

# Using Discriminant Analysis to Identify Students at Risk

Edward W. Thomas<sup>1</sup>, M. Jackson Marr<sup>2</sup>, Adrian Thomas<sup>2</sup>, Robert M. Hume<sup>3</sup>, Neff Walke<sup>2</sup>

<sup>1</sup> School of Physics, <sup>2</sup> School of Psychology, and

<sup>3</sup> Office of Minority Educational Development

Georgia Institute of Technology

Atlanta, GA 30332

## Abstract

*In designing any educational intervention one often needs to determine what factors are related to success and failure in a course, identify students at risk, evaluate the impact of any new program on student performance. In this paper we will discuss the use of discriminant analysis as a technique for addressing all of these issues.*

*Discriminant analysis is a statistical technique designed to investigate the differences between two or more groups of people with respect to several underlying variables. This technique is more appropriate than commonly used educational measures (correlations, regression weights, etc.) because the variable being predicted is categorical. Moreover, this approach results in a unit of analysis, predicted category membership, that is more useful in evaluating instructional interventions.*

*We have used discriminant analysis to predict student performance in an introductory electromagnetism course at Georgia Tech. In this course there is a high failure rate (greater than 30% make a grade of D or F) which results in a great cost to the institute and to students as success in the course is a prerequisite for all engineering majors. A technique that could identify the factors that are predictive of course performance and identify students who are at risk would be of great benefit to the design, implementation, and evaluation of any educational intervention.*

*We will present a case study with over 1600 engineering majors where discriminant analysis was used successfully to determine the predictors of course performance. We looked at over 15 possible predictor variables (SAT, GPA, etc.) and determined which variables would be most useful in identifying students at risk. Using information available from student's records we were able to successfully predict 50% of the students who eventually failed the class.*

*We will also discuss how discriminant analysis can be used as an evaluation technique. We show that this approach provides results that are both more interpretable and statistically sound than traditional measures*

## Introduction

In designing an instructional intervention of any kind one needs to accomplish at least three tasks: 1) determine the factors that are relevant to successful performance of the task at hand, 2) identify those individuals most likely to benefit from the intervention, and 3) evaluate the impact of any new program on student performance. This paper discusses the use of discriminant analysis as a useful technique in addressing all three of these critical issues.

The first task mentioned above means a careful delineation of the problem. Given the large number of possible variables, it is not surprising that individuals have different ideas based on different assumptions about what causes subpar performance or performance enhancement. Laying these different ideas out "on the table" may provide the group as a whole an opportunity to attack the problem in the best way possible. Once this discussion has taken place an educated guess about the causes of the problem at hand can be offered. This paper will show how discriminant function can be used to help determine what variables have a relationship with performance, and how such relationships can be used to help shape interventions.

The second major task, identifying the appropriate individuals with which to use the intervention, may typically mean identifying students who might be termed "at risk". These are the students who are in danger of failing a class, not understanding a certain concept, etc. However, some interventions might be targeted at different levels of performance (for example, the "gifted students" might be chosen for some supplemental instruction).

It is in this student-identification task that discriminant analysis will be seen to be most advantageous over traditional approaches. We will show how we used discriminant analysis to pinpoint students “at risk” in a class where previous attempts at such classification had been less than ideal.

Finally, it is always critical to evaluate the intervention. This step has many purposes. The most obvious is that it lets the research team (and funding agent) know whether or not a given intervention worked. In addition, evaluation can be used to shape the research program and guide it toward a more effective intervention. Unfortunately, good critical tests of a program are sometimes hard to come by. We will attempt to show how one can use discriminant analysis to assist in this difficult task.

Several statistical approaches are available to help answer the above questions. The traditional approach to data analysis in this arena has been to use simple measures of correlation. This linear-based treatment has long been recognized as the simplest and most easily interpretable measure of effect (1). However, such measures are often misleading. They give no information about the relative effects of multiple measures, nor do they provide a simple means for assessing the cumulative affects of several variables when allowed to work in combination.

Most researchers next turn to some form of multiple regression (2). This technique determines the linear relationship between a set of predictors and a criterion in terms of the model  $Y = a_1X_1 + a_2X_2 + \dots + a_kX_k + b$ . Here, Y is a criterion, such as a course grade,  $X_i$  are predictors, such as SAT or grade in a previous course,  $a_i$ 's are weights associated with each predictor, and b is a constant. This technique provides a measure of association in terms of the amount of variance accounted for by each variable, along with an estimate of their combined effect on the criterion.

While multiple regression has long been used in educational interventions, there is one major problem. It's purpose and model's equations are based on the assumption that predictors and criterion are continuous in nature. Most variables used in educational interventions are not continuous. Course grades, minority status, sex, race, etc., are clearly discrete variables. Even such variables as class size and SAT are probably better characterized, statistically, as polytomous ones since SATs range only from 200 to 800 with 5 point intervals and (instructor complaints to the contrary) class size rarely approaches infinity. On the criterion side,

most variables in educational interventions are likewise categorical (i.e., grade received). Research has shown that violations of the assumptions underlying regression modeling can have serious repercussions(3). A more appropriate technique under the current circumstances is known as discriminant analysis.

Ronald Fisher developed discriminant analysis for use with categorical data. It is based on assumptions very similar in nature to multiple regression, except that it is designed for categorical criterion. While not specifically intended for use with categorical predictors, research has shown (4,5) that it performs quite well using such data. Discriminant analysis forms linear combinations of the predictors which are used to classify cases into the various groups of the criterion. One may conceptualize discriminant analysis in terms of evaluating the centroid of a group of cases. In the present context the student cases are placed in criterion “groups” depending on their final course grade. The mean value of a discriminating variable (e.g., SAT or a preceding course grade), or predictor, for the students in a particular group is evaluated. The bigger the difference between the mean values of the predictors related to the various groups, the more discriminating is that variable. Discriminant analysis simultaneously analyzes all of these mean differences and determines which predictors are most discriminating (based on backward probabilities). The key notion here is the breaking of the criterion group into separate identifiable units. This allows for discriminant analysis to better account for discontinuous relationships among the variables.

Note, that this view of discriminant analysis is very simplistic. In reality discriminant analysis is based on solving an equation such as

$$l'y = k'x + B,$$

where, l' and k' represent matrices consisting of all possible linear combinations of responses on the criterion and all possible linear combinations of the predictors, respectively. This results in linear combinations that are solved simultaneously. They are also subject to several classification rules. The mathematics involved quickly become quite complicated. Most statistical packages include a discriminant analysis procedure. For a more detailed analysis of the technique see Huberty (6).

Discriminant analysis can be used in the same circumstances as multiple regression. Given a list of potential predictors, one can determine which are most effective in predicting performance. It

provides a discriminant function which includes only those variables that should be used in predicting performance. Unlike regression, significance tests are provided for each possible variable and the equation as a whole. Probably the biggest advantage of discriminant function over regression is that its measure of predictive ability is in terms of the percent of correct classifications. This is possible since the unit of analysis is categorical. It predicts category membership. Given the true grouping of the criterion, one can determine how many predictions produced by the equation are right. This quantity (% correct) seems much more interpretable than % of variance accounted for in regression. It also seems more translatable to educators and administrators not familiar with statistics and the concept of variance partitioning.

In order to demonstrate the methodology we now present a case study from the Georgia Institute of Technology. Statistics across the country show that a high percentage of Engineering students fail to make satisfactory grades in the required introductory physics courses. We focus specifically on a one quarter Electricity & Magnetism course taught as a central component in a three quarter course sequence. At Georgia Tech up to 30% of all students achieve an unsatisfactory grade (D or F) in this course. Since engineering majors receiving such marks are required to repeat the course, this problem represents a great cost to both students and the Institute. Our goal was to analyze the situation and determine what parameters in the students' past record might represent criteria for identifying those who are likely to be at risk of obtaining unsatisfactory grades.

We set out to accomplish this task with the help of discriminant analysis using historical data (1622 students enrolled in Professor A's class over a five year period). The first step was to determine in what ways students who performed satisfactorily differed from those who did not. Using data from student records we looked at many possible predictor variables including the following; ethnic group, gender, major, college, SATs, Overall GPA, Math GPA, and their performance in specific classes thought to be predictive of performance (chemistry, the preceding physics course, etc.). Univariate F-tests showed that all academic performance variables had relationships with performance in electromagnetism. However, discriminant analysis found that only three of these variables made significant independent contributions. These were the students overall GPA, grade in the calculus course dealing with the differential calculus of

functions and curvilinear motion, and grade in particle dynamics which is the physics course previous to electromagnetism.

Actual Grade	Predicted At Risk (F,D)	Predicted Not at Risk (C,B,A)	Total
F	144	29	173
D	175	104	279
C	203	291	494
B	114	328	442
A	125	209	234
Total	661	961	1622

*Table 1. Predicted and Actual Grades For One Instructor Over Multiple Years*

The next question is how successful these three variables are at predicting student performance. Discriminant analysis, like regression, produces a discriminant function equation that can be used for this purpose. Based on these three variables the discriminant function equation classified students as being "at risk" (predicted grade of D or F) or not at risk (predicted grade of A, B, or C). As can be seen in Table 1, 319 of the 661 individuals predicted to make a D or an F did so. This represents a hit rate of 48%. Also, of 961 individuals predicted not to be at risk, some 828, or 86%, achieved a satisfactory grade of C or better. Note that the greatest misclassification occurs for "C" students. It is interesting that discriminant analysis is highly successful in identifying these students who will perform successfully and only 50% successful in identifying students who will perform poorly. A review of the students' records shows that the better students have consistent records. The weaker students have erratic records. The data sample is inevitably skewed and cannot represent a "normal" distribution as the students with the very weakest records will have already left the system at an earlier stage and are not part of the analyzed group.

We used the above in several ways. First, these three discriminating parameters serve as a pointer for possible reasons underlying poor performance in electromagnetism. This helped confirm the groups previous suspicions that students' inability to understand, retain, and apply basic concepts

learned in particle dynamics and the Calculus was a major reason underlying poor performance in electromagnetism. It is easy to see how such an analysis would be useful in designing an educational intervention. Without the discriminant function to direct the research program one might have reached a vastly different conclusion regarding the etiology of poor performance (for example, an early theory was a lack of motivation on the part of the students which now seems mostly unfounded). Thus, with this example we see how discriminant analysis can serve as a valuable tool in determining the causes of poor performance and preventing the research team from pursuing fruitless interventions.

The next step in using discriminant analysis was to discover if we could use the discriminant function derived from the analysis of the single faculty member's classes to identify students at risk in future sections of the course. To do this the initial discriminant function, with three significant variables, was applied to a new group of students entering the course. In this case the students were instructed by four different faculty members in four separate course sections. The results are presented in Table 2. Success in identifying students not at risk was 84%, essentially the same as the historical sample. Success in identifying students at risk was 50%, again similar to the historical sample.

Actual Grade	Predicted At Risk	Predicted Not at Risk	Total
F	40	6	46
D	68	46	114
C	78	90	168
B	26	116	142
A	5	59	64
Total	217	317	534

Table 2. Predicted and Actual Grades For A Subsequent Class.

The use of this technique was more successful than previous attempts at classification. Use of the discriminant function in this manner allowed us to target our future interventions on the students who they were designed to help (i.e. those at risk).

The third use of discriminant function is in evaluation of educational interventions. In electromagnetism classes where there were no other interventions we have seen a consistent relationship between the three predictor variables and

performance. This consistency might be useful as a basis of comparison. In most educational settings it is difficult to offer a beneficial intervention to one group and set aside a true control group. The possible repercussions of such actions should be obvious. If a consistent baseline has been established via discriminant function, it is feasible to substitute this for a control group. Thus, if we offer an intervention to the entire class we can compare their performance with that of past classes. The three predictor grades essentially control for differences in the quality of students. Once again, there is no obvious way that such a system could be used with regression since error variance is inseparable from individual data.

Discriminant analysis should be the preferred method of operation in educational interventions regardless of the other benefits provided. As we have seen, it provides other benefits in addition to being the statistically correct procedure. Data are reduced more efficiently, and non-predictive variables are eliminated earlier in the analysis process. Student at risk are more reliably identified in more easily identifiable terms (i.e. predicted to fail versus predicted to pass). Discriminant analysis also allows for a more detailed analysis of errors or prediction than does regression, and does so with much more meaningful measures of effect (% correct predictions, or as confidence intervals with statements such as this person has a 70% chance of not passing the class). Finally, discriminant analysis can serve as a better basis for comparison than regression analysis for situations where control groups are not feasible.

## Acknowledgments

Supported in part by NSF grant DUE - 9455470 and by the SUCCEED Engineering Coalition

## References

- [1] Hedges, L. V. "Themeta-analysis of test validity studies: Some new approaches. In H. Wainer & H. I. Braun (Eds.). *Test Validity*, Erlbaum, pp. 191-212.
- [2] Lewis-Beck, M. S. "Applied Regression: An Introduction," Sage Publication. 1980.
- [3] Cook, R. D., and Weisberg, S *Residuals and Influence in Regression*. London: Chapman and Hall, 1982.
- [4] Gilbert, E. S. "On discrimination using qualitative variable," *Journal of the American Statistical Society*. 1968, p. 1399

- [5] Moore, D. H. "Evaluation of five discrimination procedures for binary variables," *Journal of the American Statistical Society* 1973, pp. 399-404.
- [6] Huberty, C. J. *Applied Discriminant Analysis*. New York, NY: JohnWiley, 1994.

