DEVELOPMENT AND VALIDATION OF DISCRIMINANT ANALYSIS MODELS FOR STUDENT LOAN DEFAULTEES AND NON-DEFAULTEES

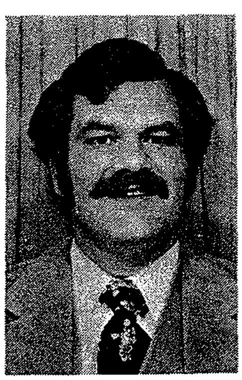
by Greeley Myers and Steven Siera

Introduction

In recent years default on guaranteed student loans has been increasing in magnitude at a rate alarming to financial aid administrators. According to USOE, the national rate of default was 4.3% in 1972, but jumped to 18.5% by 1975. While Hauptman (1977) attributes some of this change to differences in procedures for reporting and calculating default rates, the increase is still substantial.

At New Mexico State University, concern with the default rate on loans administered through the New Mexico Student Loan Program (NMSLP) derives from several sources. First is the fact that as of June 30, 1976, former NMSU borrowers had defaulted on 11.1% of the dollar amount of loans in repayment status. This led to a 5.6% reduction from the 1975-1976 allocation in the amount available to borrowers in 1976-1977. Concurrently, regulations were implemented to require that a student be making "satisfactory progress" toward a degree in order to be eligible to borrow from the NMSLP. Thus to borrow the maximum allowable amount, the student must pass at least 12 semester hours per semester with a 2.00 grade point average. If these criteria are not met, subsequent loans are reduced proportionately.

It is essentially unknown what specific characteristics differ for defaultees and non-defaultees. Hauptman (1977) reports that differences between percentages of loans received and percentage of defaults exist on the basis of family income, ethnicity, and the type and control of institution. Borrowers from low income families are more likely to default, as are those attending proprietary and specialized vocational institutions.



Greeley Myers has been Director of Student Financial Aid at New Mexico State University since August, 1967. He is active in the Southwest Association of Student Financial Aid Administrators and is a past president of the New Mexico Association of Student Financial Aid Administrators.

Steven Siera is Assistant Director of Educational Career Services at the University of California at Los Angeles. At the time the research for this article was completed, he was a Research Assistant in the Division of Student Affairs at New Mexico State University.

Dyl and McGann (1977) found that the following factors related positively to repayment of short term loans: grade point average, being married, being an engineering major, and size of the monthly payment. Factors negatively associated with repayment were: total amount of other university loans, residence in an apartment, and the size of the loan requested.

Dyl and McGann (1977) described the use of discriminant analysis as a technique to identify "good" versus "bad" student loans based on information available from the loan application. They then applied discriminant analysis to data from short term loans to demonstrate the technique. They found that 84% of cases were correctly classified by this procedure.

Before a discriminant analysis model is adopted for utilization in making decisions regarding the awarding of student loans, it must be demonstrated not only that it can classify the cases from which it was developed, but that when applied to cases other than those from which it was originally derived, it provides accurate predictions. The research by Dyl and McGann did not report such a validation. The research reported here was designed to test the ability of the analysis models to make such predictions.

Problem

In order to deal effectively with the problem of defaults, it is necessary to identify characteristics of defaultees and non-defaultees. If it is true that there are characteristics which substantially differentiate between defaultees and non-defaultees, it is possible to develop a model which will allow us to predict a borrower's probability of defaulting.

It was observed that two classes of characteristics could prove useful. First are those items which are available at the time loan application is made. Second, are events which occur subsequent to the application which might affect the borrower's likelihood of repaying. In terms of policy-making, the first set of information could be used in making decisions about loan awards, as well as for indicating intervention programs designed to lay the framework for later repayment for those students identified as high default risks. The second set of characteristics would indicate need for intervention programs for borrowers who later display a pattern associated with high default risk regardless of their initial characteristics.

Procedure

The procedure involved a statistical examination of data about students who have already exited school and have entered repayment status. Due to the improbability that defaultees would be particularly cooperative in an examination of their background characteristics, and to the fact that greatest interest was in finding indicators which would be available from the application, information from *OE Form* 1154 was used to investigate initial characteristics at the time the loan was made. Final transcripts yielded grade point average, number of hours passed, and whether a degree was earned. This latter information was used to identify educational patterns developing after the loan was made.

All students who had exited NMSU during academic years 1971-1972, 1972-1973, and 1973-1974, who had defaulted on their loan, and for whom complete records including final transcripts were available were included in the study. A total of 74 records were available. Seventy-four students exiting during the same

into the model in Table 5. The canonical correlation for this analysis of 0.643 indicates that about 41% of the variance is accounted for by the model. The classification of the 107 cases used in developing the model yielded the results shown in Table 6. The percentage of cases correctly classified was 82.2%
period who had entered repayment were included for comparison purposes. Information for one defaultee was later dropped due to illegible data on the application. Information included on the applications and transcripts of the two groups was analyzed via discriminant analysis procedures to identify variables which differentiate between the groups.

There is a likelihood that the predictor variables might not be linearly related to the default dichotomy. Therefore, variable transformations were utilized to explore the possibility of such a relationship. Variables formed from products of other variables were included to account for effects of variables having interactive effects with one another. The use of transformations and product variables leads to difficulties in interpretation. The decision to use such variables is predicated on the purpose of the study, which is to develop and empirically test a discriminant function which effectively predicts default. Interpretation is secondary.

Three basic analyses were performed. In the first analysis, data derived only from information available on the OE Form 1154 was included. In the second analysis, the information available from the students' final transcripts was included. The third analysis was performed by relaxing the usual criteria for inclusion of variables, thus creating a flooded model with a large number of variables.

For each of the analyses, a two-staged analysis was performed. Twenty each of defaultees and non-defaultees were excluded from the analysis at the model building stage. When the model was developed, the derived formula was applied to information from this group, and predictions for this known group were compared with their actual category. This tested the predictive success of the model. "All statistial procedures were performed by the SPSS procedure DISCRIMINANT (Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975).

Results

A complete list of all first-order variables is included in Table 1. Means for defaultees and non-defaultees, along with T-values for differences are included for each variable. No interpretation is made of the results of the T-tests. Their inclusion is solely to provide information for those readers seeking possible variables to consider for future study. Due to space limitations, only those higher order and product variables which entered into the analyses will be described in the text. The full set of variables included squares, cubes, square roots and inverses of selected variables as well as various products of variables.

The first analysis is that of data from the application only. The variables entered into the model are shown in Table 2. The canonical correlation of 0.530 indicates that only about 28% of the total variance is accounted for by the fitted model. The results of the classification of the 107 cases used to build the model are shown in Table 3. This represents a correct classification of 72.9% of those cases used in developing the model.

Table 1
LIST OF VARIABLES, MEANS AND T-VALUES
FOR DEFAULTEES AND NONDEFAULTEES

Variable	Mean Default	Mean Nondefault	DF	T-Value
Age at time of first loan	23.014	23.243	145	+0.20
Student's sex (male)[a]	.699	.676	145	—0.30
Ethnicity, Black[a]	.041	.014	145	—1.02
Ethnicity, Native American[a]	.027	.000	145	—1.43
Ethnicity, Oriental[a]	.000	.000	145	+0.00
Ethnicity, Spanish surnamed[a]	.247	.189	145	—0.84
Ethnicity, Other[a]	.685	.797	145	+1.56
Marital status (single)[a]	.630	.635	145	+0.06
Amount of loan requested	1019.56	1059.46	145	+0.56
Amount of other aid received	839.23	1105.80	26	+0.99
Educational debts	252.19	262.96	145	+0.09
Other debts	1053.23	931.50	145	—0.29
Dependent on parents[a]	.493	.432	145	—0.73
Separated from spouse[a]	.082	.041	145	—1.05
Father's gross income	9617.35	8270.79	60	—0.46
Mother's gross income	4498.86	4219.52	42	—0.34
Parents' joint income	10597.47	9695.56	72	—0.36
Student's gross income	2303.85	1930.08	98	—0.96
Spouse's gross income	3956.06	4358.96	40	+0.48
Student and spouse's joint income	3485.49	3769.39	105	+0.44
Family's adjusted gross income	7209.98	7206.27	145	—0.00
Adjusted family income	4157.82	4378.62	145	+0.34
Freshman at time of first loan[a]	.480	.365	145	—1.41
Sophomore at time of first loan[a]	.206	.249	145	—0.90
Junior at time of first loan[a]	.178	.162	145	—0.26
Senior at time of first loan[a]	.137	.284	145	+2.20*
Graduate student at time of first loan[a]	.000	.041	145	+1.74
Major in education college[a]	.206	.189	145	—0.25
Major in arts and sciences college[a]	.315	.378	145	+0.80
Major in engineering college[a]	.082	.135	145	+1.03
Major in agriculture college[a]	.164	.108	145	—0.99
Major in business college[a]	.178	.149	145	—0.48
Major undecided or in continuing education[a]	.055	.041	145	—0.40
Years until expected graduation	3.000	2.405	145	—3.09**
Estimated educational costs	2530.63	2645.78	145	+0.59
Cost minus other financial aid	2381.18	2464.61	145	+0.45
Need indicated by school	1001.52	1059.46	145	+0.77
Enrolled full time[a]	1.000	.986	145	+1.01
Currently enrolled at time of application[a]	.589	.770	145	+2.39*
Amount requested minus loan amount	786.14	0.00	6	—4.94**
Total amount of all loans	1399.66	1809.46	145	+2.73**
Number of hours passed	104.638	73.122	131	+3.50**
Final grade point average	2.062	2.761	130	+5.35**
Degree earned[a]	.328	.580	131	+2.98**

Note. Means with two decimal places are variables such as income, loan amounts and other financial data, and are expressed in dollars and cents.

[a] indicates a dummy variable which has a value of 1 if the characteristic is true for the individual, and 0 if it is not true. Means for dummy variates represent proportion of individuals in that category.

* $p \leq .05$.

** $p \leq .01$.

Table 2
SUMMARY OF DISCRIMINANT ANALYSIS
WITHOUT TRANSCRIPT DATA
MODEL 1

Variable	Order of Entry	F-Ratio to Remove	Standardized Discriminant Coefficients
Expected years until graduation	1	8.966	—0.546
Amount requested minus loan amount	2	6.321	—0.479
Dollar amount of total loans cubed times family's adjusted gross income	3	7.500	+0.501
Junior at time of first loan	4	4.097	—0.401
Dummy variable for separated from spouse times family's adjusted gross income	5	4.211	—0.382
Ethnicity, Other	6	3.906	+0.329

Overall Discriminant Function Characteristics
Eigenvalue = 0.391
Canonical Correlation Coefficient = 0.530
Wilks' A = 0.719 df = 6
$X^2 = 33.636$ $p \leq .001$

Table 3
CLASSIFICATION POWER OF MODEL 1

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	13	40	53
Repayment	38	16	54
Total	51	56	107

Table 4
PREDICTIVE POWER OF MODEL 1

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	10	10	20
Repayment	7	13	20
Total	17	23	40

When the data not used in developing the model was classified to validate predictive ability for the model, the classification shown in Table 4 occurs. Only 42.5% of the test cases were correctly predicted. The derived value of Chi square is 1.80 which is not significant for 1 degree of freedom.

When the data not used in developing the model was classified to validate predictive capacity of the model, the classification shown in Table 7 occurred. The percentage of test cases correctly classified was 57.5%. The value of Chi square is 3.40 which is not significant for 1 degree of freedom.

The second analysis is that of data from both the application and the final transcript. Variables entered in this analysis are shown in order of their entry

Table 5
SUMMARY OF DISCRIMINANT ANALYSIS
WITH TRANSCRIPT DATA
MODEL 2

Variable	Order of Entry	F-Ratio to Remove	Standardized Discriminant Coefficients
Grade point average squared	1	29.758	—0.663
Grade Point average squared times amount requested minus amount of loan	2	8,174	+0.387
Dummy variate for junior at time of first loan times inverse of GPA	3	7.324	+0.404
Dummy Variate for separated from spouse times family's adjusted gross income	4	5.027	+0.312
Total amount of loans cubed times family's adjusted gross income	5	4.899	—0.291
Inverse of GPA times expected number of years until graduation	6	4.423	+0.272

Overall Discriminant Function Characteristics
Eigenvalue = 0.703
Caninical Correlation Coefficient = 0.643
Wilks' A = 0.587 df = 6
x^2 = 54.323 $p \leq .001$

Table 6
CLASSIFICATION POWER OF MODEL 2

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	10	43	53
Repayment	45	9	54
Total	55	52	107

Table 7
PREDICTIVE POWER OF MODEL 2

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	6	14	20
Repayment	9	11	20
Total	15	25	40

The third analysis involved the use of a "flooded model" whereby the usual criteria of significance for entry of a variable were waived. The purpose of this was to attempt to increase the predictive capability of the model for the test cases. For this analysis, the F-ratio to enter was changed from 3.9 with an approximate probability of .05, to 1.0 with a probability of .50.

Fifteen variables were entered into the model under this condition as shown in Table 8. The canonical correlation for this analysis is 0.724, meaning that about 52% of the total variance is accounted for by the model. The classification of the 107 cases used to build the model is shown in Table 9. The percentage of these cases correctly classified was 82.2%.

Table 8
SUMMARY OF DISCRIMINANT ANALYSIS
FLOODED MODEL
MODEL 3

Variable	Order of Entry	F-Ratio to Remove	Standardized Discriminant Coefficients
Grade point average squared	1	29.758	—0.558
Amount requested minus loan amount	2	7.676	+0.372
Junior at time of first loan	3	5.767	+0.486
Dollar amount of total loans cubed	4	3.109	—0.311
Dummy variate for separated from spouse	5	4.219	+0.230
Inverse of GPA	6	3.191	+0.269
Adjusted family net income squared	7	3.887	—1.919
Ethnicity, Native American	8	3.348	+0.219
Expected years until graduation	9	3.292	+0.681
Senior at time of first loan	10	6.092	+0.415
Freshman at time of first loan	11	2.039	—0.419
Engineering major	12	2.382	—0.194
Currently enrolled at time of application	13	1.466	—0.198
Adjusted family net income cubed	14	2.099	+1.313
Family's adjusted gross income	15	1.229	+0.326

Overall Discriminant Function Characteristics
Eigenvalue = 1.101
Canonical Correlation Coefficient = 0.724
Wilks' Λ = 0.476 df = 15
$X^2 = 73.379$ $p \leq .001$

Table 9
CLASSIFICATION POWER OF MODEL 3

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	8	45	53
Repayment	43	11	54
Total	51	56	107

Table 10
PREDUCTIVE POWER OF MODEL 3

Actual Result	Predicted Result		
	Repayment	Default	Total
Default	9	11	20
Repayment	11	9	20
Total	20	20	40

When the test cases were included to check the predictive validity of the model, the classification shown in Table 10 occurred. The percentage of test cases correctly classified was 55%. The value of Chi square is 0.40 which is not significant with one degree of freedom.

Discussion

The results of the analyses indicate that the use of discriminant analysis with these variables does not lead to an accurate prediction of the likelihood of a student defaulting on a loan. Moderately adequate models for describing the data were derived, both in the current study and in the earlier study by Dyl and McGann. However, when these models are applied for the purpose of predicting "unknown" cases, prediction is not substantially different from what we might accomplish by chance. This is indicated by the nonsignificant Chi square values for the test classifications.

In interpreting the meaning of the variables included in the models, it is necessary to consider the variable entered, and sign (+ or −) of the coefficient in relationship to the variables previously entered. Interpretations of specific variables may become quite complex, which is one price we must pay for a model which can better classify and predict the likelihood of default.

For example, in analysis 1, the variable total loans cubed times family's adjusted gross income is positively related to repayment. In analysis 2, this same variable is negatively related to repayment. This results from the prior inclusion of one or more variables which are strongly correlated with the variable in question. In this case, it is probable that variables such as GPA squared (correlated with total amount of loans) and family's adjusted gross income for students separated from their spouse (correlated with family's adjusted gross income) account for the reversal of effect for this variable.

TABLE 11
CONSTRUCT GROUPING OF VARIABLES ENTERED
ANALYSES 1 AND 2

Variable	Entry Order Analysis 1	Entry Order Analysis 2
Group 1: Variables Related to Academic Success		
Expected years to graduation	1	—
Junior at time of first loan	2	—
Grade point average squared	—	1
GPA squared times amount requested minus amount of loan	—	2
Junior at time of first loan times inverse of GPA	—	3
Inverse of GPA times expected number of years until graduation	—	4
Group 2: Variables Related to Financial Condition		
Dummy variate for separated from spouse times family's adjusted gross income	5	4
GPA squared times amount requested minus amount of loan	—	2
Amount requested minus amount of loan	2	—
Total dollar amount of loans cubed	3	—
Total amount of loans cubed times family's adjusted gross income	—	5
Variable Not Related to Either Construct		
Ethnicity, other (Anglo)	6	—

Table 11 groups variables which entered into the models in analysis 1 and analysis 2. Only one of these variables fails to relate to the general constructs of either academic success or financial condition.

In examining those variables related to academic success, we find measures of grade point average appear to be most important. The higher the student's GPA, the more likely he is to repay the loan. Measures of academic success in terms of class status appear to be next importance. The closer the student is to graduation at the time of the first loan, and hence the more successful he has been, the more likely he is to repay the loan. These findings lend support to the conclusion that academic success is a substantial indicator of repayment probability.

The student with a low GPA is more likely to drop out of school, as is the student who is farther from graduation. The dropout may feel that he did not receive all that he expected from his college experience, and thus fails to repay the loan.

A similar feeling on the part of the student may be related to one of the financial conditon measures, that of the discrepancy between the amount requested and the amount of the loan. Only 7 of the 147 students had discrepancies, but all of those were students who defaulted on loans. There were two situations in which this occurred: either the student requested more than the maximum amount available under program regulations, or the analysis of need indicated substantially less need than the amount the student had requested. In either case the student could feel that he did not receive the benefits he expected, and therefore chose not to repay the loan.

Finally, measures of total amount borrowed are to some extent tied to the construct of academic success. The student who succeeds in school is more likely to secure additional loans in subsequent years and therefore ends up with a higher total amount.

It would appear from these findings that the current criterion that a student be satisfactory progress toward a degree is a valid one. The evidence supports such a conclusion. There is a need, however, for further research to confirm these findings, and to identify other predictors.

The implication of the study is that we are still unable to predict with accuracy the likelihood of default. Additional research needs to be conducted to identify those variables which are effective in making such predictions. Those financial aid administrators considering the use of discriminant analysis or similar procedures in decision making must ensure that the models used are empirically verified with their own students before they are adopted. It is essential that the predictive validity of a model be tested, not on the cases from which it is built, but rather on additional cases which were not included in the model building stage. Unless this validation is performed, there can be no confidence in the predictive capability of the model.

REFERENCES

Dyl, E. A., & McGann, A. F. Discriminant analysis of student loan applications. *The Journal of Student Financial Aid,* 1977, 7 (3), 35-40.

Hauptman, A. M. Student loan defaults: toward a better understanding of the problem. In L. D. Rice (Ed.), *Student Loans: Problems and Policy Alternatives.* New York: CEEB, 1977.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. *Statistical Package for the Social Sciences* (2nd ed.) New York: McGraw-Hill Book Company, 1975.