .001$.

Table 19: Summary of Results From Model 4

Effect	Estimate	Standard Error	Df	t value	pr |t|
Intercept	2.1744	0.4456	762	4.88	< .001
Cluster 1	-0.4014	0.4451	762	-0.90	NS
Cluster 2	0.0797	0.4494	762	0.18	NS
Cluster 3	-0.0113	0.4849	762	-0.02	NS
Acad Engagement	0.2156	0.0582	1950	3.70	< .0001
Financial Concerns	-0.3322	0.0577	1950	-5.76	< .0001
ACT Score	0.3222	0.0669	1950	4.81	< .001
High School GPA	0.4085	0.0587	1950	6.96	< .0001
Alumni Ties	-0.2987	0.1255	1950	-2.38	< .01

Examination of the covariance parameter estimates indicated that using the clusters, rather than the individual high school aggregate variables as predictors increased the variation in intercepts to .2756 (SE = .1210). Fit statistics demonstrated that the Pseudo-AIC = 13486.98, the Pseudo-BIC = 13546.04, and the generalized chi-sq / DF = 0.88. Tests of covariance parameter estimates (based on residual pseudo-likelihood) rejected the null hypothesis of no random effects χ^2 (1) = 9.99, p < .0001.

Returning to the differences in cluster means, results showed that Cluster 1 had an adjusted mean probability of 81%, while clusters 2, 3, and 4 have probabilities of 86, 86, and 87% respectively. Because the F statistic did show a significant difference in the means, sets of pairwise comparisons were made to determine where the difference would occur. Examination of the comparisons in LS Means was conducted using a Bonferroni adjustment. Results showed that the significant difference existed between clusters 1 and 2, t (762) = -3.31, p < .001.

Summary of Fit Statistics

As can be seen from Table 24, results from the fit statistics indicated that Model 3 fit the data better than the other models. Examination of the smaller pseudo BIC, as well as the smaller AIC for Model 3 indicate that this model is the best fitting, even after taking into account the additional number of parameters being included at level-2.

Table 20: Summary of Fit Statistics

Model	- 2 Res Log Pseudo Likelihood	Pseudo-AIC	Chi Sq / DF	Pseudo-BIC
Model 1	13442.67	13456.67	0.89	13498.01
Model 2	1392.60	2932.60	0.63	6506.31
Model 3	13409.00	13429.00	0.93	13487.90
Model 4	13466.98	13486.98	0.88	13546.04

Summary

The current chapter has provided the results of several analyses investigating the influence of student- and school-level characteristics on the probability of second year retention. The chapter began with describing the results for the factor analysis and internal reliability estimates for the proxy-variables used to predict retention at the student-level. Results showed that variables for financial concerns, academic engagement, self-efficacy, and institutional commitment each demonstrated high internal reliability.

Next, cluster analysis was performed to examine whether the high schools could be clustered according to specific school type, rather than treated as individual intercepts. Results showed that the high schools could be clustered primarily according to two dimensions. The first dimension captured district spending per student, and primarily represented a socio-economic proxy. The second dimension captured city size, and primarily represented an urban or rural proxy.

After performing the cluster analysis, a series of models was constructed to estimate the influence of the student- and school-level variables on predicted student retention. Results from using the student-level variables in isolation demonstrated that academic engagement, financial concerns, alumni ties, standardized test scores, and high school GPA each significantly improved model fit. Results for using the school-level variables in isolation showed that variables for free lunch eligibility and class size each significantly improved model fit.

Results from the mixed modeling indicated that, when using the aggregate high school variables at the second level, high school class size, sector, high school

GPA, standardized test score, academic engagement, and financial concerns each significantly improved model fit. Discussions and implications of these findings are presented in the following chapter.

CHAPTER VI: DISCUSSION

Overview

The following chapter summarizes and interprets the research findings, evaluates their implications to the larger field of student retention, and presents directions for future research. Topics discussed include the internal reliability of the proxy-variables, the findings of the cluster analysis (and how clustering compares to using the single high school variables), the findings of the student, school, and multi-level models (and how these findings relate to previous results), and how this dissertation may be used to direct and assist future research.

Student-Level Variables

Results from the factor analysis and internal reliability measurements indicated that the student-level variables continued to demonstrate strong internal consistency. All Cronbach's alpha scores were minimally within the mid .70 range, scores considered sufficient for low stakes testing[9]; these results suggest that it is appropriate to use the variables in measuring student retention. These findings build upon a long history of using psychosocial variables to predict student retention (Tinto, 1975; Spady, 1970; Pleitz, Terry, Campbell & Fife, 2011; Bean & Eaton 2001-2002), and suggest that strictly relying on prior academic performance to predict retention is insufficient.

The influence of psychosocial characteristics in predicting student retention has taken center stage in theories like Tinto's Interactionist model of withdrawal (where variables like commitment and engagement predict whether a student will

---

[9] See Kline (2000)

leave an institution), or in the work of Levine and Cureton (1998) who focused on the role of social learning groups in developing locus of control.

Self-Efficacy

In the current dissertation, particular psychosocial variables of interest included self-efficacy, financial concerns, academic engagement, and institutional commitment. As previously mentioned in the introduction, self-efficacy has been used to explain a variety of phenomenon, including academic performance (Komarraju & Nadler, 2013). According to these authors, students low in self-efficacy may be at an inherent disadvantage because they believe that intelligence is fixed and not able to be improved with hard work or effort. Because of this, it is highly possible that students low in self-efficacy may become frustrated when they encounter setbacks, eventually developing a sense of learned helplessness regarding their own academic capabilities.

Returning to this concept of helplessness, it is also likely that students low in self-efficacy may be likely to attribute their academic shortcomings to external situations beyond their control. As such, these students may perform very well as long as they are earning high grades, however as soon as they encounter a setback, they may be more likely to become frustrated or blame the situation on their teachers (similar to the academic entitlement phenomena; Chowning & Campbell, 2009),, the institution, or a number of other external factors.

Academic Engagement

Along with self-efficacy, a second psychosocial factor shown to influence persistence decisions was academic engagement. Interestingly, while the influence of self-efficacy became negligible after controlling for all other factors (Table 22), the influence of academic engagement continued to remain significant. Furthermore, it was interesting to note that academic engagement remained significantly predictive even after controlling for high school GPA and standardized test scores. Because of this, it seems that academic engagement is capturing a unique phenomenon that is related to, but not the same as academic performance.

The relationship between academic engagement and performance is not surprising. Certainly, it seems that many academically engaged individuals are likely to participate in behaviors known to positively influence performance. These behaviors include coming to class on time, studying for homework and exams in a diligent manner, and being challenged to perform ones best work while in high school. Given that the college curriculum is oftentimes substantially more difficult than the high school curriculum, the positive influence of academic engagement suggests that successful high schools do more than simply teach students how to perform well, they teach students how to become immersed in their learning.

The influence of immersion in the rubric is a cornerstone of the phenomenon of active learning, a discipline that promotes student responsibility for participating in their learning (Bonwell & Eison, 1991). In recent years, the approach to active learning has become a popular theme in higher education, with the 2012 President's council of advisors on science and technology encouraging the practice as a method to increase performance in STEM courses as well as improve student retention[10].

[10] President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing on million additional college graduates with degrees in science, technology, engineering and mathematics.* Retrieved April 14, 2014 from:

More recent research has been conducted examining how active learning and engagement interacts with classroom technology (Morris & Chikwa, 2014), how cooperative learning influences student engagement (Herrmann, 2013), and even how social media can be used as an effective and active teaching tool (Kassens-Noor, 2012).

Financial Concerns

Finally, the results for the psychosocial variables indicated that financial concerns played a significant role in predicting retention. The significant influence of financial concerns may be of particular interest for two reasons. First, with the cost of higher education increasing substantially, it is highly likely that students are going to become more concerned with being able to afford college. Moreover, it may be that students are not necessarily worried about paying for college per se, but rather that they no longer consider the benefits of a college education to outweigh the cost of paying for school. Because of this, it is likely that alleviating financial concerns in college students is more about emphasizing the importance and benefits to be gained from attending college, rather than by simply contributing greater amounts of financial aid.

Ironically, a college education may be more important now than ever before. A February 2014 report by the Pew Research Institute demonstrated that Americans with only a high school diploma were expected to only earn 62% of what their peers with a college degree will earn. Nearly 22% of high school graduates are currently living in poverty, compared to only 6% of college graduates. The negative

---

http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_feb.pdf effects of having only a high school diploma are again demonstrated in employment rates, where high school graduates demonstrate unemployment rates nearly four times as high as their college peers (12.2% compared to 3.8%). High school graduates are also far less likely to be satisfied with their current job, are less likely to have a career-track job, and are far more likely to lack the skills and education required to get ahead in their job (Pew Research Center, 2014).

Given the powerful influence of financial concerns in predicting student retention, it is interesting that it has only recently become a well-researched area. Certainly there exist decades of research examining how financial aid impacts retention and performance, however, again the assessment of financial concerns is not to be confused with the assessment of financial aid (although the two may certainly be related). In one of the earlier analyses of financial concerns, Tinto (1982) argues that, perhaps unlike actual financial needs, financial concerns may be malleable and adjusted according to other factors.

Returning to the interactionist nature of withdrawal, Tinto argues that financial concerns may be more important in the early college career, when students are still uncertain about their future goals. With commitment to graduation being a pillar of Tinto's theory, it then makes sense that students who do not have a strong commitment (or perhaps are lacking a direction for how to obtain such a goal) may be more likely to weigh the costs of attending college as heavily against the benefits. Interestingly as well, and furthering the interactionist dynamic, Tinto argues that students who frequently encounter positive experiences while in college are more likely to accept the heavy financial burden, because they are receiving greater benefits. In either case, what's important is that the actual financial strain is the same, though the concerns may be completely different.

School-Level Variables

Free Lunch and Class Size

Results from the school-level model also indicated that certain variables predicted student retention. However, unlike the student-level variables (which represented a variety of influences), school-level variables tended to be primarily financial in nature. Examining Table 19, results show that variables for free lunch and class size significantly improved model fit. Specifically, students were most likely to be retained if they were coming from more wealthy districts, as well as from larger schools. These results were seemingly confirmed in Table 22 (the full mixed-model) where variables for sector and high school class size again predicted student retention.

Given these findings, it certainly seems to suggest that all high schools are not created equally. Not only do certain high schools have greater financial resources, but they also appear to produce students who have a greater probability for retention in higher education. Examining figures 2 and 3, results from the clustering analysis demonstrate that the primary reasons for high schools to group (or differentiate) is due to financial variables. As seen in these figures, not only does a definite clustering pattern exist, but also there exist widespread discrepancies in the high schools themselves.

Examining the cluster means again, descriptive statistics show that approximately 28% of students in the first cluster are eligible for free lunch, with district spending per student averaging approximately $8,000. Compare this to schools in cluster 4 where only 8% of students are eligible for free lunch, and the average district spending per student is nearly three times as high at approximately $22,000. Given that financial concerns at the student level played a strong predictor of retention, it is likely that these students are coming from schools with a lower level of SES.

Results again seem to support this theory. Examining the descriptive statistics of financial concerns and retention across the clusters, results show that students in the first cluster (the cluster with the least amount of district spending per student, as well as the cluster with the highest number of students eligible for free lunch) demonstrated higher levels of financial concerns than students in the other three clusters. Additionally, these students were four percent less likely to be retained than their peers coming from more affluent districts.

The Role of Academic Engagement in Retention

The finding that district affluence strongly predicts retention at the school level is, by itself, not all that surprising. Given that districts with greater spending power can hire seemingly more qualified teachers, offer more advanced placement courses, and have greater resources for tutoring and teaching aids, it is perhaps expected that students coming from these districts are more likely to be prepared for the college rigor. What is surprising however is that students coming from these districts are actually entering college with lower incoming high school GPA's than their peers in less affluent districts. Returning to the concept of academic engagement, this seems to support the theory that it is not the grades earned per se that influence future academic performance, but rather how much effort and active learning was expended in earning the grades.

Comparing the level of academic engagement across the clusters, results showed that those students coming from the 4th cluster (the more affluent cluster) earned lower high school GPA’s, but demonstrated higher levels of academic engagement. With the greater emphasis on active learning previously discussed, it appears that high schools engaging in this process are producing students more adequately prepared for college rigor.

Implications for Policy

The current research findings provide powerful insight into the complicated dynamic between student and school characteristics in predicting college retention. As noted in the introduction, the value of a college education is becoming increasingly more important at the individual, local, and national level. Given this importance, educators and policy makers would be well justified in asking what these findings tell us about higher education retention research and how they can be used to improve student retention in higher education?

The following implications represent but a few of the multiple directions that may be taken in light of these results. As with many other psychological phenomena, the student retention decision does not exist in a vacuum. And rather than simplify the situation to an elegant formula for success, the current results may be better served as a reminder that a multitude of factors across financial, social, and academic domains are constantly acting and interacting to influence the persistence decision. With this in mind, the following suggestions may serve to provide students, administrators, and researchers in higher education with a set of potential practices to benefit and improve retention in higher education.

Implications for Students (Learned vs. Earned)

Beginning with the student-level findings, the first implication is to recognize what the results are and are not supporting. Although it is disconcerting to note that financial concerns continue to play an important role in influencing retention decisions, it is also important to note that academic preparation still plays a strong role in college success. And while a great deal of attention is (rightfully so) being paid to address the economic inequalities in American education, it is important to remember that students who are successful in high school typically are successful in college. Put more simply, if you want to know who will be successful in college, the first step is to examine who was successful in high school.

As previously mentioned however, the definition of success in high school should not be solely restricted to academic performance. Time and time again, the results have confirmed that those students who were more academically engaged in their education were more likely to be retained in college. Because of this, students should take the results as encouragement to become more active participants in their education. This not only includes developing conscientious habits (such as coming to class on time, doing the assigned reading and homework, and allowing adequate time to study for tests), but also includes developing a new approach towards their education.

This approach (again building upon the notion of active learning) should encourage students to ask questions, engage in conversations with students and teachers, take more challenging courses, and focus more on what is being learned than what is being earned. Students interested in this active approach are encouraged to read more about goal orientation theories (Elliot & Church, 1997), particularly focusing on an approach orientation to performance.

Implications for Administrators and Educators

While the implications for students primarily consist of becoming more actively involved in their educational experience, implications for administrators and educational researchers are a bit more complicated. The results above have demonstrated that student retention is the result of multiple factors, including financial, psychological, and academic domains. Because of this, the role of administrators and researchers may best be served by approaching the retention challenge from multiple perspectives.

From a financial perspective, the above results have shown that students are more likely to drop out if they are concerned about being able to pay for college. As previously stated, the cause of such concern may not necessarily be attributed to an actual lack of financial aid, but rather due to the perceived costs of attending college outweighing the benefits. Because of this, one of the first implications for administrators is to emphasize the importance of a college degree to their students.

Unfortunately, recent trends show that this value is not being preached to America's youth. In a 2011 Pew Research Survey, approximately 57% of respondents indicated that they did not feel college provided students with good value for the money. Paradoxically, amongst those surveyed that had graduated from college, nearly 75% of respondents indicated that the felt higher education had provided them with a very useful education for growing intellectually. These results, combined with the finding that more and more students are graduating with increasing college debt, indicate that perhaps now more than ever, administrators and faculty are tasked with emphasizing the importance of a college education[11].

Along with emphasizing the importance of a college education, researchers and administrators are also encouraged to use the presented results as a way to think about and potentially adjust their methods of instruction. This is not to say that many teachers in higher education are not providing a valuable and life-changing experience to their students in forms of effective instruction and assessment, but rather to encourage instructors to constantly be evaluating their methods of instruction. More specifically, as results have shown, the most effective instruction is often one that encourages an active and engaged learning style.

Fortunately, it does seem that more and more professors are approaching the classroom with an active learning perspective. New and engaging techniques for instruction are being implemented across higher education, with a greater amount of research being dedicated to how instruction can maximize active learning. This is evidenced by the prevalence of excellent research being conducted in the journal Active Learning in Higher Education and the Center for Research on Learning and Teaching at the University of Michigan.

The first teaching center in the country, CLRT at the University of Michigan is dedicated to enhancing learning and teaching at the University of Michigan, and

[11] Source: Pew Research Survey conducted Spring 2011. Retrieved April, 17, 2014 from: http://www.pewsocialtrends.org/2011/05/15/is-college-worth-it/

strives to promote a University culture that values and rewards teaching, respects and supports individual differences among learners, and encourages the creation of learning environments in which diverse students can learn and excel[12]. Research from the center provides several suggestions for improving the classroom environment, including direct links to a number of articles and resources dedicated to active learning in the classroom.

Finally, administrators and researchers are encouraged to view the results as a reminder that college students are not entering higher education as a blank slate. Rather, students bring with them a unique history, complete with a variety of experiences and expectations, each derived from a unique set of student and school-level background characteristics. Because of this, administrators and educators are encouraged to work with graduation coaches, college counselors, and other faculty to better understand how a student's background is likely to influence their persistence in higher education.

Returning again to the current discussion, implications may include developing an introductory transition course to educate students on the value of a college education, encourage students to explore majors and find an area of study they truly enjoy, and emphasize that the goal of college is to learn, rather than simply to earn high grades or get a good job.

Conclusions

Taken together, these implications represent one of the most important tenants that can be derived from the current results. Put simply, student retention is

[12] Mission statement taken from the Center For Research on Learning and Teaching website, available at http://www.crlt.umich.edu/aboutcrlt/aboutcrlt

not simple. Rather, it is an incredibly diverse and dynamic process. What used to be considered a shortcoming in academic qualifications is now being treated as a complicated dance between psychological, academic, and financial considerations. And because of this, remedying the student retention issue is far more complicated than simply addressing one of these areas. The current dissertation serves to present another helpful tool in solving this complex puzzle, yet also opens doors to a number of future research questions.

Chief amongst these questions is the greater understanding of how student and school level variables interact to predict retention. Although the current research findings suggest that financial variables play an important role at both levels, continuing research should be conducted to better understand the nature of this dynamic. For example, how do the greater financial concerns felt by students influence their priorities in college? Or how can the influence of goal commitment and major exploration interact with the cost- benefit analysis employed by students when determining whether to return to college?

Other questions may include examining how the current findings replicate in other arenas of higher education. For example, although factors such as alumni ties did show a significant improvement in model fit (see Table 21), it would be interesting to see if this influence continues across community colleges. More specifically, because alumni ties are likely to foster a sense of belonging and pride in the university (a form of social integration), and results have shown this sense of belonging to be less influential in predicting retention at community and commuter colleges (Bers & Smith, 1991; Straus & Volkwein, 2004), then it would be of value to investigate whether these findings are strictly applicable to the current environment.

REFERENCES

Abbot, M. L., & Joireman, J. (2001). The relationships among achievement, low income, and ethnicity across six groups of Washington state students. *Washington School Research Center (Technical Report No. 1).* Seattle WA: Seattle Pacific University, Washington School Research Center.

ACT. (2011). The condition of college and career readiness. Retrieved from: https://www.act.org/research/policymakers/cccr11/readiness1.html

Adelman, C. (2004). Principal indicators of student academic histories in post-secondary education. 1972-2000. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Aldenderfer, M. S., & Blashfield, R. K., (1984). *Cluster Analysis. Sage University Paper series on Quantitative Applications in the Social Sciences,* 07-011, Beverly Hills, CA: Sage.

Alt, M. N., & Peter, K. (2002). Private schools: A brief portrait *(NCES No. 2002–013).* Washington, DC: U.S. Department of Education.

American Institutes for Research, (2010). Finishing the First Lap: The cost of first-year student attrition in America's four-year colleges and universities. Retrieved from: http://www.air.org/files/AIR_Schneider_Finishing_the_First_Lap_Oct101.pdf.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. In L.S. Cermak, & F.I.M. Craik (Eds.), *Levels of Processing in Human Memory.* Hillsdale, NJ: Erlbaum.

Archibald, R. B., & Feldman, D. H. (2008). Explaining increases in higher education costs. *Journal of Higher Education, 79(3),* 268-295.

Baeck, P. & Van den Poel, D. (2012). Including the salesperson effect in purchasing behavior using PROC GLIMMIX. *Sas Global Forum 2012,* (Paper no. 350-2012.)

Baker, B. D. (2012). Revisiting the age-old question: Does money matter in education? *Albert Shanker Institute.* Retrieved from: http://www.shankerinstitute.org/images/doesmoneymatter_final.pdf

Bandura, A. (1977). Self-Efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84(2),* 191-215.

Bandura, A. (1994). Self-efficacy. In V. S. R. (Ed.), *Encyclopedia of Human Behavior* (pp. 71-81). New York, NY: Academic Press.

Bandura, A., Zimmerman, B. J., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29(3),* 663-676.

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

Bean, J. P., & Eaton, S. B. (2001-2002). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory & Practice, 3(1),* 73-89.

Birnie-Lefcovitch, S. (2000). Student perceptions of the transition from high school to university: Implications for preventative programming. *Journal of the First-Year Experience and Students in Transition, 12,* 61-88.

Bers, T. H., & Smith, K. E. (1991). Persistence of community college students: The influence of student intent and academic and social integration. *Research in Higher Education, 32(5),* 539-556.

Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning.* Maidenhead, Berkshire: Open University Press.

Bonwell, C., & Eison, J. (1991). Active Learning: Creating excitement in the classroom. *Ashe-Eric Higher Education Report (No. 1.)* Washington, D.C.: Jossey-Bass.

Braun, H., Jenkins, F., & Grigg, W. (2006). Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling (NCES 2006-461). Washington, DC: National Center for Education Statistics.

Braunstein, A., McGrath, M., & Pescatrice, D. (2000-2001). Measuring the impact of financial factors on college persistence. *Journal of College Student Retention: Research, Theory and Practice, 2(3),* 191-203.

Braxton, J. M. (1999). Theory elaboration and research and development: Toward a fuller understanding of college student retention. *Journal of College Student Retention: Research Theory & Practice, 1(2),* 93-97.

Braxton, J., Hirschy, A., & McClendon, S. (2004). Understanding and reducing college student departure. Ashe-Eric Higher Education Report (No. 30). San Francisco, CA: Wiley.

Braxton, J. M., Sullivan, A. S., & Johnson, R. M. (1997). Appraising Tinto's theory of college student departure. In J. C. Smart (Ed.), *Higher education: A handbook of theory and research* (pp. 107-164). New York, NY: Agathon Press.

Brewer, S. A., & Yucedag-Ozcan, A. (2012-2013). Educational persistence, self-efficacy and topics in a college orientation course. *Journal of College Student Retention: Research, Theory & Practice, 14(4),* 451-465.

Brown v. Board of Education, 347 U.S. 483. (1954).

Burney, V. H., & Beilke, J. R. (2008). The constraints of poverty on high achievement. *Journal for the Education of the Gifted, 31(3),* 171-197.

Caldas, S. J., & Bankston, C. L. (1999). A multilevel examination of student, school, and district-level effects on academic achievement. *Journal of Educational Research, 93,* 91–100.

Camara, W., Kimmel, E., Scheuneman, J., & Sawtell, E. (2003). Whose Grades Are Inflated? (Research Report 2003-4). College Board: New York, NY.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38,* 476-506.

Campbell, R., Hombo, C.M., & Mazzeo, J. (2000). NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance. Washington, DC: U.S. Department of Education.

Campbell, C. M. & Mislevy, J. L. (2012-2013). Early student perceptions matter: Early signs of undergraduate student retention/attrition. *Journal of College Student Retention: Research, Theory & Practice, 14(4),* 467-493.

Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in the Schools, 42(2),* 197-205.

Chowning, K., & Campbell, N. J. (2009). Development and validation of a measure of academic entitlement: Individual differences in students' externalized responsibility and entitled expectations. *Journal of Educational Psychology, 101(4),* 982-997.

Choy, S., & Carroll, C. D. (2003). How families of low and middle-income undergraduates pay for college: Full-time dependent students in 1999–2000, *(NCES Report no. 2003–162).* Washington, DC: U.S. Department of Education,

Cloud, J. (2007). Are Private Schools Really Better? *Time*. Retrieved from: http://content.time.com/time/nation/article/0,8599,1670063,00.html

Coffman, D. L., & Gilligan, T. D. (2002-2003). Social support, stress, and self-efficacy: Effects on student's satisfaction. *Journal of College Student Retention: Research, Theory & Practice, 4(1),* 53-66.

Coleman, J.S. et al. (1966). Equality of Educational Opportunity. Washington, DC: U.S. Government Printing Office.

College Board. (2005). Education pays update 2005: A supplement to education pays 2004: The benefits of higher education for individuals and society. Washington, DC: Retrieved April 14, 2014, from http://www.collegeboard.com/prod_downloads/press/cost05/education_pays_05.pdf/

College Board. (2007a). Trends in college pricing. Trends in Higher Education Series. College Board. Retrieved from: www.collegeboard.com/prod_downloads/press/cost06/trends_college_pricing_06.pdf.

Corno, L., & Mansinach, E. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18,* 88-108.

Connell, J. P. (1990). Context, self, and action: a motivational analysis of self-system processes across the life-span. In D. Cicchetti (Ed.), *The self in transition: From infancy to childhood* (pp. 61-97). Chicago, IL: University of Chicago Press.

Connor, S. L., Daugherty, D. A., & Gilmore, M. N. (2012-2013). Student retention and persistence to graduation: effects of an introductory life calling course. *Journal of College Student Retention: Research, Theory & Practice, 14(2),* 251-263.

Cotton, K. (1996). School size, school climate, and student performance. *School Improvement Research Series, Close-up #20*. Retrieved April, 6, 2014 from http://www.nwrel.org/scpd/sirs/10/c020.html.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Davidson, W. B., Beck, H. P., & Milligan, M. (2009). The college persistence questionnaire: Development and validation of an instrument that predicts student attrition. *Journal of College Student Development, 50(4),* 373-390.

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review, 18(3),* 375-390.

Deci, E., & Ryan, R. M. (2000). What is the self in self-directed learning? Findings from recent motivational research. In G. Staka (Ed.), *Conceptions of self-directed learning: Theoretical and conceptual considerations).* Munster: Waxmann.

Dettmers, S., Trautwein, U., Ludtke, I., Goetz, T., Frenzel, A. C., & Pekrun, R. (2011). Student's emotions during homework in mathematics: Testing a theoretical model of antecedents and achievement outcomes. *Contemporary Educational Psychology, 36(1),* 25-35.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics, 4,* 95–104.

Eccles, J. S. (2004). Schools, academic motivation, and stage-environment fit. In R. M. Lerner, & L. Steinberg (Eds.), *Handbook of adolescent psychology (2nd ed.).* (pp. 125-153) Hoboken, NJ: Wiley.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A. & Willms, J. D. (2001). The class size controversy, *CHERI (Working Paper #14).* Retrieved April 22 2014, from Cornell University, ILR School site: http://digitalcommons.ilr.cornell.edu/workingpapers/25/

El-Hamdouchi, A., & Willett, P. (1987). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal, 32(3),* 220-227.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72,* 218-232.

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy and inventories. *Educational Psychology Review, 16,* 325-345.

Espensade, T. J., & Chung, C. Y. (2011). Standardized Admission Tests, College Performance, and Campus Diversity. In J. A. Soares (Ed*.), Rethinking Admissions for a New Millennium: Moving Past the SAT for Social Diversity and Academic Excellence.* New York, NY: Teachers College Press.

Feldt, R. C. (2012). Social cognitive model of college satisfaction: A test of measurement and path models. *College Student Journal, 46(4),* 812-822.

Fisher, R. (1918). Studies in crop variation: An examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science, 11,* 107-135.

Floyd, K.S., Harrington, S.J. & Santiago, J. (2009). The effect of engagement and perceived course value on deep and surface learning strategies. Informing Science. *The International Journal of an Emerging Transdiscipline, 12,* 181-190.

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of your-year college outcomes. *Center for Studies in Higher Education (CSHE) Research & Occasional Paper Series: CSHE.6.07*.

Geiser, S., & Studley, R. (2004). UC and the SAT: Predictive validity and differential impact of the SAT and SAT II at the University of California. In R. Zwick (Ed)., *Rethinking the SAT: The future of standardized testing in university admissions*. New York, NY: Routledge/Falmer.

Geisinger, K. (June, 2009). A college admission question: What would we do if the ACT and SAT did not exist? *NCME newsletter, 17(2).* Retrieved January 23, 2014, from http://www.ncme.org/pubs/pdf/vol_17_num_2.pdf.

Godfrey, K. (2011). Investigating grade inflation and non-equivalence. *College Board Research Report (no. 11b-3181).* New York: College Board.

Goenner, C. F., Harris, M., & Pauls, K. (2013). Survival of the fittest: What do early behaviors tell us about student outcomes? *Journal of College Student Development, 54(1),* 43-61.

Greenwald, R., Hedges, L., & Laine, R. (1996). The Effect of School Resources on Student Achievement. *Review of Educational Research, 66(3),* 361-396.

Hanushek, E.A. (1986). Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature, 24(3),* 1141-1177.

Hanushek, E. (1999). The evidence on class size. In S. E. Mayer & P. Peterson (Eds.), *Earning and Learning: How Schools Matter* (pp. 131-168). Washington, DC: Brookings Institution Press.

Harnisch, D. L. (1987). Characteristics associated with effective public high schools. *Journal of Educational Research, 80,* 233-240.

Hermanowicz, J. C. (2006-2007). Reasons and reasoning for leaving college among the academic elite: Case study findings and implications. *Journal of College Student Retention: Research, Theory & Practice, 8(1),* 21-38.

Herrmann, K. J. (2013). The impact of cooperative learning on student engagement: Results from an intervention. *Active Learning in Higher Education, 14(3),* 175-187.

Hochstein, S. K., & Butler, R. R. (1983). The effects of the composition of a financial aid package on student retention. *Journal of Student Financial Aid, 13(1),* 21-26.

Hope, E. C., Tabbye, M. C., Jagers, R. J., & Sellers, R. M. (2013). Connecting self-esteem and achievement: diversity in academic identification and dis-identification patterns among black college students. *American Educational Research Journal, 50(5),* 1122-1151.

Hornak, A. M., Farrell, P. L., & Jackson, N. J. (2008). Making it (or not) on a dime in college: Implications for practice. *Journal of College Student Development, 51(5),* 481-495.

Horowitz, J. B., & Spector, L. (2005). Is there a difference between private and public education on college performance. *Economics of Education Review, 24,* 189-195.

Ishitani, T. T., & DesJardins, S. L. (2002-2003). A longitudinal investigation of dropout from college in the United States. *Journal of College Student Retention: Research, Theory & Practice, 4(2),* 173-201.

Jernigan, C. G. (2004). What do students expect to learn? The role of learner expectancies, beliefs, and attributions for success and failure in student motivation. *Current Issues in Education, 7(4),* available online at http://cie.asu.edu/volume7/number4/index.html.

Johnson D.W., & Johnson R.T. (1989). *Cooperation and Competition: Theory and Research.* Edina, MN: Interaction Book Company.

Johnson, D.W., & Johnson, R.T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38(5),* 365–79.

Kassens-Noor, E. (2012). Twitter as a teaching practice to enhance active and informal learning in higher education: The case of sustainable tweets. *Active Learning in Higher Education, 13,* 9-21.

Kline, P. (2000). *The Handbook of Psychological Testing (2nd ed.).* London: Routledge.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT for predicting first-year college grade point average. *College Board Research Report (No. 2008-5).* New York: College Board.

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

Kuiper, F. K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics, 31,* 777-783.

Lang, K. B. (2012). The similarities and differences between working and non-working students at a mid-sized American public university. *College Student Journal, 46(2),* 243-255.

Lee, V. E., & Burkham, D. T. (2002*). Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School.* Washington DC: Economic Policy Institute.

Lent, R. W., Brown, S. D., & Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology, 33(3),* 265-269.

Levine, A., & Cureton, J. (1998). Flaws, problems and decline: The new localism. In A. Levine & J. Cureton (Eds.), *When hope and fear collide: A portrait of today's college student.* San Francisco, CA: Jossey-Bass.

Liu, Y., Li, Z., Xiong, H., Goa, X., & Wu, J. (2010). Understanding of internal clustering validation measures Proceedings of the 10th International Conference on Data Mining (ICDM). Sidney, AU.

Lubienski, C., & Lubienski, S.T. (2006). *Charter, Private, Public Schools and Academic Achievement: New Evidence From NAEP Mathematics Data.* New York, NY: Columbia University Press.

Lynch, D. J. (2008). Confronting Challenges: Motivational Beliefs and Learning Strategies in Difficult College Courses. *College Student Journal, 42(2),* 416-421.

MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms.* New York, NY: Cambridge University Press.

MacQueen, J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.

Martin, P., Karabel, J., & Vasquez, S. (2005). High school segregation and access to the University of California. *Educational Policy, 19,* 308–330.

Massart, D.L. & Kaufman, L. (1983*), The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis.* New York, NY: John Wiley & Sons, Inc.

Milligan, G. W. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika*, *45,* 325–342.

Milligan, G. W., & Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50, 159–179.

Mounsey, R., Vandehey, M. A., & Diekhoff, G. M. (2013). Working and non-working university students: Anxiety, depression and grade point average. *College Student Journal, 47(2),* 279-389.

Morris, C., & Gladson, C. (2013). Screencasts: How effective are they and how to students engage with them? *Active Learning in Higher Education, 15,* 25-37.

Morrow, J. A. (2012). Intention to persist and retention of first-year students: The importance of motivation and sense of belonging. *College Student Journal, 46(3),* 483-491.

Munt, J. A., & Merydith, S. P. (2011-2012). The relationship of student's personality traits and psychosocial characteristics with academic retention. *Journal of College Student Retention: Research, Theory & Practice, 13(4),* 457-478.

National Center for Educational Statistics, (2011). Youth Indicators 2011, America's Youth: Transitions to Adulthood, National Center for Educational Statistics. (Report No. NCES 2012-026). Retrieved January 3, 2014 from: http://nces.ed.gov/pubs2012/2012026/index.asp

Noble, J. P., & Sawyer, R. L. (2004). Is high school GPA better than admission test scores for predicting academic success in college. *College and University Journal, 79(4),* 17-22.

Onwuegbuzie, A.J., & DaRos-Voseles, D. A. (2001). The role of cooperative learning in research methodology courses: A mixed-methods analysis. *Research in the Schools, 8(1),* 61–75.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *Journal of Higher Education, 51,* 60-75.

Pascarella, E. T., & Terenzini, P. T. (2005*). How college affects students (Vol. 2): A third decade of research.* San Francisco, CA: Jossey-Bass.

Paulsen, M. B., & St. John, E. P. (1997). The financial nexus between college choice and persistence. In R. A. Voorhees (Ed.), *Researching student aid: Creating an action agenda,* (pp.65-82). San Francisco: Jossey-Bass.

Paulsen, M., & St. John, E. P. (2002). Social class and college costs: Examining the financial nexus between college choice and persistence. *Journal of Higher Education, 73(2),* 189-236.

Peña, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization for the K-means algorithm. *Pattern Recognition Letters, 20,* 1027-1040.

Pekrun, R. Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire. *Contemporary Educational Psychology, 36(1),* 36-48.

Pew Research Center. (2014). The rising cost of not going to college. Retrieved April 14, 2014 from http://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology,* 33-40.

Pleitz, J., Terry, R., Campbell, N. J., & Fife, D. (2011). Rethinking at-risk. In R. Hayes (Ed.), *Proceedings of the 7th Annual National Symposium on Student Retention* (pp. 178-186). Norman, OK: The University of Oklahoma.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93(3),* 223-232.

Plessy v. Ferguson. 163 U.S. 537 (1896).

Ramos-Sanchez, L., & Nichols, L. (2007). Self-efficacy of first-generation and non-first-generation college students: The relationship with academic performance and college adjustment. *Journal of College Counseling, 10(1),* 6-18.

Raudenbush, S. W., & Bryk, A. S. (2012). *Hierarchical linear models: Applications and data analysis methods (Advanced Quantitative Techniques in the Social Sciences).* Thousand Oaks: Sage Publications.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846-850.

Reason, R. D. (2003). Using an ACT-based merit-index to predict between-year retention. *Journal of College Student Retention: Research, Theory & Practice, 5(1),* 71-87.

Riconscente, M. M., & Seli, H. (2012). Predicting end-of-semester interest and self-efficacy from course and professor ratings. *College Student Journal, 46(1),* 34-39.

Robb, C. A., Moody, B. M., & Abdel-Ghany, M. (2011-2012). College student persistence to degree: The burden of debt. *Journal of College Student Retention: Research, Theory & Practice, 13(4),* 431-456.

Rocca, K.A. (2010). Student participation in the college classroom: An extended multidisciplinary literature review. *Communication Education, 59(2),* 185–213.

Rocconi, L. (2013). Analyzing multilevel data: comparing findings from hierarchical linear modeling and ordinary least squares regression. *Higher Education, 66(4),* 439-461.

Sarle, W. S. (1982), "Cluster Analysis by Least Squares," *Proceedings of the Seventh Annual SAS Users Group International Conference,* 651–653.

Sarle, W. S. (1983). Cubic clustering criterion (Tech. Rep. A-108). Cary, N.C.: SAS Institute.

Sawyer, R. (2010). Usefulness of high school average and ACT scores in making college admissions decisions. *(ACT Research Report Series 2010-2).* Iowa City, IA: ACT.

Schuh, J. H. (1999-2000). Examining the effects of scholarships on retention in a fine arts college. *Journal of College Student Retention: Research, Theory & Practice, 1(3),* 193-202.

Schunk, D. H. **(**1991**)**. Self-efficacy and academic motivation**.** *Educational Psychologist, 26(3-4),* 207-231**.**

Sedaghat, M, A., Alizera, A., Hejazi, E., & Hamidreza, H. (2011). *Procedia-Social and Behavioral Sciences, 15,* 2406-2410.

Shapiro, D., Dundar, A., Ziskin, M., Yuan, X., & Harrell, A. (2013). Completing College: a national view of Student Attainment Rates—Fall 2007 Cohort. National Student Clearing House Research Center, Signature Report 6. Retrieved February 24, 2014 from: http://nscresearchcenter.org/wpcontent/uploads/NSC_Signature_Report_6.pdf.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24(4),* 323-355.

Slate, J. R., & Jones, C. H. (2005). Effects of school size: A review of the literature with recommendations. *Essays in Education, (13).* Retrieved April 14, 2014, from http://www.usca.edu/essays/vol132005/slate.pdf

Solberg, V. S., O'Brien, K., Villareal, P., Kennel, R., & Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the College Self-Efficacy Instrument. *Hispanic Journal of Behavioral Sciences, 15,* 80-95.

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

Sparkman, L. A., Maulding, W. S., & Roberts, J. G. (2012). Non-Cognitive predictors of student success in college. *College Student Journal, 46(3),* 653-664.

St John, E. P., Paulsen, M. B., & Carter, D. F. (2005). Diversity, college costs, and postsecondary opportunity: An examination of the financial nexus between college choice and persistence for African Americans and whites. *The Journal of Higher Education, 76(5),* 545-569.

Stampen, J. O., & Cabrera, A. F. (1988). The targeting and packaging of student aid and its effect on attrition. *Economics of Education Review, 7(1),* 29-46.

Strauss, L. C., & Volkwein, J. F. (2004). Predictors of student commitment at two-year and four-year institutions*. The Journal of Higher Education, 75(2),* 203-227.

Suarez-Orozco, C., Pimentel, A., & Martin, M. (2011). The significance of relationships: Academic engagement and achievement among newcomer immigrant youth*. Teachers College Record, 111(3),* 712-749.

Sum, A., Khatiwada, I. McLaughlin, J. & Palma, S. (2009). The consequences of dropping out of high school. Retrieved from: http://www.northeastern.edu/clms/wpcontent/uploads/The_Consequences_of_Dropping_Out_of_High_School.pdf

Tabachnick, B. G., & Fiddell, S. L. (2012). *Using Multivariate Statistics (6th Ed.).* New York: Pearson.

Tierney, W.G., & Venegas, K.M. (2007). The cultural ecology of financial aid. *Readings on Equal Education, 22,* 1–35.

Tierney, W. G., & Venegas, K. M. (2010). Finding money on the table: Information, financial aid, and access to college*. The Journal of Higher Education, 80(4)*, 363-388.

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

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

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

U.S. Census Bureau. (2010). Educational Attainment by Selected Characteristics. Retrieved January 3, 2014 from: http://www.census.gov/population/www/socdemo/educ-attn.html

U.S. Department of Education, Office of Educational Research and Improvement. (1999). Schools and Staffing Survey, 1999-2000: Overview of the data for public, private, public charter and bureau of Indian affairs elementary and secondary schools. (Report no. NCES 2002-313). Retrieved from: http://nces.ed.gov/surveys/sass/dataprod9901.asp.

U.S. Government Accountability Office. (2002). Per-Pupil spending differences between selected inner city and suburban schools varied by metropolitan area. (Report No. GOA-03-234). Retrieved from: www.gao.gov/cgi-bin/getrpt?GAO-03-234.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research, 78(4),* 751-796.

Voelkl, K. E. (1997). Identification with school. *American Journal of Education, 105,* 294-319.

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

Wang, X., & Kennedy-Phillips, L. (2013). Association with the academic and social involvement of second-year college students. *Journal of College Student Development, 54(5),* 541-548.

Wang, M., & Eccles, J. (2013). School contest, academic motivation, and academic engagement. A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction, 28,* 12-23.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58,* 236–244.

Weaver, R. R., & Qi, J. (2005). Classroom organization and participation: College students' perceptions. *Journal of Higher Education, 76(5),* 570–601.

Wenglinsky, H. (1997) How Money Matters: The effect of school district spending on academic achievement. *Sociology of Education, 70(3),* 221-237.

Wenglinsky, H. (2007). Are private high schools better academically than public high schools? Center on Education Policy. Retrieved from: http://www.cepdc.org/cfcontent_file.cfm?Attachment=Wenglinsky%5FReport%5FPrivateSchool%5F101007%2Epdf

Woodruff, D. J., & Ziomek, R. L. (2004). High School Grade Inflation from 1991 to 2003. (Research Report Series 2002-04), Iowa City, IA: ACT, Inc.

Woosley, S. A., & Miller, A. L. (2009). Integration and institutional commitment as predictors of college student transition: Are third week indicators significant? *College Student Journal, 43(4),* 1260-1271.

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

Zwick, R., & Sklar, J. C. (2005). Predicting College Grades and Degree Completion Using High School Grades and SAT Scores: The Role of Student Ethnicity and First Language. *American Educational Research Journal, 42(3),* 439-464.

Zubin, J. A. (1938). A technique for measuring likemindedness. *Journal of Abnormal and Social Psychology, 33,* 508-516.

APPENDIX 1: Summary of Variables

Table 21: Variables in the Current Analysis

Variable Name	Variable Description
Need	I need to work to afford to go to school
Aid Rec	How important was financial aid received in your decision to attend OU
Money	How difficult do you think it will be having enough money during your first year
Costs	How important were the costs in your decision to attend OU
Resource	I am worried about having enough financial resources
HsStudy	While in high school, the amount of hours I spent studying outside of the class room was:
NewQuestion	A lot of students have to work hard to get good grades in high school while other students do not. Which best describes you?
Worked Harder	I feel that I worked harder than most students while in high school*
Challenged	While in high school, I was challenged to do my best academic work*
Rarely	While in high school, I rarely studied outside of class
Wentto	While in high school, I went to class without doing the assigned reading
Homework	While in high school, I went to class without doing the assigned homework
Late	While in high school, I came late to class
Bored	While in high school, I felt bored in class
Skip	While in high school I skipped class
WaitedAssignments	While in high school, I wanted until the last minute to do my assignments
WaitedExam	While in high school, I waited until the last minute to study for an exam.
Doubts	On occasion, I have had doubts about my ability to succeed in in life*
ConfidentSucceed	I am confident in my abilities to succeed at OU
ConfidenceAbilities	I have confidence in my abilities to succeed at OU
Dowell	I am worried about my ability to do well academically
Important	It is important that I graduate from OU as opposed to another university
ConfidentOu	I am confident I made the correct choice when choosing to attend OU*
OuChoice	In selecting a college, OU was my _____ choice
Transfer	I plan to transfer from OU

Table 22: Correlations Amongst School-Level Variables

	Free Lunch	Class Size	Sector	Grad Rate	Student Teacher	District Spending	City Type
Free Lunch	1.00	-	-	-	-	-	
Class Size	-0.164	1.00	-	-	-	-	-
Sector	-0.496	-0.391	1.00	-	-	-	-
Grad Rate	-0.578	-0.071	0.452	1.00	-	-	-
Student Teacher	0.174	0.233	-0.324	-0.171	1.00	-	-
District Spending	-0.267	0.075	0.074	0.202	-0.057	1.00	-
City Type	-0.188	0.086	0.367	0.214	-0.069	0.085	1.00

APPENDIX 2: Additional Canonical Variate Graphs

Figure 5: Canonical Variate Axis 2 by Graduation Rate

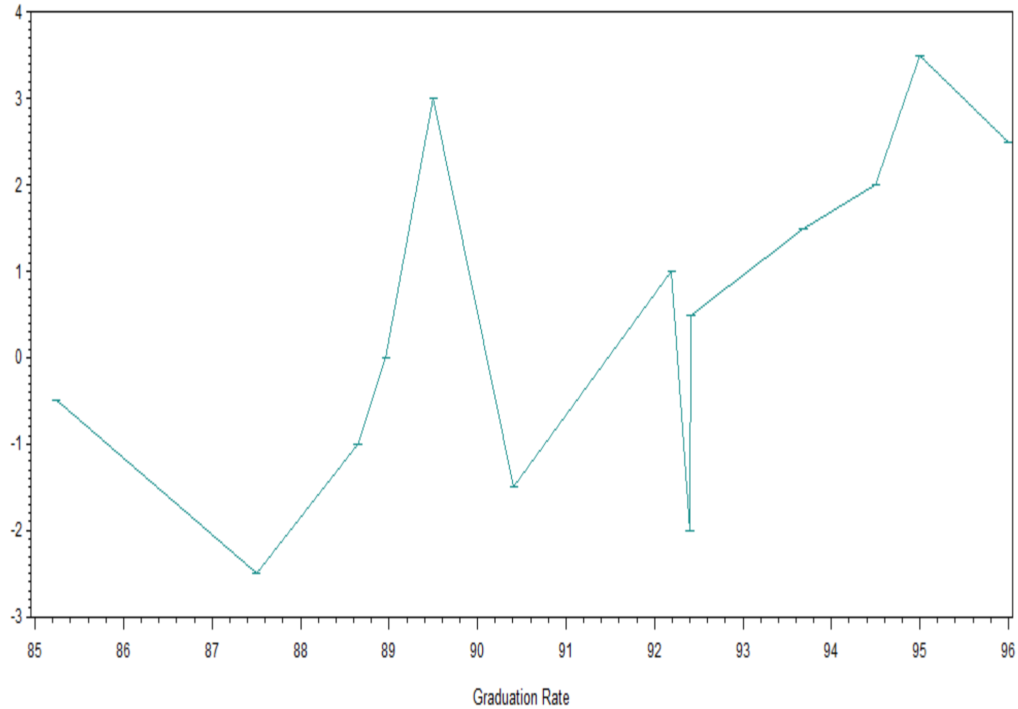


Figure 6: Canonical Variate Axis 2 by Free Lunch

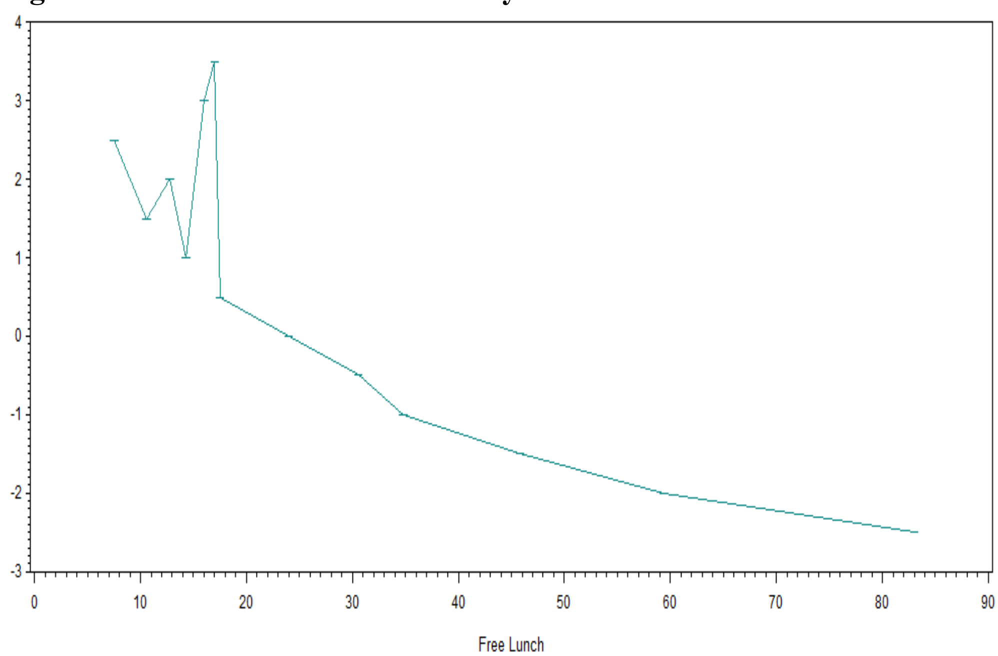
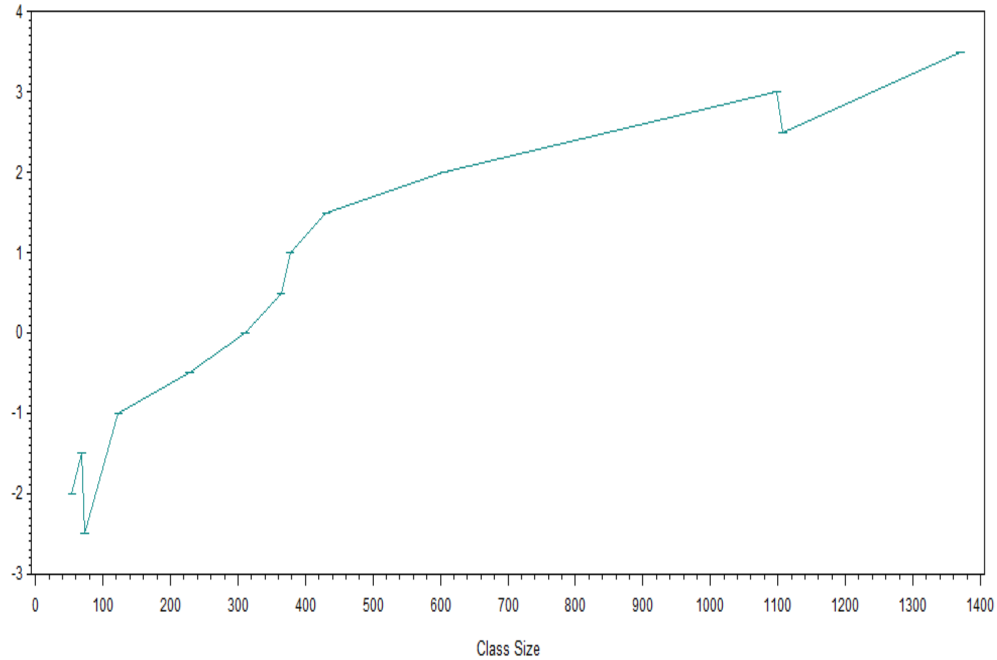


Figure 7: Canonical Variate Axis 2 by Class Size



Appendix 3: Results and Description of Model 5

Because district spending was such a heavy influence in creating the clusters, an additional model (Model 5) was created to examine how using district spending and free lunch at the school level, and the optimal student model at the student level influenced predictive accuracy. Results from this model indicated that, although district spending itself was not significant, using district spending and free lunch at the school level, in place of the clusters, did significantly contribute towards improving model fit. Specifically, results showed that variables for free lunch eligibility, t (731) = -4.29, p < .0001, academic engagement, t (1933) = 3.27, p < .0001, financial concerns, t (1933) = -5.12, p < .0001, ACT scores, t (1933) = 3.94, p < .0001, high school GPA, t (1933) = 7.02, p < .0001, and alumni ties t (1933) = -1.92, p < .0001, each significantly contributed towards improving model accuracy. A summary of results is presented in Table 23 below.

Table 23: Summary of Results from Model 5

Effect	Estimate	Standard Error	Df	t value	Pr > |t|
Intercept	2.2061	0.2714	731	8.13	< .0001
Free Lunch	-0.013	0.0031	731	-4.29	< .0001
District Spending	0.000	0.0002	731	0.31	0.7584
Academic Engagement	0.1920	0.0586	1933	3.27	< .0001
Financial Concerns	-0.2995	0.5784	1933	-5.12	< .0001
ACT Score	0.2686	0.0681	1933	3.94	< .0001
High School GPA	0.4086	0.0582	1933	7.02	< .0001
Alumni Ties	-0.2385	0.1245	1933	-1.92	0.0556

The estimate of the variance in intercepts indicated that the model intercept variance = 0.1842 (SE=0.1133), placing the variance somewhere between Model 3 and Model 4. As such, it seems like using district spending is a good proxy for the clusters, although it does not predict as accurately as using the aggregate variables. Examination of the fit statistics indicated that -2 Res Log Pseudo-likelihood = 13232.13, the Pseudo-AIC = 13250.13, and the Pseudo-BIC = 13303.12. These results again confirm that using this model does not fit as well as using the aggregate variables, however it does provide a greater fit than using the clusters.

Appendix 4: Missing and Final Data File Comparisons

Because the data were cleaned and reduced to the final file, it is helpful to examine the descriptive statistics between the two files. Tables 24 and 25 below present the descriptive statistics from the full and final file. As can be seen, reducing the variables did not significantly change the descriptive statistics between the two files.

Table 24: Descriptive Statistics from Full File

Variable	N	Mean	Std Dev	Min	Max
Academic Engagement	2789	3.44	0.58	1.00	5.16
Financial Concerns	2856	2.87	0.77	1.00	4.60
Institutional Commitment	2835	4.16	0.51	1.50	4.83
Self Efficacy	2856	2.10	0.60	1.00	4.50
ACT Score	3875	25.88	4.07	13.00	36.00
High School GPA	3882	3.59	0.33	1.70	4.00
Alumni Ties	2901	0.66	0.47	0.00	1.00
Grad Rate	931	90.61	9.68	26.00	100.00
Student Teacher	926	15.82	10.71	4.00	313.00
Free Lunch	907	25.76	21.40	0.00	100.00
District Spending	922	10924.7	319.44	2335.00	36024.00
High School Class Size	934	280.68	235.45	5.00	1433.00
City Type	935	1.82	1.89	0.00	5.00
Sector	935	0.17	0.37	0.00	1.00

Table 25: Descriptive Statistics from Final File

Variable	N	Mean	Std Dev	Min	Max
Academic Engagement	2647	3.44	0.58	1.00	5.16
Financial Concerns	2647	2.87	0.77	1.00	4.60
Institutional Commitment	2647	4.16	0.50	1.50	4.83
Self Efficacy	2647	2.09	0.60	1.00	4.50
ACT Score	2647	25.42	3.78	13.00	36.00
High School GPA	2647	3.57	0.32	2.14	4.00
Alumni Ties	2647	0.66	0.47	0.00	1.00
Grad Rate	734	91.36	7.78	26.00	100.00
Student Teacher	734	15.93	5.86	4.00	313.00
Free Lunch	734	23.73	19.56	0.00	100.00
District Spending	734	9643.07	2836.46	2335.00	36042.00
High School Class Size	734	380.02	274.04	0.00	1440.00
City Type	734	2.33	1.86	0	5.00
Sector	734	0.14	0.35	0.00	1.00