PREDICTING GRADUATION RATES: A STUDY OF LAND GRANT, RESEARCH I AND AAU UNIVERSITIES

Rick Kroc, University of Arizona
Doug Woodard, University of Arizona
Rich Howard, Walsh College
Pat Hull, University of Arizona

Introduction

Graduation rates are becoming increasingly important as measures of effective undergraduate education and for institutional accountability. At local, state and national levels, these rates are subject to intense scrutiny. The implementation of the *Student Right-to-Know and Campus Security Act is* only the latest, if most difficult, manifestation of this trend. Because the most visible presentations of graduation rates have been superficial and to a large extent misleading *(US News and World Report,* October 4, 1993; *The Chronicle of Higher Education,* March 27, 1991; *NCAA Report on Graduation Rates,* 1993), colleges and universities need to have a deep and thorough understanding of student retention in general and of the graduation rates on their own campus. Fortunately, a substantial literature has developed to assist with this process. Models have been developed, tested and revised (Astin, 1971, 1993; Tinto, 1975, 1987; Bean, 1983, 1990; and Nora, Casteneda, & Cabrera, 1992).

This research draws on the Astin's (1993) work, which examined the predictability of graduation rates from the entry characteristics of students in light of his tripartite model: input, environment, and outcome (Astin, 1971). We focused exclusively on public land grant and research universities, institutions that have received the most withering criticism for their poor performance with undergraduates. The project proceeded in two phases. First, we examined the adequacy of the input side of the model: how well can graduation rates be predicted from student background characteristics, particularly for research and land grant universities? Second, we added institution level variables to the model and assessed their impact. In establishing the best analytic technique, we added our perspective to the debate about using logistic or linear regression (Dey and Astin, 1993). A third, qualitative, phase is planned to examine remaining differences between actual and predicted graduation rates.

The Database

Seventy-five land grant, Research I and AAU universities were asked if they would be willing to participate in a study of freshman graduation rates. Specifically, they were asked to create and send a unit record file of data about their 1988 freshman cohort, including high school grade point average, SAT and ACT test scores, class rank, sex, ethnicity, residency, four year graduation and persistence, and five year graduation and persistence. Data were kept confidential and reports on study findings were promised to participating universities. The first report was sent in September 1994.

Usable files were received from 53 institutions on more than 160,000 students. In addition to providing the foundation for our regression analysis, these data yielded an interesting array of descriptive data, as discussed below.

Data at the university level was collected from a number of sources, including NCES and various guide books. These variables as well as the student variables are described in Table 1.

Student Background and Graduation Rates

The simple descriptive data gathered from the 53 land grant and research universities during the first phase of the study provided some interesting baseline findings. For example:

- The four year graduation rate for women was 12% higher than the men's rate (36% compared to 24%). The five year rate favored women by 8% (59% vs. 51 %). These differences were unchanged after accounting for variations in high school grade point average, test scores, ethnicity and domicile using logistic regression.

- Underrepresented minority students (African American, Hispanic and Native American) were 9.5% of the study population. Their graduation rate was about 17% percentage points lower than the white student rate.

- In-state freshmen had a 2.4% lower five year graduation rate than out-of-state freshmen, but the five year graduation and persistence rate (including both graduated and enrolled students) favored residents by 2.1%.

- The mean four year graduation rate for the study population was 29.6%; the five year rate was 54.5%; an additional 9.7% of the students were still enrolled after five years; and the mean SAT/ACT score was 1023.

To examine the relationship between student background and graduation rates, we replicated and extended Astin's (1993) work, where student characteristics were regressed on graduation rates to produce a predicted graduation rate for a particular institution, which was then compared with the actual rate. Concordance tables were used to convert ACT scores to the SAT scale, and Chisholm's method (1992) provided a concordance to estimate high school grade point average from class rank for those universities that only had rank. Data were sufficient for more than 130,000 students at 44 of the universities to complete this phase of the analysis. We estimated graduation and persistence rates (four and five year) using high school GPA, test scores, sex, and ethnicity, comparing our equations and results with Astin's. In addition, we compared linear with logistic regression and added residency (domicile) to the analysis. Comparisons of our prediction equations with Astin's (1993) revealed interesting differences. Whereas Astin obtained the strongest correlation with four year graduation rates (R = .34), our best results were obtained using five year rates (r

= .32). Astin's equation over-predicted the four year rate for 93% of the universities in our sample; our equation produced a better fit, including 10% better concordance with the actual individual Student data (see Figure 1). For four year graduation and persistence, however, Astin's results and ours were essentially equivalent. We concluded from these findings that Astin's prediction equations, particularly for four year graduation rates, provide a relatively poor fit for land grant and research universities, most likely because much of the data were obtained from smaller, private institutions, where graduation in four years is much more frequent.

To illustrate the impact of using student background to predict graduation rates, the results for two universities are useful. University A had an actual five year graduation rate of 66.8%, more than 10 percentage points higher than the study mean. With student variables in the regression equation, however, the predicted rate was 67.9%, slightly above the actual rate for this university. University B had an actual rate of 31.3%, nearly 25% below the mean, but only about 12% below its predicted rate of 43.4%. These examples indicate how student background, as quantified and collected in most institutional admissions offices, is related to universities' graduation rates. The column labeled "predicted with student variables" in Table 2 displays the predicted graduation rates using only student variables for the 44 universities with sufficient data for the regression. Actual rates are also shown.

Table 3 shows the standardized regression coefficients for the student variables. High school grade point average was the best predictor of graduation rate, followed by test scores and gender; ethnicity and domicile had smaller, but appreciable effects. The regression correctly classified 68.4% of the students in the study as having graduated or not graduated after five years. Significance testing is not particularly relevant in this study since the large number of cases will result in almost any non-zero coefficient becoming statistically significant.

In all of our comparisons, logistic regression performed better than linear regression. Residual analysis showed a better fit, particularly at the extremes of the distribution (see Figure 2); concordance of predicted and actual graduation at the individual student level was 4% better; and 68% of the universities had a closer fit between their actual and predicted rates with logistic regression. Although logistic regression adds some complexity to the analysis and interpretation, we concluded that the benefits outweigh the drawbacks for this type of analysis.

Institutional Variables and Graduation Rates

As shown in Table 1, we identified 22 university level variables for inclusion in the study. Because of the number and disparate nature of these variables, we decided to apply factor analysis to simplify the data. The orthogonal rotation identified four primary factors: a cost factor loading heavily on tuition and cost of attendance variables; a size factor with high loadings on enrollment and library volume variables; a quality factor that loads heavily on student selectivity and faculty credential variables; and a budget factor reflecting ratios of budget categories. We created four factor scores as new variables and, in addition, we retained two other variables that correlated with graduation rate, SAT test score variance and percent of freshmen in residence halls.

We then added the six institution level variables to the student background variables in a multi-level logistic regression. Although we intend to use hierarchical linear modeling techniques to refine this analysis in the future, at this point we have repeated the institutional variables across all students at a university and applied logistic methods to the individual student records.

Adding institution variables improved the prediction somewhat. The overall equation correctly predicted the graduation (or non-graduation) of 70.2% of the students and the correlation between predicted and actual rates rose to .35. Fifty eight percent of the university graduation rates were more accurately predicted using both student and institutional variables than with only student variables. As shown in Table 3, the cost factor had the largest coefficient among the University level variables, second only to high school grades in the overall analysis - the higher the cost, the higher the graduation rate. Cost and residence hall percentage also had appreciable coefficients, followed by test score variance and budget ratio. Because of its colinearity with student variables, the institutional quality factor had no influence in the model.

Returning to our examples, University A had a predicted five year graduation rate of 69.9% using both student and institution level variables, slightly above its actual rate of 66.8%. University B had a predicted rate of 32.3%, very close to its actual rate of 31.3%. The column labeled "predicted with institutional and student variables" in Table 2 displays the predicted graduation rates using both student and institutional variables for the 44 universities with sufficient data for the regression. Although the overall correlation rose only modestly when institutional variables were added, the predicted graduation rate for many universities changed substantially.

Implications

With regard to our ability to predict graduation rates in this study, we found that the glass was both half full and half empty. On one hand, these rates were more predictable than we thought they would be - our seemingly low correlation of .35 resulted in better prediction at the university level than might be expected. On the other hand, the actual rate for most universities still differed, sometimes by large amounts, from the predicted rate. As described in the next section, we intend to explore further these differences using more qualitative methods.

We also concluded that the decision as to which institutions to include in an analysis of this kind is critical. Predicted graduation rates for land grant, Research I and AAU universities were quite different from those in Astin's (1993) study pointing out the necessity of understanding the sample used to develop the regression equation.

Adding institution level variables to the student variables improved our prediction, although the distinction between these two levels is somewhat blurred in several cases. For example, the tuition/cost of attendance variable may be simply a proxy for the socio-economic status of the student. In other words, if we knew student SES, the cost factor might not have significantly entered into the analysis. We hope to identify both additional student and institutional factors in the next phase of the study. Moreover, it is clear that in order to change the educational culture and create a learning environment that focuses on student learning (outcomes), researchers need to examine the influence of institutional policies and practices on student retention and subsequent graduation rates. Very few ethnographic studies have been done to identify the influence of institutional culture (Kuh's institutional ethos) in student retention and graduation.

With regard to analytic technique, we believe logistic regression is preferable to linear regression when analyzing graduation rates. Although both techniques perform equally well in the center of the distribution, logit is clearly superior at the extremes. This could be particularly important in analyses of students with particularly high or low graduation rates.

Finally, we learned that establishing a comprehensive national data base with individual student record data is both possible and valuable. This data collection and management process, which is the subject of another presentation during this AIR Forum, was interesting and important in its own right. We hope to expand this data base by adding more universities, other cohorts, and additional variables. We will continue to provide reports back to participating universities.

Further Study: Qualitative Analysis

To study the remaining differences between predicted and actual graduation rates for the 53 universities in our study, we will likely need to employ qualitative methods. Although our primary focus will be on identifying institutional factors that are within a university's control, we need also to ferret out student factors, such as SES, that may not be well accounted for in the current analysis, but may also further illuminate graduation rates.

We are considering two alternative strategies, both involving case studies of individual universities:

- Study outliers, those universities with the largest differences between predicted and actual graduation rates; or

- Study pairs of universities that have similar predicted graduation rates, but substantially different actual rates.

In either case, the strategy would be to look beyond simple quantitative variables. Clearly, quantifiable student and institutional variables help us understand graduation rates. Just as clearly, other factors, perhaps less easily measured and more idiosyncratic, are important.

References

Astin, A. (1971). *Predicting academic performance in college.* New York: Free Press.

Astin, A. (1993). How good is your institution's retention rate? Los Angeles: Higher

Education Research Institute.

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

Bean, J. P. (1990). Why students leave: Insights from research. In Hossler,D., Bean,J. P. and Assoc. (Eds.), *The strategic management of college enrollments.* San Francisco: Jossey-Bass.

Chisholm, M. P. (1993). An evaluation of a statewide admission standards policy. Association for Institutional Research Annual Forum, Atlanta.

Dey, E. L. and Astin, A. (1993). Statistical alternatives to studying college retention: A comparative analysis of logit, probit and linear regression. *Research in Higher Education, 34,* 569-582.

Durkin, K.F., Griff iths, K.L., and McLaughlin, G. W. (1992). Institutional Models for

Retention and Graduation Rates. Paper presented at the SAIR/SCUP Conference, Myrtle Beach, South Carolina.

Kuh, G.D. (1995). The Other Curriculum. Journal of Higher Education, 66, pp. 123-155.

Nora, A., Castenda, M.B., & Cabrera, A.F. (1992). Student persistence: The testing of a comprehensive structural model of retention. Paper presented at the 1992 ASHE Annual Meeting, Minneapolis, Minnesota.

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

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition..* Chicago: University of Chicago Press.

Table 1

STUDENT AND INSTITUTIONAL VARIABLES

STUDENT VARIABLES	INSTITUTIONAL VARIABLES
High School Grade Point Average	**Undergraduate acceptance rate**
SAT / ACT Test Score	**Percent in top 10% of high school class**
Sex	**Student / faculty ratio**
Class Rank	**Percent faculty with Ph.D.**
Residency	**Tuition - in and out of state**
Ethnicity T Test Score	**Ratio of budget to gifts, grants, contracts**
Sex	**Cost of attendance - in and out of state**
Class Rank	**Financial aid: percent of cost of attendance**
Residency	**Percent budget for instruction**
Ethnicity	**Percent budget on academic/student support**
	Geographic location
	SAT/ACT test score variance
	Enrollment: undergrad, grad and total
	Grad/undergrad ratio
	Percent freshmen receiving financial aid
	Library volumes
	Percent freshmen in residence halls

Table 2

Logistic Regression Analysis on 5-Year Graduation Rate
(data from 44 Institutions)

University ID	Number of Freshmen
24	2,612
45	4,659
27	3,266
9	1,810
5	3,272
42	2,864
17	3,416
47	5,387
35	5,882
39	4,002
43	2,774
15	1,605
40	3,190
2	3,643
3	3,561
4	3,566
54	3,319
52	3,123
37	3,597
31	7,697
13	2,693
25	3,946

16	2,534
28	1,149
34	3,055
21	3,795
8	2,313
6	1,901
18	3,029
7	3,634
14	2,071
1	5,015
44	3,352
53	1,541
32	2,979
49	2,938
38	1,378
26	3,532
41	4,076
50	2,508
10	1,310
20	2,845
46	1,694
51	1,824

5-Yr. Graduation Rate

Actual Rate (sorted in descending order)	Predicted with Student Variables	Predicted with Institutional & Student Variables	Sort Rank
90.8	72.5	80.2	1
83.1	69.1	83.2	2
78.6	65.5	77.1	3
74.6	62.0	79.3	4
70.5	56.2	61.0	5
66.8	67.5	69.8	6
66.3	62.8	68.9	7
66.2	59.7	64.5	8
65.4	61.7	65.4	9
65.2	58.4	60.7	10
65.0	60.0	63.8	11
63.9	54.7	61.7	12
61.5	57.7	55.4	13
60.8	61.6	67.9	14
58.4	59.0	56.3	15
58.4	51.4	52.9	16
58.2	64.0	66.3	17
57.0	63.5	64.2	18
56.2	56.6	60.9	19
55.9	60.6	55.3	20
55.3	56.0	53.5	21
54.8	54.0	56.0	22

54.5	54.5	51.9	23
52.5	49.2	40.2	24
49.2	54.7	59.1	25
49.1	50.6	44.6	26
48.5	52.5	55.2	27
48.2	50.8	52.7	28
47.8	51.7	39.0	29
47.3	47.3	46.9	30
45.6	49.2	35.4	31
45.3	53.0	48.9	32
42.8	48.2	44.4	33
42.4	51.3	41.0	34
41.8	59.0	51.9	35
40.6	53.8	53.0	36
39.4	53.5	42.8	37
39.3	44.4	37.7	38
39.1	52.4	51.7	39
35.4	50.6	38.2	40
34.9	50.5	39.8	41
33.8	50.6	45.5	42
31.3	43.0	32.2	43
28.2	46.8	38.1	44

Table 3

STANDARDIZED REGRESSION COEFFICIENTS

Student Variables	Coefficient: Student Variables Only in Equation	Coefficient: All Variables in Equation
High School GPA	.28	.26
SAT Math	.12	.08
Sex (male=0; female=1)	.11	.10
Residency (res=0; non-res=1)	.05	.02
Ethnicity (White=1; other=0)	.04	.05
Ethnicity (Native American=1)	-.03	-.03
Ethnicity (Hispanic=1)	-.03	-.03
SAT Verbal	.02	.02
Ethnicity (African American=1)	-.01	-.03
Institutional Variables		
Cost Factor		.13
Percent in Residence Halls		.06
Size Factor		.04
SAT / ACT Variance		.05
Budget Ratio Factor		.03
Quality Factor		.00

Institutions with sufficient data to include in regression:		
Arizona State U.	U. of Calif.-Davis	U. of Minnesota
Colorado State U.	U. of Calif.-San Diego	U. of Missouri
Florida State U.	U. of Colorado-Boulder	U. of Nebraska
Kansas State U.	U. of Connecticut	U. of New Mexico
Louisiana State U.	U. of Delaware	U. of Oklahoma
Mississippi State U.	U. of Florida	U. of Oregon
New Mexico State U.	U. of Hawaii	U. of Rhode Island
N. Carolina State U.	U. of Idaho	U. of Tennessee

Penn. State U.	U. of Indiana	U. of Texan-Austin
Rutgers. U.	U. of Iowa	U. of Vermont
S. Dakota State U.	U. of Kentucky	U. of Virginia
Texas A&M U.	U. of Maine	U. of Washington
U. of Alabama	U. of Maryland	U. of Wisconsin
U. Arizona	U. of Massachusetts	U. of Wyoming
U. of Arkansas	U. of Michigan	Washington State U.