Factors related to college student success: The case of a state university in the Midwest

Kevin Paterson
University of the Incarnate Word

Adam Guerrero
University of the Incarnate Word

ABSTRACT

The current research examines the determinants of college student retention for first-time freshmen enrolled full-time at a moderately selective regionally accredited four-year public university in the Midwest. Before elaborating upon the methodological results estimated using models for categorical dependent variables, a review of the extant literature that integrates insights gleaned from sociological, psychological, and economic perspectives is provided. Key contributions to the literature include a unique analysis sample, a unique set of independent variables, and statistical estimation methodologies that build upon traditional models of college student persistence to include machine learning techniques. Furthermore, multivariate models are expanded to include a program evaluation of academic learning community membership with particular emphasis on how leaders in higher education can leverage analytical results to improve upon data-informed decision-making.

Keywords: higher education, student success, college retention, student persistence

INTRODUCTION

63.4 percent of first-time freshmen (FTF) were enrolled in 4-year institutions designated as "Public" in 2018, with 74.8 percent retained after one year (NCES, 2020b). Retention and graduation rates are ranking factors that differentiate colleges and influence where students apply for admission and how institutions make their selections (Sanoff et al., 2007). Student persistence is still one of the most widely studied areas in higher education, leading to higher retention and graduation rates for students and heightened revenue for colleges (Tinto, 2006, 2017). Factors can predict whether FTF applicants will be successful, including scores on standardized tests, which is one of the primary criteria for admissions (Cai, 2020; Kuh et al., 2006). In addition, explanatory sociological, psychological, and economic variables inform students and administrators why FTF students do not persist (Siedman, 2005).

Statement of the Problem

In the United States, four-year degrees (typically bachelor's degrees) are offered by various public, nonprofit, or for-profit institutions, thereby providing prospective students with a wide range of choices. The total number of four-year institutions admitting first-year undergraduate students in the United States in 2019 was 2,330 (NCES, 2020a). Of these 2,330 institutions, 730 were designated public institutions and competed for the same students as their private and for-profit counterparts. Over the years, there has been a noticeable increase in public and nonprofit institutions, while for-profit institutions have decreased in number. Moreover, the under-18 years old population in the United States has been decreasing over the past decade, as demonstrated by Ogunwole et al. (2021), which perpetuates the need for higher education institutions to offer a competitive landscape for attracting first-time freshmen (FTF) and have them persist through graduation. Therefore, improving the retention rates of FTF students provides an institution of higher education with a competitive advantage and remains an essential topic of inquiry.

Purpose Statement

This study aims to examine methods that ascertain the results for explaining and predicting student persistence. It will contrast them as to which can better determine retention among FTF students, particularly in a moderately selective accredited Midwestern regional public 4-year institution not including international students (MDHE, 2022). First, this paper discusses the history of explanatory variables in the context of 4-year institutions in the US and presents evidence from the literature about how sociological, psychological, and economic perspectives help explain student persistence. After that, various analyses regarding the appropriate use of traditional models and a machine learning method that include explanatory variables from each category, including sociological, psychological, and economic, are discussed based on the existing literature to offer suggestions on how public institutions can utilize them to impact their retention rates positively.

Definitions

Institute of higher education – an institution that offers a post-high school level of education that awards either a regional or national accredited 4-year degree in the United States.

First-time freshmen – students are attending an institute of higher education in the United States for the first time.

Student persistence – a student, remaining enrolled or retained, to the next year at the same institute of higher education.

International students – "non-immigrant" visitors who come to the United States temporarily to take classes or online courses virtually.

Moderately selective – first-time, full-time students and transfer students with a combined percentile of high school percentile rank and student ACT composite score greater than or equal to 100, or an ACT score of at least 21 or equivalent score on the SAT.

Limitations/Delimitations

The study is based on ex post facto data collected from a moderately selective accredited Midwestern regional public 4-year institution of higher education for the periods 2017-2018 and 2018-2019. These periods were covered in the time provided to prepare this research study for the Doctor of Business Administration culminating experience courses at the University of the Incarnate Word while still having an adequate sample size that can be studied for impact on student retention rates. This study also makes a case for examining whether the machine learning algorithm, discriminant analysis (DA), predicts better student persistence. In addition, further investigation of other machine learning methods could be explored, including regularized logistic regression.

A REVIEW OF THE LITERATURE

Due to the complexity of student retention and the limitations of earlier models, many models have developed across sociological, psychological, and economic perspectives (Tinto, 2006).

Sociological Perspective

The sociological perspective is grounded in Durkheim's (1951) theory of suicide and the role of social integration. Spady (1970) and Tinto (1975) argued that a student's decision to drop out of college is like an individual's decision to commit suicide. Both decisions require the individual to break their social ties from a social structure, likely motivated by low levels of social integration. In contrast, a student whose academic and social values align closely with an institution's increases their level of academic and social integration and the likelihood of persisting (Spady, 1970; Tinto, 1975, 1993). Tinto's (1975) interactionalist theory is based on a longitudinal model concerned with the quality of students' academic and social interactions. Student background characteristics, including gender, race, primary and secondary education experiences, aptitude, social capital, and others, shape their commitment to obtain a degree at a particular college. Over time, a student's commitment to getting their degree is strengthened or weakened depending on their academic and social integration level. For example, suppose a student’s social integration with their peers deepens to the point that their ability to study is inadequate. In that case, this can increase the student's probability of dropping out of college. Tinto’s interactionalist theory remains a seminal study widely cited due to the positive effect of background characteristic variables that explain a student’s decision to persist (Reason, 2003).

The background characteristic variables have expanded Tinto’s model to include other contextual variables, including classroom experience and size. For example, Langbein and Snider (1999) investigated the relationship between college course evaluations and student persistence. This added the dimension of classroom experiences to the model, which influences students' academic and social integration levels. Moreover, first-year students' classroom experiences may be affected by classroom size (Montmarquette et al., 2001).

Psychological Perspective

Researchers have primarily depended on sociological models to explain student retention (Bean & Eaton, 2000). As a result, psychological theories developed due to the complexity surrounding student retention and earlier model limitations. For example, Astin’s (1999) theory of involvement, initially published in 1984, found that the amount of psychological energy spent on their educational experience influences their academic and social participation level.

The modification of Mueller and Price’s (1981) model of employee turnover in a for-profit organization by Bean (1980, 1983) resulted in the industrial model of student attrition. Student persistence is like employee turnover in market-based organizations and includes psychological concepts such as student satisfaction, perceptions, and intentions. Indeed, Bean’s model has ten exogenous variables that affect student satisfaction. Five variables match Price and Muller’s model, including routinization, participation, instrumental communication, integration, and distributive justice. The remaining five are a modification and extension of Price and Muller’s model, including grades, practical value, development, courses, and membership in campus organizations. When a student’s external environment, including marital status and transfer opportunities, is combined with the ten exogenous variables, a student’s level of satisfaction affects their intentions which jeopardizes their likelihood of persisting (Bean, 1983).

Bean and Eaton (2000) provided the foundation for understanding the decision to drop out of college in terms of psychological theories and determinants. These theories include attitude-behavior theory, self-efficacy theory, and attribution theory (Bean, 2005; Bean & Eaton, 2000). Specifically, Fishbein and Ajzen (1975) attitude-behavior proved that beliefs lead to behaviors that affect student departure. In addition, Bandura (1997) demonstrated a direct relationship between self-efficacy and goal accomplishment, including student persistence to the next academic year. Finally, attribution theory and locus of control play an integral role in the psychological perspective. Students with an internal locus of control can perceive how their attributes, including aptitude and hard work, contribute to outcomes (Bean & Eaton, 2000). This contrasts with a student having an external locus, which leaves them believing that outcomes are out of their control, leaving them less likely to respond to a given situation. To summarize, Bean and Eaton’s psychological model of student retention demonstrated that students enter college with diverse personal characteristics. As each student interacts with their academic environment, psychological processes occur that for successful students result in positive self-efficacy, reduced stress, increased efficiency, and internal locus of control, leading to academic and social integration and student persistence (Bean & Eaton, 2000).

Psychosociological Integrated Perspective

Tinto’s interactionalist theory, the student integration model, is significant in understanding external factors, including others’ influence, finances, and college preparedness on student persistence (Cabrera et al., 1992; Reason, 2003). According to J. Braxton and Hirschy (2005) and Cabrera et al. (1992, 1993), the combination of Bean and Eaton's psychological model of student retention with Tinto's student integration model produced an integrated model that provided a more comprehensive understanding of the interplay between individual, environmental, and institutional factors. The integrated model of student retention consists of all the statistically confirmed variables from both theories, including GPA, goal commitment, social integration, financial attitude, encouragement from friends and family, academic integration, institutional commitment, and intent to persist (Aljohani, 2016).

Economic Perspective

Researchers have studied student persistence using an economic approach by conducting a cost-benefit analysis (J. Braxton & Hirschy, 2005). According to Becker (1992), economics is concerned with how individuals allocate limited resources like income or time to maximize their well-being regardless of their motivations. Rational utility-maximization postulates that individuals' tastes are consistent, cost calculations are correct, and they seek the highest satisfaction level from their economic decisions (McCormick, 1997). Therefore, the student weighs the current discounted benefits and costs of pursuing a particular college or university versus transferring to another institution, entering the workforce, or seeking the next-best opportunity (Kerkvliet & Nowell, 2005; Leppel, 2001). Light (1995) demonstrated that young men re-enroll in college when the costs are relatively low, and the benefits are relatively high using a semiparametric, proportional hazard model represented by variables including test scores, family income, tuition costs, and predicted earnings. Other economic approaches have been utilized to study student persistence, including price-response theory. Indeed, Singell (2004) found that students already enrolled in college are less likely to persist to the next year when the price of education increases due to reductions in financial aid.

METHODOLOGY

Studies with all the categories of variables represented, including sociological, psychological, and economic, significantly explain college student persistence. However, many colleges cannot collect student retention data from each main category. Due to insufficient permission to use it in research or other limitations, such studies focus on explanation rather than prediction, even though the ability to explain is diminished due to omitted-variable bias (OVB). For example, researchers interested in student persistence do not always have permission to use financial aid information due to access limitations due to their institution's policies, which are part of the economic perspective. Yet, according to Singell (2004), financial aid is a determinant in students enrolling in college the following year. Moreover, the ability of a student to pay for college can moderate the effects of other variables of student persistence (Cabrera et al., 1992).

Since most studies can explain only a percentage of why students leave college and are missing key categories of variables, their focus should transition from explanation to prediction since parameter coefficients in those models will be biased. However, there are still significant benefits associated with using analytical results to predict whether a student will be retained from one year to the next. In short, researchers can still get robust predictions, which can then be used to conduct cost-benefit analyses related to retention maximization or attrition minimization programs. Along these lines, it is essential for researchers to examine the predictive power of traditional student retention models compared to models that integrate machine learning algorithms.

Research Questions

The goal of this study will be to attempt to answer the following questions and identify other potential questions addressed by this study. In addition, questions identified that cannot be addressed adequately by this study should be covered by further research:

(1) What are the factors that affect student success at a moderately selective 4-year Midwestern public institution of higher education?

(2) How can the research methodology chosen uncover insights related to the explanation and prediction of college student persistence decisions?

Data Collection and Participants

A Stata program pulls the population and sample by leveraging two ex post facto datasets obtained from a moderately selective accredited Midwestern regional public 4-year institution of higher education for the periods 2017-2018 and 2018-2019 (Stata, 2022). First-time freshmen admitted for fall 2017 comprise Dataset cohort_17, including 1,210 participants. Dataset cohort_18 consists of 1,301 participants that are first-time freshmen admitted for fall 2018 identified. Of the 2,511 participants, 78 international students were filtered out since they were not the focus of this study. A sample of 2,414 participants was drawn from a population of 2,433 by removing 19 participants, less than a fraction of a percent, due to missing data. Stata is used to pool Dataset cohort_17 (variable cohort==17) and Dataset cohort_18 (variable cohort==18) into one sample analysis, Dataset nw_retention, with the provided "cohort" variable available to identify which dataset each observation originated as indicated in Table 1 (Appendix).

Data Analysis

The sample analysis, Dataset nw_retention, was pooled from Datasets cohort_17 and cohort_18 collected by the same Midwestern regional institution of higher education. The same method for data collection was used, recording the same variables for each observation, including student background characteristics. A categorical variable that indicates cohort membership was included in the final model to account for potential differences between groups. Furthermore, student background variables were examined to determine whether pooling observations from cohorts into one dataset was appropriate. No statistically significant differences between cohorts on background variables are included in the multivariable analyses below.

This study centers on models used in the literature to study college student dropout propensities, including the linear probability model and logit regression analysis. The combination of ordinary least squares (OLS) and logistic regression was used by Leppel (2001, 2005) to study the relationship between social integration and student retention in higher education. The linear probability model (LPM) revealed that the likelihood of persistence increases as students more frequently engage in collegiate activities like meeting with faculty outside of class and joining school clubs. LPM is useful but requires researchers to correct heteroscedasticity and account for other limitations, including nonsense predictions, errors not normally distributed, and the constant effect independent variables have on the dependent variable. Nevertheless, the researchers of this study will begin by exploring the linear probability model.

A logit model can be leveraged since this study is concerned with a binary dependent variable defined by whether a student persists to the next year at the same institution. Furthermore, the logit model is not susceptible to the limitations of LPM, so scholars, including Tinto (1993) and Wetzel et al. (1999), recommend using logit regression analysis to study college student retention. Logit regression analysis is based on maximum likelihood estimation (MLE), which estimates the parameters of the logistic probability distribution. It provides results that are easier to interpret and asymptotically efficient, consistent, and normally distributed estimators, which the researchers will discuss.

DESCRIPTIVE STATISTICS

First-time freshmen (FTF) consist of 2,433 non-international students from an accredited Midwestern regional public 4-year institution of higher education. For the variables included in the study, 19 students were found to have missing data, like no high school percentile rank. Listwise deletion was used to control for the minuscule percentage of missing data resulting in a complete dataset of 2,414 participants. Variable categorization and measurements with descriptive statistics are provided and included in the empirical analyses as indicated in Table 2 (Appendix).

There were 549 students who did not persist to the following year out of the 2,414 observed. This corresponds to a 22.7 percent drop rate closely correlated to public 4-year institutions with an admissions acceptance rate of 75 to 89.9 percent between 2017 and 2019 (NCES, 2020b). Variables are represented in each category, including sociological, psychological, and economical, which is helpful for the explanation of the dependent variable. In addition, like most studies on student retention, variables are missing from one or more categories. For example, the dataset for this study does not include classroom size from sociological in Montmarquette et al. (2001), instrumental communication from the psychological perspective in Bean (1983), and tuition costs from economics in Singell (2004).

RESULTS

Linear Probability Model

The linear probability model (LPM) is a linear regression model applied to categorical dependent variables. Ordinary least squares (OLS) estimate the parameters of LPM, which leverages a linear function of the independent variables. In this case, the dependent variable is “drop,” which is dichotomous and equal to one if the student does not remain enrolled for the next year at the same institution and zero otherwise. The independent variables for each category, as indicated in Table 2 (Appendix), are placed sequentially into each regression model and bound by the estimated probabilities between [0, 1], as shown in Table 3 (Appendix). For example, student background variables are included in base-case Model 1 and remain in the other models as categories of variables are added.

The base-case model has an R-squared (r2) value of 0.007, which explains less than 1 percent of the variation in the dependent variable. As a result, the base-case model is elaborated upon to include variables from the categories of determinants of college student persistence outlined in the extant literature. Specifically, academic integration variables join student background variables in Model 2 to further explain student persistence. The R-squared value is 0.349, significantly improving by adding academic integration variables to the base-case model. As social integration and economic and finance variables are added into Models 3 and 4, the R-squared remains similar, fluctuating between 0.349 and 0.352. Each variable that is statistically significant, like “HSPercentile” and “probation," remains substantial at the same level as additional categories are added, except for the variable “black” in Model 1. For example, as social integration variables are added in Model 3, the variable “h_prpay” becomes significant with a p-value less than or equal to 0.05 percent, and this is also the case in Model 4. Therefore, Model 4 is the preferred model, with the highest R-squared and adjusted R-squared (ar2) while including each variable of statistical significance across each model with one exception.

As a result, students with a higher high school percentile rank (HSPercentile), college grade point average (gpa), or student ACT composite scores (act) have an improved probability of persisting. Furthermore, students who prepay for student housing (h_prpay) or have a specific major (specific) are also more likely to persist to the following year. In contrast, female students (female), those on academic probation (probation), or those receiving a Federal Pell Grant (pell) are more likely to drop out. The ACT composite scores squared (act2) were significant with a coefficient value of zero. As a student’s ACT composite scores (act) go up, its benefit on student persistence is diminished. These findings are consistent with the existing literature, including Pell Grants' adverse effect on student retention (Chen, 2012).

Logit Regression Analysis

A comparison of LPM results with those of a traditional model like logit regression is recommended due to its weaknesses, such as heteroscedasticity, nonsense predictions, errors that are not normally distributed, and the constant effect independent variables have on the dependent variable. Moreover, logit regression provides results that are easier to interpret since it is asymptotically efficient, consistent, and usually provides distributed estimators.

The logit model is a statistical model that models the probability of one event with two alternatives taking place. Moreover, logit regression is the appropriate analysis since the dependent variable "drop" is dichotomous. The independent variables for each category, as indicated in Table 2 (Appendix), are placed sequentially into each logit regression model, as shown in Tables 4 and 7 (Appendix). The same technique is indicated in Table 3 (Appendix) for adding categories of variables to each LPM model. For example, student background variables are included in base-case Model 1 and remain in the other models as categories of variables are added, as indicated in Table 4 (Appendix).

The pseudo-R-squared (r2_p) improves as categories of variables are added to each model, with Model 4 equal to 0.314. Each variable that is statistically significant, like "female," "gpa," and "probation," remains substantial at the same level as additional categories of variables are added, except for the variable "black" in Model 1 and "HSPercentile" in Model 3. For example, as social integration variables are added in Model 3, the variable "h_prpay" becomes significant with a p-value less than or equal to 0.10 percent, and this is also the case in Model 4. Therefore, Model 4 is the preferred model, with the highest pseudo-R-squared (r2_p) containing the most variables of statistical significance while maintaining each variable of importance across each model with two exceptions.

A comparison between the coefficients of LPM and logit models is completed before analyzing logit models that present odds. Since the logit regression model is non-linear in terms of odds and probabilities, its coefficients, as indicated in Table 4 (Appendix), must be normalized, as shown in Table 5 (Appendix), using the "mfx" command in Stata (Stata, 2022).

The marginal effects of the logit model, as indicated in Table 5 (Appendix), are displayed in the "dy/dx" field, where they are normalized so that a change in probability based on one unit of change in an explanatory variable like "female" can be compared to the corresponding LPM coefficient as indicated in Table 3 (Appendix). The sign of the coefficients is consistent when comparing the models, as shown in Table 6 (Appendix), reducing the likelihood of nonsense predictions. Upon closer inspection, both models share the same statistical significance with one exception; the coefficient "HSPercentile" takes on $P>|z| = 0.136$ for logit. In the logit model, economic variables are also emphasized over other categories. For example, the coefficient "specific" increases from 3.1 to 4.3 percent while "female" decreases from 5.3 to 4.6 percent, as indicated in Table 6 (Appendix). Overall, the analysis shows that both models display a similar marginal effect.

For the remaining interpretation, the researchers will use the logit models that present the odds ratios as indicated in Table 7 (Appendix) instead of the raw coefficients as shown in Table 4 (Appendix). The likelihood ratio Chi-square value of 811.632 is large, and its p-value is less than 0.000, meaning the model's fit is statistically significant. Furthermore, since the pseudo-R-squared is 0.314, which is between 0.2 and 0.4, this also represents a strong fit. In Table 7 (Appendix), the logit regression reports odds ratios that provide the odds' multiplicative effects rather than the additive effects on the log odds or logits.

Table 7 (Appendix) shows that the odds of being a female (female) and dropping out are 1.407 times higher than a male student. In contrast, each unit of increase in a student's ACT composite scores (act) is 0.955 times less likely to drop out. The interpretation of a variable using the odds ratio differs according to whether it is dichotomous, such as "female," or a continuous variable, such as "act". For example, students who receive a Federal Pell Grant (pell) are 1.270 times more likely to drop out, which can be explained by the fact that these grants are issued to those from disadvantaged backgrounds. Similarly, a one-unit change in college grade point average (gpa) is associated with a 1-0.279, or 72.1% reduction in the likelihood of dropping out from one year to the next, while a student on academic probation (probation) is 1.559 times more likely to drop out.

Academic Learning Community Membership

Since membership in an academic learning community (alc) did not show up as significant in the regression results, further examination is warranted. Variance inflation factor (VIF) is used to assess how much the variance of the estimated regression coefficient increases if the predictors are correlated. Table 3 (Appendix) shows the linear probability models (LPM), and running the command "estat vif" provides the uncentered variance inflation factors as indicated in Table 8 (Appendix). Unfortunately, student ACT composite scores (act) and ACT composite scores squared (act2) have a VIF value greater than 10. Hence, multicollinearity is present, which can distort the findings by making it challenging to examine the effects of variables that are highly correlated with other independent variables.

As a result, two-sample t-tests with unequal variances for college grade point averages (gpa) and dropout rates (drop) were conducted. A member of an academic learning community (alc) is positively associated with the college grade point average (gpa), as indicated in Table 9 (Appendix), and negatively associated with dropping out, as shown in Table 10 (Appendix). This finding was statistically significant and consistent with the extant literature. Moreover, statistically significant differences for students who were members of an assist program (assist) were associated with higher grade point averages and lower dropout rates than those who were not members (Hoffman, 2020).

CONTRIBUTIONS AND LIMITATIONS

This study is essential because it performed retention analysis on a unique sample, included variables standard in the retention literature in addition to institution-specific variables, and uncovered methodological challenges with traditional multivariate models for explaining and predicting college student success. Variables were represented from all the major categories of determinants for college student persistence, but there are limitations, including a relatively small sample with limited data. Furthermore, there are variables in the literature that were not included in this study, and a cross-sectional analysis was performed but could be expanded to include controls for time. For example, financial aid variables might be more critical for students closer to graduation as the repayment of student loans is near. In contrast, first-time freshmen might not be as focused on their current financial situation because graduation is far off. Another limitation of the study is using a binary dependent variable (drop) instead of a multinomial variable so multiple decision outcomes can be examined, like students dropping out and re-enrolling.

RECOMMENDATIONS

Institution leadership should consider revising their data collection procedures to include more data, including time series data and additional variables on the front end through data entry. In addition, accounting for multiple decision outcomes is important because many students drop out and re-enroll, referred to as "stop out" in the literature. If that is not accounted for, the effects of the determinants could be biased, making explanation difficult. Multicollinearity also makes explanation challenging, so it is recommended that decision-makers evaluate institution-specific programs using a better design on the front end combined with straightforward quantitative methods, including analysis of variance (ANOVA), two-sample t-tests, and chi-square results from cross-tabulations. Another option would be to provide the institution-specific program to some students, but not others, to better understand how impactful those programs are in maximizing student success. For example, give the program to those students that qualify in cohort 1 but not those that qualify in cohort 2. Finally, it is recommended that leaders continue to conduct retention analyses using traditional models for explanation but explore other quantitative methods, including machine learning techniques like discriminant analysis, to maximize the efficacy of prediction.

REFERENCES

Aljohani, O. (2016). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6*(2), 1–18. https://doi.org/10.5539/hes.v6n2p1

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Developerment*, *40*(5), 518–529. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/5007c2/TN_cdi_proquest_journals_195180247

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. Worth Publishers.

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. https://link.springer.com/article/10.1007/BF00976194

Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *Review of Higher Education*, *6*, 127–148.

Bean, J. P. (2005). Nine Themes of College Student Retention. In A. Seidman (Ed.), *College Student Retention* (pp. 215–243). American Council on Education and Praeger Publishers. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/1lggg93/01INCARNATE_ALMA2126042830002031

Bean, J. P., & Eaton, S. B. (2000). A Psychological Model of College Student Retention. In J. M. Braxton (Ed.), *Reworking the Student Departure Puzzle* (pp. 48–61). Vanderbilt University Press.

Becker, G. S. (1992). *The economic way of looking at life*. https://www.nobelprize.org/uploads/2018/06/becker-lecture.pdf

Braxton, J., & Hirschy, A. S. (2005). Theoretical Developments in the Study of College Student Departure. In A. Seidman (Ed.), *College Student Retention* (pp. 61–87). American Council on Education and Praiger Publishers.

Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). The Convergence Between Two Theories of College Persistence. *The Journal of Higher Education*, *63*(2), 143–164. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/5007c2/TN_cdi_proquest_journals_205332444

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention. *Source: The Journal of Higher Education*, *64*(2), 123–139. https://www-jstor-org.uiwtx.idm.oclc.org/stable/pdf/2960026.pdf?refreqid=excelsior%3Ab64485d8800ef38f8967fab08da28a83&ab_segments=&origin=

Cai, L. (2020). Standardized Testing in College Admissions: Observations and Reflections. *Educational Measurement, Issues and Practice*, *39*(3), 34–36. https://doi.org/10.1111/emip.12389

Chen, R. (2012). Institutional Characteristics and College Student Dropout Risks: A Multilevel Event History Analysis. *Research in Higher Education*, *53*(5), 487–505.

Durkheim, E. (1951). *Suicide* (G. Simpson, Ed.). The Free Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley.

Hawley, T. H., & Harris, T. A. (2005). Student characteristics related to persistence for first-year community college students. *J. College student retention*, *7*(1–2), 117–142.

Hoffman, A. (2020). Success coaching program revisions and the impact on suspension and retention. The University of the Incarnate Word, 1-23.
Kerkvliet, J., & Nowell, C. (2005). Does One Size Fit All? University Differences in the Influence of Wage, Financial Aid, and Integration on Student Retention. *Economics of Education Review*, *24*(1), 89–95.
Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What Matters to Student Success: A Review of the Literature Commissioned Report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success*. https://nces.ed.gov/npec/pdf/kuh_team_report.pdf
Langbein, L. I., & Snider, K. (1999). The Impact of Teaching on Retention: Some Quantitative Evidence. *Social Science Quarterly*, *80*(3), 457–472. https://primo-tc-na01.hosted.exlibrisgroup.com/permalink/f/5007c2/TN_cdi_proquest_journals_204355482
Leppel, K. (2001). The Impact of Major on College Persistence among Freshmen. *Higher Education*, *41*(3), 327–342.
Leppel, K. (2005). College Persistence and Student Attitudes toward Financial Success. *College Student Journal*, *39*(2), 223–241.
Light, A. (1995). Hazard Model Estimates of the Decision to Reenroll in School. *Labor Economics*, *2*(4), 381–406.
McCormick, K. (1997). An essay on the origin of the rational utility maximization hypothesis and a suggested modification. *Eastern Economic Journal*, *23*(1), 17–30.
MDHE. (2022). *Admissions Selectivity Categories*. DHEWD Policies & Guidelines. https://dhewd.mo.gov/policies/admissions-selectivity.php
Montmarquette, C., Mahseredjian, S., & Houle, R. (2001). The Determinants of University Dropouts: A Bivariate Probability Model with Sample Selection. *Economics of Education Review*, *20*(5), 475–484.
Mueller, C. W., & Price, J. L. (1981). A causal model of turnover for nurses. *Academy of Management Review*, *24*(3), 543–565. https://www.jstor.org/stable/255574
NCES. (2020a). *COE - Characteristics of Degree-Granting Postsecondary Institutions*. Annual Reports and Information Staff (Annual Reports). https://nces.ed.gov/programs/coe/indicator/csa
NCES. (2020b). *Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected years, 2006 to 2019*. Digest of Education Statistics. https://nces.ed.gov/programs/digest/d20/tables/dt20_326.30.asp
Ogunwole, S. U., Rabe, M. A., Roberts, A. W., & Caplan, Z. (2021). *U.S. Adult Population Grew Faster Than Nation's Total Population From 2010 to 2020*. United States Census Bereau. https://www.census.gov/library/stories/2021/08/united-states-adult-population-grew-faster-than-nations-total-population-from-2010-to-2020.html
Reason, R. D. (2003). Student Variables that Predict Retention: Recent Research and New Developments. *NASPA Journal*, *40*(4).
Sanoff, A. P., Usher, A., Savino, M., & Clarke, M. (2007). College and University Ranking Systems. *Institute for Higher Education Policy*, 1–53. https://files.eric.ed.gov/fulltext/ED497028.pdf
Siedman, A. (2005). *College Student Retention : Formula for Student Success*. American Council on Education and Praiger Publishers.

Singell, L. D. (2004). Come and stay a while: does financial aid effect retention conditioned on enrollment at a large public university? *Economics of Education Review*, *23*(5), 459–471.

Spady, W. G. (1970). Dropouts from Higher Education: An Interdisciplinary Review and Synthesis. *Interchange*, 64–85.

Stata. (2022). *Stata/MP*. https://www.stata.com/statamp/

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. The University of Chicago Press.

Tinto, V. (2006). Research and Practice of Student Retention: What Next? *Journal of College Student Retention : Research, Theory & Practice*, *8*(1), 1–19. https://doi.org/10.2190/4YNU-4TMB-22DJ-AN4W

Tinto, V. (2017). Reflections on Student Persistence. *Student Success*, *8*(2), 1–8. https://doi.org/10.5204/ssj.v8i2.376

Wetzel, J. N., O'toole, D., & Peterson, S. (1999). Factors Affecting Student Retention Probabilities: A Case Study. *Journal of Economics and Finance*, *23*(1), 45–55.

APPENDIX

Table 1: 2017-2019 Tabulation of Student Participants

Group	Freq.	Percent
cohort_17	1,142	47.30%
cohort_18	1,272	52.70%
Total	2,414	100.00%

Table 2: Midwestern Regional Institution Database Records

Variable and Category	Measurement	Mean	SD
Dependent Variable:			
drop	1 if a student dropped, 0 otherwise	0.227	0.419
Student Background Variables:			
female	1 if a student is Female, 0 otherwise	0.576	0.494
black	1 if a student is Black, 0 otherwise	0.070	0.255
asian	1 if a student is Asian, 0 otherwise	0.010	0.097
Academic Integration Variables:			
act	Student ACT composite scores	22.150	4.330
act2	ACT composite scores squared	509.345	179.716
gpa	College grade point average	3.389	0.456
HSGPA	High school grade point average	65.546	21.815
HSPercentile	High school percentile rank	2.744	1.128
probation	1 if a student was on probation, 0 otherwise	0.135	0.341
Social Integration Variables:			
alc	1 if a member of a learning community, 0 otherwise	0.071	0.257
assist	1 if a member of the "assist" program, 0 otherwise	0.091	0.287
h_prpay	1 if prepaid student housing, 0 otherwise	0.908	0.288
Economic and Finance Variables:			
specific	1 if student major is specific, 0 otherwise	0.530	0.499
pell	1 if a student received a Federal Pell Grant, 0 otherwise	0.401	0.490
Additional Control Variable:			
Cohort	1 if a member of the 17-18 academic year, 0 otherwise	0.473	0.499

Table 3: Linear Probability Models

	Model 1	Model 2	Model 3	Model 4
female	-0.011	0.051***	0.052***	0.053***
	(0.017)	(0.015)	(0.015)	(0.015)
black	0.129***	0.009	0.006	-0.005
	(0.033)	(0.029)	(0.029)	(0.029)
asian	-0.002	-0.047	-0.048	-0.052
	(0.088)	(0.071)	(0.071)	(0.071)
act		-0.009*	-0.009*	-0.009*
		(0.006)	(0.006)	(0.006)
act2		0.000*	0.000*	0.000*
		(0.000)	(0.000)	(0.000)
HSGPA		-0.018	-0.016	-0.011
		(0.029)	(0.029)	(0.029)
HSPercentile		0.001**	0.001**	0.001**
		(0.001)	(0.001)	(0.001)
gpa		-0.215***	-0.214***	-0.214***
		(0.008)	(0.008)	(0.008)
probation		0.097***	0.096***	0.096***
		(0.023)	(0.023)	(0.023)
alc			-0.014	-0.009
			(0.027)	(0.028)
assist			0.039	0.035
			(0.028)	(0.028)
h_prpay			-0.050**	-0.048**
			(0.024)	(0.024)
specific				-0.031**
				(0.014)
pell				0.025*
				(0.015)
Cohort				-0.001
				(0.015)
_cons	0.225***	0.823***	0.830***	0.822***
	(0.013)	(0.098)	(0.104)	(0.104)
N	2414	2414	2414	2414
r2	0.007	0.349	0.350	0.352
ar2	0.005	0.346	0.347	0.348

Standard errors in parentheses
=* p<0.10 ** p<0.05 *** p<0.01

Table 4: Logit Regression Models

	Model 1	Model 2	Model 3	Model 4
drop				
female	-0.063	0.338***	0.345***	0.342***
	(0.099)	(0.129)	(0.130)	(0.130)
black	0.644***	0.019	0.003	-0.095
	(0.170)	(0.224)	(0.226)	(0.229)
asian	-0.012	-0.275	-0.270	-0.306
	(0.508)	(0.645)	(0.642)	(0.635)
act		-0.048*	-0.046*	-0.046*
		(0.055)	(0.056)	(0.057)
act2		0.001*	0.001*	0.001*
		(0.001)	(0.001)	(0.001)
HSGPA		-0.278	-0.281	-0.230
		(0.241)	(0.242)	(0.242)
HSPercentile		0.007	0.009*	0.007
		(0.005)	(0.005)	(0.005)
gpa		-1.270***	-1.272***	-1.276***
		(0.068)	(0.068)	(0.068)
probation		0.443***	0.439***	0.444***
		(0.165)	(0.166)	(0.166)
alc			-0.104	-0.040
			(0.249)	(0.261)
assist			0.156	0.120
			(0.212)	(0.213)
h_prpay			-0.381*	-0.363*
			(0.200)	(0.200)
specific				-0.308**
				(0.124)
pell				0.239*
				(0.127)
Cohort				-0.039
				(0.129)
_cons	-1.239***	2.550***	2.753***	2.708***
	(0.077)	(0.838)	(0.898)	(0.913)
N	2414	2414	2414	2414
r2_p	0.006	0.308	0.310	0.314
chi2	14.350	797.021	801.333	811.632

Standard errors in parentheses
=* p<0.10 ** p<0.05 *** p<0.01

Table 5: Marginal Effects after Logit Model 4

y = Pr(drop) (predict)
0.1652269

variable	dy/dx	Std. err.	z	P>z	[95% C.I.	]	X
female*	0.046	0.017	2.650	0.008	0.012	0.081	0.576
black*	-0.013	0.030	-0.430	0.668	-0.071	0.046	0.070
asian*	-0.038	0.071	-0.540	0.591	-0.177	0.101	0.010
act	-0.006	0.008	-0.810	0.420	-0.022	0.009	22.150
act2	0.000	0.000	0.880	0.379	0.000	0.001	509.345
HSGPA	-0.032	0.033	-0.950	0.341	-0.097	0.034	3.389
HSPercentile	0.001	0.001	1.490	0.137	0.000	0.002	65.546
gpa	-0.176	0.011	16.480	0.000	-0.197	-0.155	2.744
probation*	0.068	0.028	2.420	0.016	0.013	0.123	0.135
alc*	-0.005	0.035	-0.150	0.877	-0.074	0.064	0.071
assist*	0.017	0.031	0.550	0.584	-0.044	0.079	0.091
h_prpay*	-0.055	0.033	-1.660	0.097	-0.120	0.010	0.908
specific*	-0.043	0.017	-2.470	0.013	-0.077	-0.009	0.530
pell*	0.033	0.018	1.860	0.064	-0.002	0.069	0.401
Cohort*	-0.005	0.018	-0.310	0.760	-0.040	0.029	0.473

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Table 6: Compare LPM to Logit

	Model 4	
variable	LPM	Logit
female	0.053***	0.046***
black	-0.005	-0.013
asian	-0.052	-0.038
act	-0.009*	-0.006*
act2	0.000*	0.000*
HSGPA	-0.011	-0.032
HSPercentile	0.001**	0.001
gpa	-0.214***	-0.176***
probation	0.096***	0.068***
alc	-0.009	-0.005
assist	0.035	0.017
h_prpay	-0.048**	-0.055*
specific	-0.031**	-0.043**
pell	0.025*	0.033*
Cohort	-0.001	-0.005

Standard errors in parentheses

=* p<0.10 *** p<0.01" ** p<0.05

Table 7: Logit Regression Models Odds Ratios

	Model 1	Model 2	Model 3	Model 4
drop				
female	0.939	1.402***	1.412***	1.407***
	(0.093)	(0.181)	(0.183)	(0.184)
black	1.904***	1.019	1.003	0.909
	(0.324)	(0.229)	(0.227)	(0.208)
asian	0.988	0.759	0.764	0.737
	(0.502)	(0.490)	(0.490)	(0.468)
act		0.954	0.955	0.955
		(0.052)	(0.053)	(0.054)
act2		1.001	1.001	1.001
		(0.001)	(0.001)	(0.001)
HSGPA		0.757	0.755	0.794
		(0.182)	(0.183)	(0.192)
HSPercentile		1.007	1.009*	1.008
		(0.005)	(0.005)	(0.005)
gpa		0.281***	0.280***	0.279***
		(0.019)	(0.019)	(0.019)
probation		1.558***	1.551***	1.559***
		(0.257)	(0.257)	(0.259)
alc			0.901	0.961
			(0.225)	(0.251)
assist			1.169	1.128
			(0.247)	(0.240)
h_prpay			0.683*	0.696*
			(0.137)	(0.139)
specific				0.735**
				(0.091)
pell				1.270*
				(0.161)
Cohort				0.961
				(0.124)
_cons	0.290***	12.809***	15.691***	15.003***
	(0.022)	(10.739)	(14.092)	(13.696)
N	2414	2414	2414	2414
r2_p	0.006	0.308	0.310	0.314
chi2	14.350	797.021	801.333	811.632

Exponentiated coefficients; Standard errors in parentheses
=* p<0.10 ** p<0.05 *** p<0.01

Table 8: Variance Inflation Factors for LPM

Variable	VIF	1/VIF
act2	15.170	0.066
act	14.290	0.070
HSGPA	3.740	0.267
HSPercentile	3.440	0.290
gpa	1.540	0.651
assist	1.350	0.741
probation	1.300	0.767
black	1.190	0.839
female	1.120	0.894
alc	1.110	0.898
Cohort	1.110	0.900
pell	1.080	0.927
specific	1.030	0.973
h_prpay	1.030	0.974
asian	1.000	0.995
Mean VIF	3.3	

Table 9: Two-sample T-test with Unequal Variances for GPA

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	2,243	2.726	0.024	1.140	2.679	2.773
1	171	2.982	0.070	0.919	2.844	3.121
Combined	2,414	2.744	0.023	1.128	2.699	2.789
diff		-0.256	0.074		-0.403	-0.110

diff = mean(0) - mean(1) t = -3.4469
H0: diff = 0 Satterthwaite's degrees of freedom = 212.017

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.0003 Pr(T > t) = 0.0007 Pr(T > t) = 0.9997

Table 10: Two-sample T-test with Unequal Variances for Drop

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	2,243	0.232	0.009	0.422	0.215	0.250
1	171	0.164	0.028	0.371	0.108	0.220
Combined	2,414	0.227	0.009	0.419	0.211	0.244
diff		0.069	0.030		0.010	0.127

diff = mean(0) - mean(1) t = 2.3038
H0: diff = 0 Satterthwaite's degrees of freedom = 205.08

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9889 Pr(T > t) = 0.0222 Pr(T > t) = 0.0111