Signaling Student Retention With Prematriculation Data

Joseph G. Glynn
Paul L. Sauer
Thomas E. Miller

Logistic regression is employed to develop a model that seeks to provide information to enhance early identification of freshmen at risk of attrition. The early identification is accomplished shortly after freshman orientation. The dependent variable of interest is the binary and nominal variable of persistence. Students who proceed from freshman matriculation to graduation without ever having dropped out are classified as persistors, and freshman matriculates who leave college either temporarily or permanently are classified as dropouts. The independent variables employed to predict attrition include demographics; high school experiences; and attitudes, opinions, and values as reported on a survey administered during freshman orientation. The model and its results will be presented along with a brief description of the institutional intervention program designed to enhance student persistence.

While extensive research efforts have been used to develop and improve theoretical models of student retention or persistence (Bean, 41

Joseph G. Glynn is an associate professor of management at Canisius College; Paul L. Sauer is an associate professor at Canisius College; and Thomas E. Miller is the dean of students in the Division of Student Affairs at the University of South Florida.
1982; Cabrera, Nora, & Castaneda, 1993; Milem & Berger, 1997; Tinto, 1975, 1993), a concern of many administrators is the ability to predict as early as possible the likelihood of a student dropping out of school. This concern is particularly critical at the school where the data in this study was collected—a small college that depends upon tuition payments as the primary source of operating funds. For this and similar schools, retention of students is financially critical. The purpose of this research is to develop a logistic regression model of persistence that maximizes predictability of attrition of incoming freshmen by employing data obtained as early as possible in the application and matriculation process. The earlier the likelihood of dropping out of school is detected, the sooner intervention can occur to reduce that likelihood.

We begin with a review of the findings of research that includes the theoretical models of retention. These models provide insight into variables that have been shown to significantly relate, directly or indirectly, to persistence to maintain enrollment status.

### Theoretical Background

Three theoretical models account for the majority of published research on student retention: the Student Attrition Model (SAM) (Bean, 1982), the Student Integration Model (SIM) (Tinto, 1975), and Astin’s Theory of Involvement (1975).

The key theoretical premise in Tinto’s (1975) model is that as students are integrated into and become more interdependent with both the academic and social elements of a college or university, the probability that a student will leave the university declines. Attribution is a direct function of the congruency between the student and the institution. For most students this congruency is positively related to the time spent at the institution and is especially critical during the freshman year (Pascarella & Terenzini, 1980).

This congruency between the student and institution involves developing compatibility between a student’s motivation, drive, and academic ability and the academic and social characteristics of the college or university. In addition, a student’s commitment to an educational
goal plus a commitment to stay at the school are effective factors in the student’s decision to persist.

The most serious threat to the validity of Tinto’s model resulted from a potential specification error owing to failure to include a series of external factors (Bean, 1985). Working from a different theoretical base, Bean (1982) proposed an alternative comprehensive model that included external factors.

Bean’s (1982) model was based on both process models of organizational turnover and the Fishbein and Ajzen (1975) model of attitude—behavior relationships. Attitudes formed from beliefs result in behavioral intentions that lead to the overt behavioral act of persisting or dropping out.

Though SAM is driven by these theoretical relationships, a major contribution is the inclusion of external factors as antecedent constructs. External factors include attitude constructs that both directly (e.g., finances) and indirectly (e.g., influence of parents and friends mediated by institutional fits) affect intent. In addition, some constructs (e.g., finances and opportunities to transfer to other schools) have both a direct and indirect (mediated by intent) effect on behavior.

Astin (1975) found that involvement was critical to a student’s decision to persist or drop out. Astin (1984) concluded that a policy was effective to the extent that it increased student involvement.

Integration of the Theories and Models
Efforts to integrate both the SIM and SAM models (Cabrera, Nora, & Castaneda, 1993) and the SIM and Astin models (Milem & Berger, 1997) indicate that the three models are complementary. Most empirical research in this area employs longitudinal data collected at two or more time intervals after student matriculation. The premises of the SIM model require use of longitudinal data because the process of choice is based on the evolution of academic and social integration over time. Findings indicate that while in general the antecedent variables explain persistence, the tests of the complexities of the causal patterns and interactions of the variables in the model vary across studies. Part of this variation is attributed to the population from
Both SAM and SIM have been tested under a variety of conditions using a variety of subjects from a variety of institutions (for a review see Cabrera, Castaneda, Nora, & Hengstler, 1992). Comprehensive tests using structural equation modeling techniques have provided insight into comparative performance of the SIM and SAM (Cabrera et al., 1992; Cabrera, Nora, & Castaneda, 1993).

Cabrera et al. (1992, p. 159) found empirical evidence that: (a) persistence was mediated through intent for both models; (b) the models (i.e., theories) were complementary, not mutually exclusive; (c) external factors had a significant effect on persistence, supporting the SAM inclusion of them; and (d) two constructs were found to be equivalent across models: Courses (SAM) = Academic Integration (SIM), and Institutional Fit and Quality (SAM) = Institutional Commitment (SIM).

In SAM the construct “courses” are indicated by a single variable—satisfaction with the course curriculum. In SIM “academic integration” is indicated by two measures, one of which is satisfaction with the academic experience. In SAM, the construct “institutional fit and quality” is indicated by four measures, one of which is the feeling that the student belongs at that school. In SIM, “institutional commitment” is indicated by the single item, namely, the confidence the student had that the right decision was made to attend the chosen school. In Cabrera et al. (1992), only construct convergence was statistically evaluated across models. No attempt was made to statistically compare the nomological validity of the two models. “Nomological validity is the extent to which the scale correlated in theoretically predicted ways with measures of different but related constructs” (Maholtra, 2002, p. 294).

Cabrera, Nora, and Castaneda (1993) provide a comprehensive statistical test of the nomological validity of SAM and SIM. They achieve this by designing a single baseline model that integrates the structural relationships of the constructs in the two models. To achieve this integration they performed confirmatory analysis on the measurement properties of the two sets of constructs that were found to be identical.
in the Cabrera et al. (1992) study. Confirming this identity facilitated integration of the two models. Based on empirical evidence of a “significant other” effect on persistence (Nora, 1987; Nora, Attinasi, & Matonak, 1990), they further expanded the integrated model by incorporating a “significant other” construct effect to measure the impact on dropout decisions. Results of the statistical test of this expanded integrated model revealed that:

1. The external factor effects were stronger and more complex than portrayed and confirmed in empirical tests of the SAM model.
2. The “significant other” construct had a significant effect.
3. Intervention efforts cannot be simple and singly applied, but must combine a variety of support services to respond to both direct and indirect effects observed in the model (Cabrera, Nora, & Castaneda, 1993, pp. 135–136).

Milem and Berger (1997) formulated an integration of Tinto’s SIM model and Astin’s Theory of Involvement by modifying Tinto’s model to include measures of involvement. They noted that tests of SIM tend to use behavioral and perceptual measures, both of which differ from measures of involvement. Using research designed to collect longitudinal perceptual and involvement data and path analytic techniques to analyze relationships, they found strong support for their theorized relationship between involvement and perception, and support for their integrated model.

The benefit of these models is not only that the intervention efforts can be prescribed but also that the causes of persistence and attrition can be defined. These models do not necessarily provide the high levels of predictability that are desired by administrative staff in identifying, as early in the matriculation process as possible, the individuals who are most likely to drop out of school. The purpose of our study is to improve on this early identification of at-risk students by developing a model that optimizes predictability. To this end we build on the knowledge of the results of the empirical tests of these models to interpret the results of the logistic regression model we formulate to predict probability of attrition. Because logistic regression has not often been used to model persistence (Dey & Astin, 1993; Wetzel, O’Toole, & Peterson, 1999), a contribution of our study is an empirical test of
a model of persistence using a stepwise logistic regression procedure with an extensive set of variables comparable to the set of variables employed in the theoretical models previously discussed. The variables included in the student retention survey employed in our study were not based on any of the theoretical studies summarized earlier. The theoretical background discussion is provided to enable the reader to relate our study to past theoretical developments in the area of student retention.

Model Development

In this section, we address two issues: operationalization of variables and functional form of the model. With respect to the operationalization of variables, results of past studies lead us to conclude that:

1. Persistence is a function (direct or indirect) of background variables, goal commitment, institutional commitment, academic and social integration (Tinto, 1975), external factors (Bean, 1982), involvement (Astin, 1975) and financial factors (Wetzel, O'Toole, & Peterson, 1999). Explanatory variables should include measures of all of these.

2. Persistence is an evolutionary process that will be influenced to a greater or lesser extent by a variety of variables and constructs that increase or decrease in their relative importance over a student's tenure. At any point in time during a student’s educational process the effect of these influences needs to be evaluated. For example, a sophomore may be influenced to withdraw for different reasons than a junior.

The functional form of the model should represent the probability of persistence. The dichotomous dependent variable is persistence, and its value (persist or drop out) is the outcome of a choice process. Given the dichotomous dependent variable, the lack of longitudinal data, and the goal of maximizing predictability, the independent variables will all be modeled as having direct effects in a logistic regression.

The self-report measures used as independent variables are indicators of underlying latent characteristics of the students such as attitude and level of involvement.
Procedure

Because most dropouts at the school where the study was undertaken occur during the freshman year, our goal was to maximize the predictive validity of the model while employing data collected prior to or concurrent with matriculation. The potential predictor variables were limited to demographics concerning the student and the parents of the student, high school academic and social experiences, and attitudes and opinions over a wide range of topics collected via a survey administered during freshman orientation. Thus, all data were measures of characteristics, opinions, attitudes, and values formed prior to the college experience. Therefore our modeling approach differed from most previous empirical works that used data collected at multiple times during and after matriculation. From the Tinto model, only Background Variables, Goal Commitment 1, and Institutional Commitment 1 would be included in our model. Academic and Social Integration as well as Goal Commitment 2 and Institutional Commitment 2 designed to be measured during the freshman year would not be available at the time our data was collected. To provide a proxy for some of these measures obtained during the freshman year and thereafter, our model incorporated variables that measure similar experiences during high school. For example, Pascarella and Terenzini (1983) used measures of contact with university faculty during the freshman year to operationalize academic integration and social integration. Our model used similar measures, but with respect to high school teachers.

Unlike in the causal models of attrition behavior (Tinto, 1975; Bean, 1982; Cabrera et al., 1992; Cabrera, Nora, & Castaneda, 1993), the factors are modeled as having only direct effects by including them as independent variables in a logistic regression model. In fact, Wetzel, O’Toole, and Peterson (1999) included only academic integration, social integration, and financial support in their logistic regression model. They found that financial support offered little if any explanation of persistence.

Variable Development

Beginning in the late 1970s, the dean of students and senior student affairs officer at a medium-sized private college in the northeastern
part of the United States directed student-oriented research employing a combination of an established external instrument and a series of ad hoc studies that required the development of in-house survey instruments. Expertise gained from these studies contributed significantly to the development of the freshman student retention survey instrument in 1984.

External Survey Research

**CIRP Survey**
This national study of freshmen is part of the Cooperative Institutional Research Program (CIRP), and was administered annually by the institution to incoming students prior to the beginning of classes. The focus was upon freshman social, political, religious, and moral values; high school experiences; and expectations concerning academic and career goals.

In-House Survey Research

**CIRP Senior Follow-Up Survey**
A survey including portions of the CIRP instrument, and a few additional items added by the college, was administered to seniors as a part of their graduation experience. The goal was to track and assess changes in student perceptions between the freshman and the senior year.

**Dropout Survey**
Students who dropped out of the college were surveyed with respect to the reason(s) for their withdrawal. Results were compiled in a newly created dropouts database.

**Focus Groups**
The college contracted with an independent research organization to conduct focus groups wherein students were asked to evaluate their educational experience with respect to issues that address student services.

**Ad Hoc Studies**
Several ad hoc studies were undertaken to survey students over a wide array of topics potentially related to retention/attrition. Some of the
projects were: (a) residence hall preferences study, (b) faith and justice study (attempted to assess moral values), (c) residence hall retention study, and (d) student “work” research (analyzed characteristics of students working varying amounts of time at jobs on and off campus).

**Student Retention Instrument and Attrition Database**

In 1984 the senior student affairs officer and a faculty colleague developed a new survey instrument to be administered to all freshman matriculates. Survey items were derived in part from the CIRP Study and also from information gleaned from the ad hoc studies described above. Additionally, many institution-specific items were included in the instrument. For example, there are 20 items that require the respondent to rate “How important each item was in your decision to attend this college.” The survey was pilot-tested in the fall of 1984, revised, and has been used to poll freshman matriculates since 1985 (Miller, Glynn, & Neuner, 1988).

The survey data were combined with selected data from various campus administrative databases—financial aid, registrar, admissions, bursar, and the aforementioned new dropouts file. The resulting attrition database houses approximately 250 variables on all freshmen who completed the survey. The list of all 250 variables and the freshman survey are available from the authors.

The intent of the researchers was to identify a set of variables that would be of value in predicting probability of attrition. Decisions about the initial inclusion of specific survey-based items, demographic variables, and other data collected by the institution were based upon our judgment of their potential value as variables that might be associated with risk of attrition.

Freshmen were required to fill out the survey as a part of freshman orientation. This was the only path to becoming a member of the attrition database. Through 1995 (the last year of the analysis sample), the response rate had been about 95%, meaning that approximately 95% of freshman matriculates are represented in the attrition database.
Logistic Regression Model

Our operational definition of attrition is the act of dropping out at any time in a student's college experience. Thus a student who leaves the college during any semester, or completes a semester but fails to register for the following semester (without having graduated), is a dropout. Students who never drop out are referred to as persistors. Our fundamental goal was, at the time of matriculation, to assess the probability of attrition.

The logistic regression model was built on data collected from the freshman classes of 1988 to 1995 and was employed to assign a chance or probability of attrition to each freshman matriculating in the fall of 2000. Because the logistic regression model permits the estimation of each individual student's probability of dropping out of college (Stage, 1998; Wetzel, O'Toole, & Peterson, 1999), it is especially useful where an office of retention must prioritize students for intervention interviews with the goal of maximizing retention. A problem, however, is that the distribution of the dependent variable is likely to be highly skewed toward persistence. For example, if 85% of the analysis sample were persistors, a classification model that classified every student as a persistor would have a success rate of 85%, or would classify 85% of students correctly. To resolve this issue the maintenance of relative balance between the number of dropouts and the number of persistors (i.e., about 50% dropouts and 50% persistors) in the analysis sample was desirable. An appropriate number was randomly selected from the population of persistors. The result was an overall analysis sample of 5,221 students, with 50% persistors and 50% dropouts.

Principal Components Analysis of Survey Data

Because of the large number (250) of potential independent variables available for inclusion in our model, the potential for multicollinearity and subsequent parameter bias was high. Furthermore, many of the survey items had been identified in previous studies as multiple item indicators of underlying latent constructs (e.g., goal and institutional commitment, attitude, involvement). The student retention survey
contained several sections that sought to elicit respondent attitudes/opinions with respect to a host of variables such as physical fitness or study habits in high school; and expectations with respect to college and future career, family values, politics, religion, social responsibility, financial considerations, social relationships, and so forth, which were the same as or quite similar to the measures found in previous studies. These data, in combination with the responses to the importance of selected attributes in the decision to attend this college, formed the subset of variables that were initially analyzed via principal components analysis (PCA). The exploratory PCA model was used for two specific purposes: (a) data reduction and (b) the identification of factors to serve as uncorrelated potential predictor variables.

**Analysis Sample**

The attrition database used in the logistic regression included the selected 5,221 freshmen who matriculated between 1988 and 1995 inclusive. Our model does not focus upon freshman attrition exclusively. Rather, our intent is the early identification of students who are most likely to drop out at some point during their college experience. Students in the analysis sample fall into one of two groups: (a) persistors—students who have never dropped out and (b) dropouts—students who have dropped out at least once in their undergraduate career.

All students in the analysis sample had the opportunity to have dropped out any time during at least a four-year period, and thus were of an equivalent risk from an attrition perspective.

**Other Potential Predictor Variables**

In addition to the factors derived from the PCA model, a set of potential predictor variables for the logistic regression model consisting of background variables, external factors, financial factors, and influences of significant others were taken from other internal college databases (e.g., admissions, registrar) as well as from ad hoc studies previously cited.
Results

Principal Components Analysis of Survey Items

Seventy-nine items from the survey were included in the PCA model. All items were scaled one-to-five, with higher response numbers always signifying more agreement or more importance. Variables that did not appear to comport with the concept of simple structure were dropped from the analysis. A scree plot was used to determine that 12 factors would be retained for subsequent rotation. Coincidentally, the decision to retain 12 factors corresponded to a criterion of a minimum eigenvalue of 1.0 for inclusion in the factor solution. No cases were dropped as outliers, and variable means were substituted for missing values. The final model used 37 of the 79 survey items, and identified 12 principal components (factors) that accounted for 62.8% of the total variance among the 37 variables. A varimax rotation was used, and the resulting factor loading matrix met all of our expectations with respect to parsimony and a clear, interpretable factor pattern. The 12 rotated factors (SF1-SF12) derived from the PCA of the survey items, their eigenvalues, and the variables that loaded significantly on them, are listed in the Appendix. Factor scores (to be used as inputs to the logistic regression model) were computed by the regression method.

Other Predictor Variables

Selection of other independent variables for inclusion in the logistic regression model was accomplished via a series of tests comparing persistors versus dropouts on: (a) mean differences (t-tests) for interval/ratio and assumed-interval/ratio variables, or (b) the likelihood of a relationship (chi-square tests-of-independence) for nominal variables.

For example, interval/ratio and assumed-interval/ratio scaled variables such as high school average, math SAT score, age, and an expression of concern for financing their education (scaled 1–5) were evaluated via independent samples t-tests comparing averages for persistors and dropouts. Variables with observed significance levels of .05 or higher (p ≤ 0.05) were included as independent variables in the logistic regression. Chi-square tests-of-independence were run on the nomi-
nally scaled variables gender, religious preference, commuter/resident status, and marital status of parents versus the dichotomous dropout variable. Nominal variables displaying an observed significance level of .05 or better were entered into the original runs of the logistic regression model.

The final tally of independent variables to be included in the logistic regression model was 62 and was comprised of: (a) the 38 interval/ratio or assumed-interval/ratio variables (identified via t-tests), (b) the 12 categorical/nominal variables (identified via chi-square tests), and (c) the 12 factors: SF1-SF12 (identified by the PCA of the survey data).

Many of these 62 variables had high multicollinearity. For example, consider the combination of the three SAT variables—verbal SAT, math SAT, and total SAT scores. All 62 variables were originally used in exploratory backward stepwise logistic regression runs. We relied upon the backward stepwise procedure to eliminate variables that displayed high correlations with variables or combinations of variables in the model.

Logistic Regression

The selected 5,221 freshmen who matriculated between 1988 and 1995 formed the original analysis sample. The initial logistic regression run included all 62 potential predictor variables and all 5,221 cases. A two-part procedure was employed in an iterative manner to arrive at the final model of significant predictor variables: (a) the backward stepwise logistic regression procedure to remove variables on the basis of the Wald statistic being less than 0.10 for removal and (b) our own criterion of a studentized residual greater than 1.90 for removal of a case as an outlier. A residual is simply the difference between the observed probability of attrition (0.0 for persistors and 1.0 for dropouts) and the model's predicted probability of attrition (which must range between 0.0 and 1.0). The studentized residual can be computed for each student by dividing the residual for that student by the estimate of the standard deviation of the residuals that varies from point to point as the values of the predictor variables differ from their means. On successive runs variables were eliminated on the basis of
poor predictive power (i.e., Wald < 0.10 to leave the model), and cases were disqualified as outliers on the basis of a studentized residual greater than 1.90.

This alternating removal of variables, cases, then variables, and so on was continued until the “best” set of predictor variables had been identified. Once this set of predictor variables was determined, all cases that had been removed as outliers were restored to the analysis sample. A logistic regression was run on the full 5,221 case analysis sample employing the set of “best” predictors described above. On this and successive runs, outliers were once again identified and removed on the basis of studentized residuals greater than 1.90. Cases that had missing data for any predictor variable(s) were excluded from the analysis. The genesis of the analysis sample appears in Table 1.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Composition of Analysis Sample for Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Removed as outliers</td>
</tr>
<tr>
<td>Original cases</td>
<td>5,221</td>
</tr>
</tbody>
</table>

Our goal was the achievement of high predictive validity, and we adopted the following guidelines to determine the identification of the best set of significant variables in the final model:

1. Achieving at least 80% correctly classified in the analysis sample.
2. Limiting the excluded outliers to approximately 20% of the analysis sample.
3. Achieving a significant model chi-square (p < 0.01). This statistic tests the null hypothesis that all logistic regression coefficients (except the intercept) equal 0. This is analogous to the overall-F statistic in linear regression.
Ultimately, nine variables and seven factors (SFs) from the attrition database surfaced as potentially the most effective predictors of student attrition. These variables are listed in Table 2.

**Table 2**

**Independent Variables and Factors (SF) Used in the Logistic Regression Model**

<table>
<thead>
<tr>
<th>Background</th>
<th>Factor Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsave</td>
<td>Eight-term high school academic average</td>
</tr>
<tr>
<td>offwork1</td>
<td>Off-campus hours worked per week at time of matriculation</td>
</tr>
<tr>
<td>matrcage</td>
<td>Age at time of matriculation</td>
</tr>
<tr>
<td>edparnts</td>
<td>Total of mother’s and father’s education (range=1-10)</td>
</tr>
<tr>
<td>zipcode</td>
<td>Live in local MSA (1) or not (2)</td>
</tr>
<tr>
<td>parmar</td>
<td>Parents alive, married, and living together? Yes (1) or no (2)</td>
</tr>
<tr>
<td>gender</td>
<td>Male (1) or female (2)</td>
</tr>
<tr>
<td>Financial</td>
<td></td>
</tr>
<tr>
<td>concern1</td>
<td>Concern for financing education at time of matriculation</td>
</tr>
<tr>
<td>Goal commitment</td>
<td></td>
</tr>
<tr>
<td>SF5</td>
<td>Moral/religious values</td>
</tr>
<tr>
<td>SF6</td>
<td>Bad academic attitude in high school</td>
</tr>
<tr>
<td>SF7</td>
<td>Good study habits in high school</td>
</tr>
<tr>
<td>SF8</td>
<td>Good relationships with high school teachers</td>
</tr>
<tr>
<td>SF9</td>
<td>Concern for the disadvantaged</td>
</tr>
<tr>
<td>SF11</td>
<td>Expect academic problems in college</td>
</tr>
<tr>
<td>marfut</td>
<td>Expectations concerning marriage within one year of graduation</td>
</tr>
<tr>
<td>Institutional</td>
<td></td>
</tr>
<tr>
<td>SF1</td>
<td>Academic reasons for choosing this school</td>
</tr>
</tbody>
</table>
The results of the logistic regression to predict student attrition are represented in Table 3. Each predictor variable is listed, along with: (a) the R statistic—a measure of the partial correlation between the variable and the likelihood of dropping out, (b) the B value—the logistic regression coefficient, (c) the SE B—the standard error of B, and (d) the Wald Sig—the observed significance level for the B coefficient via the Wald statistic.

Variables in Table 3 have been listed in order of the magnitude of the R statistic, the strength of the partial correlation between the predictor variable listed and the dependent variable—likelihood of dropping out. Please note the signs of the R values and the signs of the regression coefficients B. The signs of R and B are negative for “hsave” (the eighth-term high school average). The interpretation is that as high school average increases, the likelihood of dropping out decreases. The magnitude of the R statistic for hsave is larger than that of any other predictor. This means that hsave is more closely associated with chances of dropping out than are any of the other predictor variables. The second variable listed in Table 3 is SF6, the factor variable that represented a bad academic attitude in high school. The positive sign of its coefficient B (and R value) indicates that higher scores on this variable are associated with higher chances of dropping out. Alternatively, bad academic attitudes in high school are likely to lead to attrition in college.

Examination of the R statistic (partial correlation coefficient) and the B value (logistic regression coefficient) reveals that variables positively related to the probability of dropping out included SF6, parmar, SF9, offwork1, concern1, matrcage, SF11, and zipcode. Variables that were positively related to greater likelihood of persistence included hsave, SF7, gender, SF5, edparnts, marfut, and SF8.

The criterion validity of the results of the logistic regression can be assessed by the percent of students correctly classified as dropouts or persistors as depicted in Table 4. In our model 83% were correctly classified. To assess criterion validity, a cut point of 0.5 was used to form a classification table. Each student whom the model indicated had a probability of 0.5 or greater of dropping out was classified as a predicted dropout. Each student who was forecast to have a probability of any value less than 0.5 was classified as a predicted persistor.
Table 3  
Logistic Regression Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>R</th>
<th>B</th>
<th>S.E. B</th>
<th>Wald Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-</td>
<td>22.367</td>
<td>2.049</td>
<td>.000</td>
</tr>
<tr>
<td>HSAVE</td>
<td>-.326</td>
<td>-0.381</td>
<td>0.017</td>
<td>.000</td>
</tr>
<tr>
<td>SF6</td>
<td>.252</td>
<td>1.405</td>
<td>0.083</td>
<td>.000</td>
</tr>
<tr>
<td>PARMAR(^a)</td>
<td>.212</td>
<td>2.132</td>
<td>0.149</td>
<td>.000</td>
</tr>
<tr>
<td>SF9</td>
<td>.181</td>
<td>0.818</td>
<td>0.067</td>
<td>.000</td>
</tr>
<tr>
<td>SF7</td>
<td>-.179</td>
<td>-.800</td>
<td>0.066</td>
<td>.000</td>
</tr>
<tr>
<td>GENDER(^a)</td>
<td>-.179</td>
<td>-1.578</td>
<td>0.131</td>
<td>.000</td>
</tr>
<tr>
<td>SF5</td>
<td>-.178</td>
<td>-.799</td>
<td>0.067</td>
<td>.000</td>
</tr>
<tr>
<td>EDPARNTS</td>
<td>-.165</td>
<td>-.375</td>
<td>0.034</td>
<td>.000</td>
</tr>
<tr>
<td>OFFWORK1</td>
<td>.161</td>
<td>0.712</td>
<td>0.065</td>
<td>.000</td>
</tr>
<tr>
<td>CONCERN1</td>
<td>.145</td>
<td>0.516</td>
<td>0.053</td>
<td>.000</td>
</tr>
<tr>
<td>MARFUT(^a)</td>
<td>-.115</td>
<td>-1.287</td>
<td>0.165</td>
<td>.000</td>
</tr>
<tr>
<td>SF8</td>
<td>-.108</td>
<td>-.454</td>
<td>0.061</td>
<td>.000</td>
</tr>
<tr>
<td>MATRCAGE</td>
<td>.090</td>
<td>0.529</td>
<td>0.085</td>
<td>.000</td>
</tr>
<tr>
<td>SF11</td>
<td>.086</td>
<td>0.368</td>
<td>0.062</td>
<td>.000</td>
</tr>
<tr>
<td>ZIPCODE(^a)</td>
<td>.072</td>
<td>0.743</td>
<td>0.147</td>
<td>.000</td>
</tr>
<tr>
<td>SF1</td>
<td>-.018</td>
<td>-0.113</td>
<td>0.060</td>
<td>.061</td>
</tr>
</tbody>
</table>

\(^a\)Categorical Variables
The classification table shows the number of students who were classified as dropouts who actually dropped out, and the number of students who were classified as persistors who never dropped out.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Persistors</th>
<th>Dropouts</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistors</td>
<td>1319</td>
<td>273</td>
<td>82.9%</td>
</tr>
<tr>
<td>Dropouts</td>
<td>279</td>
<td>1373</td>
<td>83.1%</td>
</tr>
<tr>
<td>Overall Percent</td>
<td>83.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

The purpose of this research was to develop a model of student retention that maximized the predictive validity while using data collected as early in the matriculation process as possible (i.e., at the time of or before matriculation into the college). The predictive validity was 83% as measured by the percent of students correctly classified as dropouts or persistors. This is a very high level of predictive validity when compared to previous efforts to model persistence. Our model serves to provide an early warning signal of potential likelihood of a student dropping out. Once a student is flagged as a dropout by our model, the wealth of literature on the process of assimilation into the college community along the lines of social and academic integration provided by numerous empirical tests of variations of the Tinto, Bean, and Astin models serves to guide the staff in directions that are most likely to be effective in intervention efforts (Newman & Newman, 1999; Tinto, 1997).

If we consider the two main components of the Tinto model, namely goal commitment and institutional commitment (Tinto, 1993), plus the external factors such as financial solvency and parental relation-
ships, we see that SF5, SF6, SF7, SF8, SF9, and SF11 are high school measures that serve as proxy indicators of college goal commitment and concern1 is a measure of financial stress that the student feels. Thus weaknesses in goal commitment tend to dominate with respect to increasing the probability of attrition. Institutional commitment is not a factor. This indicates that at the particular institution where the survey was conducted, students on the verge of attrition are not likely to be disenfranchised with the institution as much as they have not matured to a level of commitment to goals in their lives necessary to persevere with work at a college level. The results of the follow-up interviews of at-risk students should also provide information that can be used to confirm the basis for planned attrition, be it lack of goal or lack of institutional commitment.

The significance of SF6, the bad attitude toward academics in high school, is a condition of the student that is not likely to change after the student matriculates in college. It is indicative of students who have failed to integrate academically in high school and now face the same problem in college. The student embodies the problem and takes it with him or herself to college.

The effect of parmar also seems to corroborate a finding by Nora (1987) concerning the impact of significant others on the decision to drop out or persist. The likelihood of dropping out for a student from a family in which the parents were divorced (or at least one was deceased) was significantly greater than if the student were from a family in which the parents were still married and lived together. In this case, the most significant “other” would be one or two of the parents in a divorced family situation. It is also possible a divorce (death) may place a financial burden on one or both parents resulting in an inability to cover tuition and other costs of college.

**Intervention Program**

Recall that the primary goal of this research was the early identification of at-risk students. The logistic regression model provides the director of student retention with invaluable information concerning which students are at highest risk, and which variables may be influencing that high likelihood of attrition.
The institution has used the model to predict risk of attrition for individual students and then to intervene with those who are at risk greater than 50%. As soon as a satisfactory logistic regression model has been attained, each student in the new freshman class is assigned a predicted probability of attrition based upon responses and scores on the predictor variables. The director evaluates the projected at-risk students beginning at the top of the ordered list, or those students with assigned probabilities of or near 1.0. The director examines the scores on the predictor variables and creates a physical file on each student. In addition to the scores on the logistic regression predictors, many student attributes are noted. Some examples are athletic status, major, commuter/resident status, health considerations, extracurricular activities, job commitments, and actual financial need.

Students with the higher chances of attrition are contacted by telephone to make an appointment with the Office of Student Retention. The purposes of the face-to-face meetings with at-risk students are to: (a) identify problems, (b) provide support, and (c) reduce the potential of attrition. The objective of the director is to assist the student to identify and resolve any real or perceived problems, and to provide the student with a sense of involvement and ownership in the educational process. Schematically, the Office of Student Retention performs the activities depicted in Figure 1.

![Figure 1](image-url)

**Figure 1**

Proactive Approach to Providing Support for At-Risk Students
Application of the model has resulted in significant improvements in retention and graduation rates. In the ten years (1984–1993) leading up to the use of the model, the average rate of persistence from the freshman year to the sophomore year was 75%, and ranged between 74% and 76%, a matter of substantial consistency. However, as Table 5 shows, that rate has clearly risen since the initial application of the model to the freshman class in 1994, and resultant interventions.

The freshman retention rate increased substantially from 74.6% in 1993 to 80.9% in 1994. The freshman retention rate continued to increase to 84.6% in 1999 before falling (for reasons as yet undetermined) to 81.2% in 2000. The retention rates for years beyond the freshman year are not directly reported in Table 5. Rather, we chose to report four- to six-year graduation rates.

In conclusion, it can be fairly stated that the improvement in these freshman retention rates and four- to six-year graduation rates after the use of the model (1994 and beyond) are illustrative of the powerful impact of the model on the welfare of the institution and its students.

Table 5
Freshman Retention Rates and Four-, Five-, and Six-Year Graduation Rates (All Percents)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr. Ret. Rate</td>
<td>74.6</td>
<td>80.9</td>
<td>80.0</td>
<td>83.0</td>
<td>83.9</td>
<td>83.7</td>
<td>84.6</td>
<td>81.2</td>
</tr>
<tr>
<td>Grad. Rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-yr</td>
<td>37.3</td>
<td>45.9</td>
<td>47.6</td>
<td>53.0</td>
<td>52.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5-yr</td>
<td>53.9</td>
<td>58.0</td>
<td>57.5</td>
<td>64.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6-yr</td>
<td>56.0</td>
<td>60.6</td>
<td>60.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Limitations

The model is not able to provide any information with respect to the changes in likelihood of dropping out over time. That is, we were unable to track differences in the likelihood of attrition from the freshman to sophomore to junior to senior years. Hence the results reflect an average of effects over a four-year period during which the student may have elected to drop out.

The research provided pertains to a private university in an urban location with a significant, though declining number of commuter students. For freshman matriculates, the percent living in college residences has risen from about 37% in the late 1980s to 57% in 2000. Research by Bean and Metzner (1987) raises awareness of the likelihood that even the comprehensive baseline model may not be sufficient to handle the variety of institution and student types likely to be encountered. Bean and Metzner (1987) proposed a variation of the Bean (1982) model to explain dropout behavior by “nontraditional” students. Nontraditional students were defined as students over 24 years of age who do not live in campus residences, attend classes on a part-time basis, and hence are influenced significantly less by the social environment on campus. These students are primarily concerned with academic offerings. Dropout decisions are more significantly affected by external environmental influences (i.e., outside the campus/school environment).

References


---

**Appendix:**

**Factors (Independent Variables) Derived From Principal Components Analysis**

Included are a general description or factor name, the eigenvalue of the factor (in parentheses), and a verbatim listing of the survey items [with factor loadings from a varimax rotation in brackets] that comprised the factor. Some of the items on the college’s Attrition Survey instrument were borrowed directly from the CIRP Values instrument—these items are signified with ** below. Items denoted by a single * are very close to items on the CIRP instrument—only minor changes in wording were made on these items.

**SF1—Academic reasons for choosing this school (4.92)**

*The five items below have responses ranging from 1 (Not At All Important) to 5 (Extremely Important).*

- Quality of faculty [.75]
- Small class size [.67]
- Individual attention from faculty and administration [.76]
- Specific academic programs [.59]
- Teaching reputation [.76]
SF2—Unsure of career (2.79)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I know what career I plan to work in [-.82]

The two items below have responses ranging from 1 (Not At All Likely) to 5 (Extremely Likely).

*Change major field of study [.91]
*Change career plans [.92]

SF3—Enjoy politics (2.26)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I enjoy discussing political issues [.83]

The two items below refer to the respondent personally and have responses ranging from 1 Not At All Important to 5 Extremely Important.

**Influencing the political structure [.81]
**Keeping up to date with political affairs [.84]

SF4—Cost was an important factor in choice of this college (2.10)

The three items below concern choice of college and have responses ranging from 1 (Not At All Important) to 5 (Extremely Important).

Tuition, room, and expenses [.85]
Quality of service in financial aid program [.74]
Expected cost to you and your family [.87]

SF5—Moral/religious values (1.75)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

It is OK for couples to live together before marriage [-.61]

The one item below concerning activities while in high school has responses ranging from 1 Never) to 5 (Frequently).

Attended religious service [.70]

The two items below concerning choice of college have responses ranging from 1 (Not At All Important) to 5 (Extremely Important).

Jesuit academic tradition [.73]
Religious tradition [.82]
SF6—Bad academic attitude in high school (1.68)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I was very concerned with maintaining good grades in high school [.50]

The three items below concerning activities while in high school have responses ranging from 1 (Never) to 5 (Frequently).

Turned assignments late [.70]
Made careless mistakes on tests [.72]
Had difficulty concentrating on assignments [.71]

SF7—Good study habits in high school (1.60)

The three items below concerning activities while in high school have responses ranging from 1 (Never) to 5 (Frequently).

Saw teachers for help after class [.73]
Studied with friends [.68]
Did extra-credit assignments [.68]

SF8—Good relationships with high school teachers (1.46)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I had very good relationships with most of my high school teachers [.66]

The two items below concerning activities while in high school have responses ranging from 1 (Never) to 5 (Frequently).

Had poor teacher in class [-.71]
Had excellent teacher in class [.72]

SF9—Concern for the disadvantaged (1.30)

The one item below has responses ranging from 1 Strongly Disagree to 5 Strongly Agree.

**Students from disadvantaged social backgrounds should be given preferential treatment in college admissions [.72]

The two items below refer to the respondent personally and have responses ranging from 1 (Not At All) Important to 5 (Extremely Important).

**Helping to promote racial understanding [.73]
*Helping people who are in need [.61]
SF10—Physical fitness (1.19)

The one item below has responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I exercise or work out regularly to stay in shape [.82]

The one item below concerning choice of college has responses ranging from 1 (Not At All Important) to 5 (Extremely Important).

Athletics (facilities and programs) [.85]

SF11—Expect academic problems in college (1.13)

The two items below dealing with potential college experiences have responses ranging from 1 (Not At All Likely) to 5 (Extremely Likely).

**Fail one or more courses [.71]  
*Drop out temporarily [.81]

SF12—Computer skills (1.07)

The two items below have responses ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

I am able to write fairly detailed computer programs [.83]  
I feel comfortable using a personal computer (FOR OTHER THAN VIDEO GAMES) [.81]