DOCUMENT RESUME

ED 273 672 TM 860 523

AUTHOR	McGuire, Michael D.
TITLE	Identification of Students at Risk for Early Withdrawal: A Hierarchical Clustering Model.
PUB DATE	Apr 86
NOTE	11p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cluster Analysis; *Dropout Characteristics; Dropout Research; Factor Structure; Higher Education; *High Risk Students; *Persistence; *Potential Dropouts; Predictor Variables; Research Design; Two Year Colleges; Two Year College Students
IDENTIFIERS	Hierarchical Cluster Analysis

ABSTRACT

The hierarchical clustering technique was used to differentiate college dropouts from persisters, and to determine how student clusters differ from one another on relevant outcome variables. Subjects were 618 freshmen who entered a community college with the intention of completing a two-year associate degree. There were 432 persisters who returned for the last three semesters, and 186 non-persisters who did not. Four variables which were found to differentiate persisters and non-persisters were entered into a hierarchical clustering program (PROC CLUSTER in SAS): (1) age; (2) degree objective; (3) scores on writing placement tests; and (4) scores on mathematics placement tests. Five clusters were developed, according to academic readiness, performance, and persistence. Analysis of variance and chi square tests were performed. Dependent variables included age; sex; ethnic status; grade point average; placement test scores in reading, mathematics, and writing; and percentage of first semester courses completed. A comparison of the predictive value of the clustering technique and of placement tests indicated that clustering was not superior in diagnostic sensitivity or specificity. Although the distribution of persisters and non-persisters was significantly different across clusters, the hierarchical clustering patterns did not appear to have sufficient predictive validity as a screening method. (GDC)

IDENTIFICATION OF STUDENTS AT RISK FOR EARLY WITHDRAWAL: A HIERARCHICAL CLUSTERING MODEL

Michael D. McGuire

College of Lake County

INTRODUCTION

The past ten years have witnessed the development of a substantial base of theoretical and practical information on the subject of student retention in higher education. Relatively stable and comprehensive models of academic persistence and withdrawal behaviors have been postulated (e.g., Tinto, 1975) and substantiated (e.g., Chapman & Pascarella, 1983). On the practical issues of retention enhancement, a large number and wide variety of theory-driven strategies have evolved for helping colleges and universities to reduce student attrition (e.g., Noel, 1978; Lenning et al., 1980). Among the more promising actions is the building of an "early warning system" to identify those students who appear most likely to drop out prematurely.

As with high-risk targeting of individuals in other fields, issues of diagnostic sensitivity and specificity must be addressed before a formal identification and intervention system can be implemented in an ethical and cost-effective manner. For example, the common finding that ethnic minority populations suffer a higher attrition rate than white middle-class populations does not provide sufficiently sensitive (i.e., many non-persisters are not minority students) or specific (i.e., many minority students are persisters) for efficient intervention purposes. The possibility that indiscriminate "marking" of students as being at risk for attrition may actually impede the academic progress of persisters, or stigmatize them in the eyes of other students, suggests the need for refinement of the identification process. Use of global identification strategies followed by voluntary participation in retention-enhancing activities may not effectively reach most of the highest-risk students.

Another problem in the identification of students with a high probability of withdrawing prematurely is the lack of information available at time of initial enrollment. As a result, many students have already withdrawn by the time "marker" information (e.g., mid-term attendance or grade rosters) can be gathered. Even in those cases where attendance or grade data are available prior to midterm, the momentum that may have carried high-risk students into their first college enrollment may have shifted irreversibly under the weight of early confusion, failure, intimidation, and conflicting responsibilities.

This situation is especially acute at public community colleges, which are often characterized by "easy access" and "open door" policies that simplify the admission process and thus hinder the collection of sufficient data for sophisticated retention modeling. Unfortunately, these institutions are also characterized by an extremely high attrition rate, and are thus in greatest need of effective identification and intervention strategies (Lenning et al., 1980). Ideally, high-risk students should be identified prior to initial enrollment (hence, using only pre-enrollment information). Given the practical barriers to a lengthy application process and extensive mandatory testing and advisement to all beginning students, especially at large community colleges with a high percentage of evening and part-time enrollees, there is a need to develop screening methods that can provide as much predictive leverage as possible from a minimal amount of admission and placement information.

Finally, there is the problem of analytical complexity in making a binary decision (high risk vs. low risk) about a student on the basis of multiple risk criteria with independent thresholds. For example, how would one categorize a student who is above the risk threshold on three of six marker variables but well below the threshold on the other three? Given the reality that most high-risk students do not conform to a single profile, one approach would be to study the retention patterns and other characteristics of groups of students formed on the basis of similarities on these marker variables.

Hierarchical clustering techniques constitute a potentially useful method for grcuping together students with similar pre-enrollment characteristics and determining whether some clusters are comprised of students at high risk for premature withdrawal from college.[1] Hierarchical clustering is a common statistical procedure in taxonomic sciences, and involves the classification of objects (i.e., students) into subgroups that

[1] It is of interest that in a recent review of methodologies in 78 retention studies (Bean & Metzner, 1985), no study used hierarchical clustering techniques.

are as homogeneous as possible (Bachelor & Buchanan, 1984). The purpose of the present study was to test the hypothesis that hierarchical clustering can reliably differentiate persisters from non-persisters, and to determine how student clusters differ from one another on relevant outcome variables.

METHOD

The sample consisted of 618 students who first enrolled at a large suburban Midwestern community college in the Fall of 1983 with the stated goal of obtaining a two-year associate degree. All first-time students from this cohort with complete data were included in the study. There were 432 persisters and 186 non-persisters in the sample; persisters were those students who returned for each of the next three major semesters (1984 Spring, 1984 Fall, 1985 Spring), while non-persisters did not enroll in any of the next three terms.

A stepwise discriminant analysis revealed that 4 pre-enrollment variables significantly differentiated between persisters and non-persisters: age, degree objective, and scores on writing and mathematics placement exams. The raw scores for these variables were entered into a hierarchical clustering program (PROC CLUSTER in SAS) using Ward's clustering method (SAS Institute, 1982); the resulting cluster tree was then plotted (cubic clustering criterion, or CCC, by number of clusters) to determine a "best" configuration (defined jointly as a local CCC peak, a minimum R^2 of .80, and a local maximum R^2 increment from the next smallest configuration). An effort was also made to keep the number of clusters reasonably small, to avoid fragmenting the sample to such a degree that high-attrition or high-persistence clusters would occur on the basis of chance alone. The optimum configuration was judged to consist of 5 clusters (R^2=.80).

Analysis of variance and chi-square tests were performed to determine if the 5 clusters of students differed significantly on persistence status and on other characteristics, especially those not included in the cluster analysis. Dependent variables included degree objective, sex, ethnic status, age, grade point average, placement test scores in reading, writing and mathematics, and percentage of first semester courses completed.

Finally, identification accuracy was calculated for the three student clusters with the highest percentage of non-persisters, and for the mathematics and writing placement test scores at five cutoff scores. This last comparison was to determine whether the clustering technique resulted in greater precision in identifying high-risk students than simple placement tests alone.

RESULTS

Table 1 lists the characteristics of each group of students identified by the hierarchical clustering model. There were significant group differences on each of the placement exams, course completion rate, persistence status, G.P.A., ethnic category, and degree objective. There were no significant group differences on age or sex.

TABLE 1

Means and Frequencies, by Cluster

			CLUSTER				
Variable	1	2	3	4	5	F	p
MEANS							
Writing Exam	80.6	44.5	20.6	28.8	72.7	435.7	.0001
Reading Exam	83.6	62.7	44.4	39.8	71.1	105.5	.0001
Math Exam	92.1	82.2	57.6	20.7	52.4	779.2	.0001
% of Courses Completed	92.3	80.9	78.2	82.6	84.1	5.0	.001
Age	22.9	21.7	23.9	22.0	22.9	1.8	N.S.
G.P.A.	2.82	2.31	2.01	2.02	2.43	16.9	.0001

			CLUSTER				
Variable	1	2	3	4	5	(X^2)	p
FREQUENCIES							
Persisters	149	101	37	50	95	27.0	.0001
Non-persisters	32	45	32	35	42		
Males	84	87	38	39	65	7.8	N.S.
Females	97	59	31	46	72		
White	175	133	55	58	129	81.5	.0001
Non-White	6	13	14	27	8		
A.A.	20	11	7	8	22	21.0	.01
A.S.	46	29	5	10	22		
A.A.S.	115	106	57	67	93		

Significant Pairwise Comparisons (Tukey's HSD, alpha=0.05):

WRITING: Cluster 1 > Cluster 5 > Cluster 2 > Cluster 4 > Cluster 3.

READING: Cluster 1 > Cluster 5 > Cluster 2 > Clusters 3 & 4.

MATH: Cluster 1 > Cluster 2 > Cluster 3 > Cluster 5 > Cluster 4.

COURSE COMPLETION: Cluster 1 > Clusters 2 & 3.

G.P.A.: Cluster 1 > Clusters 2-5; Cluster 5 > Clusters 3 & 4.

Cluster 1 was characterized by a high level of academic readiness, performance, and persistence. This group had the highest proportion of persisters (82.3%) and baccalaureate-oriented students of all clusters. Cluster 2 was characterized by relatively high mathematics placement scores but relatively low reading and writing scores. Its members assumed the middle-ground on the other dependent measures, including a 69.2% persistence rate. Cluster 3 was characterized by a relatively low level of academic readiness, especially in writing, and by the lowest G.P.A., course completion rate, and persistence rate (53.6%) of all clusters. Cluster 4 resembled Cluster 3 in several respects, with somewhat lower math and reading placement scores but higher course completion and persistence rates (58.8%). Cluster 5 had the second-highest readiness (except in mathematics), performance, completion, and persistence (69.3%) indicators.

TABLE 2

Comparison of Identification Accuracy

	SENSITIVITY	SPECIFICITY
	% of total non-persisters identified	% of identified who were non-persisters
Cluster 2	24.2%	30.8%
Cluster 3	17.2%	46.4%
Cluster 4	18.8%	41.2%
Writing Test		
20th %ile:	16.1%	43.5%
25th %ile:	29.6%	48.7%
30th %ile:	35.5%	48.9%
35th %ile:	39.3%	45.1%
40th %ile:	43.6%	42.4%

TABLE 2 (Continued)

	SENSITIVITY % of total non-persisters identified	SPECIFICITY % of identified who were non-persisters
Math Test		
20th %ile:	10.8%	48.8%
25th %ile:	15.6%	50.9%
30th %ile:	19.9%	43.5%
35th %ile:	22.0%	41.4%
40th %ile:	25.3%	40.9%

The identification accuracy data in Table 2 indicate that the clustering technique was not superior to placement tests in terms of diagnostic sensitivity or specificity.

DISCUSSION

The results of the present study revealed distinctly different patterns of academic readiness and achievement among students in different empirically-derived clusters. Of special interest were significant differences on outcome variables, reflective of academic performance and persistence, that were not used to create the clusters. The strength of the clustering model overall was clearly suboptimal, however, based on the negative value of the cubic clustering criterion for the 5-cluster model chosen for further analysis. Although the distribution of persisters and non-persisters was significantly different across clusters, the hierarchical clustering patterns in this study do not appear to have sufficient predictive validity be useful as a screening method.

Scrutiny of these data suggests four recommendations for the further study of hierarchical clustering as a method for analyzing the attrition patterns of students. First, the addition of more and/or more powerful predictor variables in the cluster analysis should result in a positive CCC and a stronger model overall. Of special interest might be the inclusion of non-academic (e.g., socialization) factors that may influence many students' educational persistence. Second, the selection of a model at a higher CCC peak may yield smaller but more homogeneous clusters that include one or more predominantly high-risk subgroups. While larger cluster configurations increase the risk of Type I error, a manageable and reasonably "safe" configuration of 10-12 clusters may produce sufficiently high sensitivity and specificity ratings to be of practical usefulness.

Third, the study needs to be replicated using more samples from diverse institutions. If the cluster patterns prove not to be stable or generalizable, at least within the domain of community colleges, then their significance will be diminished. As in the cluster analysis of other human characteristics, the potential problems of non-representative samples or shifting populations could limit the reliability of the findings presented above. Finally, the meaningfulness of sub-groups identified by clustering techniques should be validated. Do the groups really differ from one another, and are they internally similar, in fundamental ways? The use of focus groups or structured interview techniques with selected clusters might shed further light on student sub-types and the possible predictability of their enrollment behaviors.

REFERENCES

Bean, J.P. & Metzner, B.S. (1985). A conceptual model of nontraditional undergraduate student attrition. Review of Educational Research, 55 (4), 485-540.

Bachelor, P. & Buchanan, A. (1984). Using cluster analysis to solve real problems in schooling and instruction. ERIC Research Report No. SWRL-TR-85.

Chapman, D. & Pascarella, E. (1983). Predictors of academic and social integration of college students. Research in Higher Education, 19, 295-322.

Lenning, O.T., Beal, P.E. & Sauer, K. (1980). Retention and attrition: Evidence for action and research. Boulder: NCHEMS.

Noel, L. (Ed.) (1978). Reducing the dropout rate. San Francisco: Jossey-Bass.

SAS Institute (1982). SAS User's Guide: Statistics. Cary, NC: SAS Institute.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45 (1), 89-125.