81-3

A Series of Occasional Papers in Draft Form Prepared by Members of the Research Department for Review and Comment.

STAFF MEMORANDA

AN EXAMINATION OF THE CONCEPTUAL ISSUES INVOLVED IN DEVELOPING CREDIT SCORING MODELS IN THE CONSUMER LENDING FIELD

Alan K. Reichert

FEDERAL RESERVE BANK OF CHICAGO

An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models in the Consumer Lending Field

by

Alan K. Reichert
Federal Reserve Bank of Chicago

Chien-Ching Cho
The MITRE Corporation, Boston

George M. Wagner
Lincoln National Bank
Fort Wayne, Indiana

The views expressed herein are solely those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Chicago or the Federal Reserve System. The material is of a preliminary nature, is circulated to stimulate discussion, and is not to be quoted without permission.

Table of Contents

List of Tables

List of Figures

An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models in the Consumer Lending Field

Introduction

The dramatic growth in consumer debt coupled with the increasing concern regarding discriminatory lending practices has generated considerable interest on the part of lenders in developing statistically based credit scoring models. The development of such models has evolved over the past ten years from relatively crude measures of credit risk to highly sophisticated broad-based management information systems employing a variety of advanced multivariate statistical techniques. This paper discusses the application of the most widely used of these statistical models--multiple discriminant analysis (MDA).

As widespread as it has become, the application of multiple discriminant analysis is not without its problems and limitations. Recent evidence indicates that the statistical assumptions and data requirements of the general MDA model are not consistently being met in a wide range of research applications [13]. In certain instances a misspecified model that violates one or more of these assumptions may produce inaccurate and misleading results. The objective of this paper is to see if MDA techniques can in fact be properly and rigorously applied to an actual lending situation. The study attempts to put these theoretical requirements into their proper perspective and illustrates the nature and extent of the bias one is likely to encounter when these theoretical assumptions are not fully achieved.

Section I discusses the growing importance of credit scoring models in the consumer lending field and illustrates how such models are frequently employed. Recent legislation intended to insure non-discriminatory credit decisions is reviewed, with special attention given to statistical credit granting systems. Section II briefly discusses the theory of multiple discriminant analysis and describes a number of the conceptual problems frequently encountered in the use of MDA techniques in applied research. Section III describes the sampling procedures, data sources, and variable descriptions involved in developing the credit scoring model. Section IV examines a number of the conceptual issues identified in Section II and attempts to demonstrate the degree of importance associated with each assumption. Finally, Section V summarizes the findings and indicates a number of areas where further research and empirical analysis are warranted.

Section I

Literature Review

The application of mathematical and statistical models to the credit granting decision can be traced back to the early 1940s when Durand [11] first employed discriminant analysis to measure credit risk. Sporadic attempts to investigate statistical credit systems followed but it wasn't until the mid-1960s that the credit industry became quite serious about developing credit scoring models. The tremendous growth in aggregate consumer credit and the increasing number of individual borrowers necessitated the development of some form of automated credit evaluation system. The widespread popularity of both bank and retail credit cards has added a new high risk dimension to the credit decision. More recently, inflation and recession have plagued a growing number of borrowers, requiring more frequent and careful assessments of changing borrower conditions. Expanding government involvement in credit regulation has generated further interest in statistically based credit granting systems of one type or another.

Not surprisingly most of the early scoring systems employed unsophisticated statistical techniques. The early models frequently began with a simple "good" versus "bad" customer profile to identify unique borrower characteristics. Arbitrary points or weights were then attached to individual characteristics and a composite score or risk index was calculated for each potential borrower. A cut-off value was established to separate potentially good and poor credit risks[5]. While the early attempts to model credit behavior appear somewhat crude when compared to today's advanced multivariate statistical techniques, they consistently identified several tangible management benefits associated with numerical credit scoring systems [2,3,6,8,9,10]. These benefits can be categorized into five major groups: (1) As training aids automated credit scoring models are being used to help new lending personnel identify and weigh important credit factors. These systems function as a check device to insure that correct decisions are being made throughout the training period; (2) Credit scoring models have been used to evaluate the performance of experienced credit personnel and to identify unacceptable behavior as quickly as possible. This function becomes critical as the size of the credit operation expands and branching necessitates decentralized lending authority; (3) Credit scoring models can lead to more effective management control and a quicker response to changing economic and credit policy. For example, a decision to tighten credit terms can be readily and consistently implemented by simply raising the credit score cut-off value to a suitably higher level. Thus, all credit personnel have a well-defined measure as to exactly how much credit conditions are to be restricted; (4) Credit scoring is also seen to play an important role in the marketing of credit services since it provides an objective measure of the customer's past borrowing behavior and can be used to target new promotional activities towards the bank's preferred customer base; and (5) Since most practitioners employ statistical credit scoring techniques in an adjunct fashion to their established lending procedures, ensuing economic benefits relate to the cost savings which accrue to the quick identification of clearly good and bad credit risks. This allows seasoned credit personnel to concentrate on the marginal credit applications.

Today, hundreds of commercial banks and consumer finance companies employ some type of statistical credit evaluation model. Outside the financial sector the recent growth in credit cards has encouraged major retailers such as Montgomery Wards and Sears to develop highly sophisticated credit scoring models to evaluate the hundreds of thousands of card applications received each year. For example, each of the approximately 900 Sears stores offers two types of credit card plans. A separate credit scoring model is estimated for each charge card plan for each individual Sears store. These roughly 1,800 credit scoring models are the primary means of determining whether or not a card is to be issued and are used to establish the appropriate credit authorization limit. The models are reestimated periodically as general economic and credit conditions change.

The recent passage of the Equal Credit Opportunity Act (ECOA) and subsequent amendments forbids the use of such factors as race, color, nationality, sex, marital status, age, and receipt of public assistance monies as factors in the credit granting decision [16]. Growing concern has developed throughout the credit industry regarding the lender's ability to realistically identify nondiscriminatory variables that can effectively differentiate between potentially good and bad borrowers. To the extent that such differentiation is possible it becomes the obligation of the lender to explicitly identify and legally justify the reason(s) which led to the denial of credit. The act offered at least one avenue of relief when it specified that lenders can employ credit scoring techniques that are "demonstrably statistically sound" and "empirically derived" if they do not include those factors explicitly prohibited.

The Board of Governors of the Federal Reserve System (Board) has been given the responsibility for developing the criteria for determining what constitutes a statistically sound and empirically derived credit scoring system. A credit system can be either empirical or judgmental in design. In a judgmental system, the selection of the credit variables and their respective credit weights (points) is based primarily upon the past credit experience or subjective judgment of the institution's lending personnel. The Board considers scoring systems to be "empirically derived" if they evaluate an applicant's creditworthiness by distributing points to specific applicant attributes on the basis of an empirical comparison of the institution's actual experience with both creditworthy and non-creditworthy borrowers. The Board's ruling does not specify the precise statistical methodology for making the necessary empirical comparison but requires the procedure to be reasonably effective in terms of predictive accuracy.

To be classified as "demonstrably statistically sound" three criteria must be met:

(1) The credit data used in system development should represent either the institution's entire population of applicants or a properly drawn sample from this population that would include both accepted and rejected applicants.

(2) Prior to implementation the predictive accuracy of the credit scoring system must be validated using actual applicant data. That is, the system must be found capable of distinguishing between creditworthy and non-creditworthy applicants at a statistically significant rate (to date no specific statistical test or level of significance has been determined by the Board).

(3) Once the system has been made operational, the model must be revalidated at appropriate time intervals to insure that the applicant attributes and the associated scoring points currently in use are still appropriate. Once again, no precise time interval is specified but logic would suggest that revalidation is appropriate whenever there has been a substantial change in underlying economic and credit conditions.

Section II

The Theory and Practice of Multiple Discriminant Analysis

Discriminant analysis is the study of differences between two or more groups. As Orgler points out, "discriminant analysis. . . first distinguishes among groups and identifies group differences; second, it classifies existing and new observations into predetermined groups, finally, it identifies the key variables that contribute the most to the discrimination among groups" [20]. Assume that we have two basic groups, A and B, for which we feel that two variables, X_1 and X_2, will provide substantial discriminatory power. That is, information concerning characteristics X_1 and X_2 in a specific case will enable us to correctly classify that case into either group A or B. Furthermore, the group assignment procedure may frequently be simplified by forming a linear combination of the discriminatory variables X_1 and X_2. The value of the weighted sum is then used to classify individual cases into specific groups.

For example, assume that the distribution of groups A and B for variables X_1 and X_2 is as depicted in Figure 1 below.

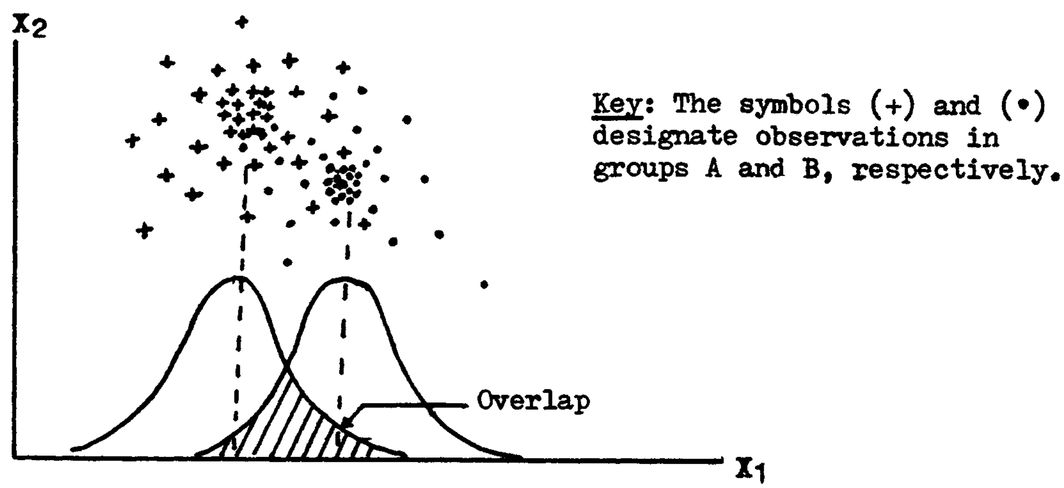

Figure 1. Two Group Z-Scores Projected on Horizontal Axis (X_1)

It is important to note that the groups overlap to a considerable extent. If variable X_1 was employed solely to discriminate between the two groups, the results would likely be rather poor. This is illustrated in Figure 1 where groups A and B are assumed to be distributed in a bivariate normal manner with maximum density at the center (or centroid) of each cluster of points. Projecting both clusters onto X_1 generates two normal distributions with substantial overlap.

A similar set of overlapping distributions would result if both groups were projected on X_2. Thus, neither X_1 or X_2 individually represents a powerful discrimination since each leads to a significant amount of group overlap. On the other hand, it may be possible to define a unique linear combination of X_1 and X_2 which will lead to greater group separation. For example, define

$$Z = d_1X_1 + d_2X_2$$

where d_1 and d_2 represent relative weights for the discriminatory variables and Z represents the weighted discriminant score used in making the actual group assignment. Figure 2 illustrates the projection of groups A and B on the newly defined Z axis. Note the substantial reduction in group overlap.

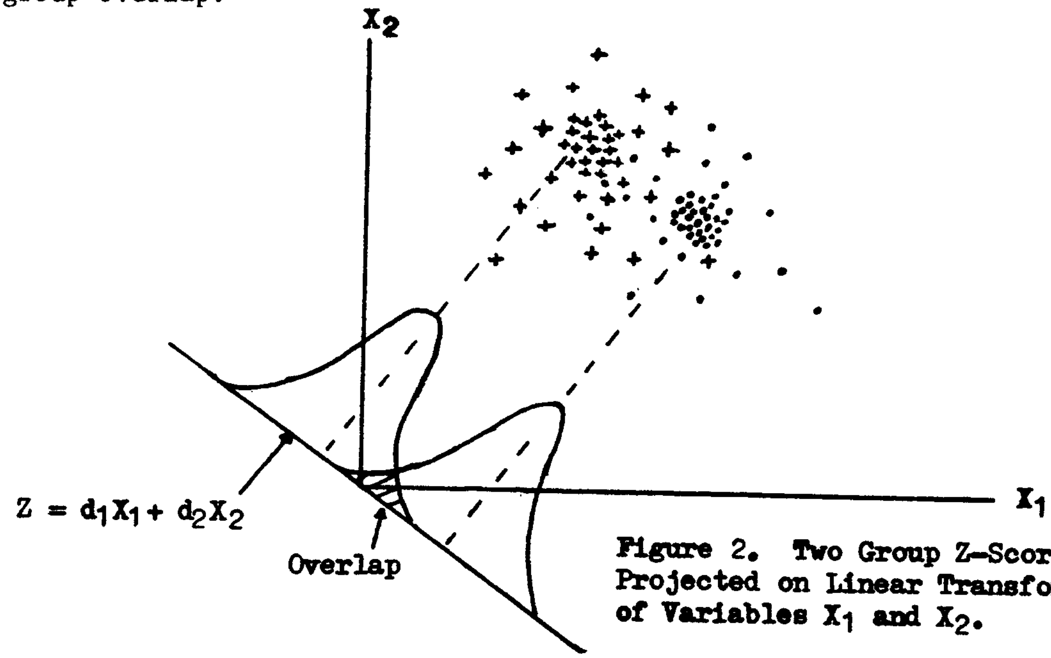

Figure 2. Two Group Z-Scores Projected on Linear Transformation of Variables X_1 and X_2.

While an infinite number of linear combinations are possible, only one unique combination will serve to provide maximum group separation. The greater the group separation along the Z axis, the more powerful the discriminant equation is in classifying observations into their respective groups. Although various measures of overlap are possible, assume that "f" is our measure of separation (nonoverlap) and f is a function of our discriminate weights d_1 and d_2. The function $f(d_1, d_2)$ can be differentiated in such a manner that yields the set of values for d_1 and d_2 which maximize the value f and hence lead to maximal group separation. In this sense, the linear function serves to reduce the bivariate problem to a univariate "Z-score" which forms the basis for subsequent group classifications.

In a more general model starting with "n" possible predictors or discriminant characteristics, step-wise computational procedures have been devised which select that unique subset of variables that forms the basis for the final discriminant equation. Thus, discriminant analysis serves to both identify and weight the proper discriminant variables and classifies observations into predetermined groups.

Several recent articles have identified various theoretical problems frequently encountered in the application of multiple discriminant analysis to credit scoring and other related topics. Eisenbeis, for example, identifies at least six broad categories of potential statistical problems as follows [13]:

(1) Virtually all multiple discriminant programs assume that the discriminant variables are distributed within each major classification group in a multivariate normal fashion. Significant departures from normality can severely distort the final classi-

fication results.

(2) The proper specification for most credit scoring systems requires that the group definitions in the estimation sample be identical to the group definitions in the underlying population. Hence, a three-group model is generally most appropriate since it includes both good and bad borrowers plus rejected (turned down) applications.

(3) Dichotomous variables are common in credit-related data and clearly violate the multivariate normal density function assumptions generally employed in calculating probabilities of group membership.

(4) When validating the predictive accuracy of a credit scoring model, a classification sample that is statistically independent of the estimation sample should be employed. Attempts to validate the model by classifying the estimation sample generate biased results.

(5) In order to minimize the model's misclassification error rate, the classification function should be weighted by the a priori probability of group membership. Failure to incorporate a priori probabilities in the classification process will not minimize prediction errors and frequently has a severe disruptive effect on between-group classification results.

(6) The linear discriminant model assumes equality between the variance-covariance matrices across all groups. Failure to test and adjust for inequality in the dispersion matrices makes inefficient use of the available credit information and can severely distort the classification results.

Section IV of the paper examines each of these six potential problems in the context of a realistic lending situation. Results are presented both before and after the aforementioned assumptions are met to give the reader an appreciation of the general nature and importance of each of these issues. It should be pointed out that the following results are based upon a relatively small sample of consumer loan data collected at a single point in time. Thus, the reader is cautioned against over-generalizing the findings.

Section III

Sampling Procedures and Model Development

Consumer loan data were obtained from the credit files of a medium-sized commercial bank. A sample of 405 closed loans and 243 rejected applications was gathered during the last half of 1977 and early 1978. The data were taken from standard loan applications submitted for a wide range of consumer goods and services such as automobiles, household furniture, vacation, and educational loans, etc. While documentation for a large number of closed loans proved to be incomplete and inadequate for credit scoring purposes, the sample identified above includes only those applications where complete documentation was available. The same condition holds for the rejected loans except that many of these applications were rejected before contract length or loan payment amount was determined. These missing values were estimated using credit terms appropriate for the type of loan requested.

Lenders typically employ two general criteria to guide them in the credit granting decision--the potential borrower's ability and willingness to repay. Ability to repay is usually measured by comparing borrower income and current debt requirements to ascertain the applicant's ability to service additional levels of debt. In addition, information regarding the applicant's employment history is frequently used to measure the stability of future earnings. Measures of financial wealth, such as bank deposits, ownership of stocks and bonds, etc., may be considered important but frequently are not scored since the item being purchased on credit is usually the only asset held to collateralize the loan. Willingness to repay is determined from the applicant's previous credit history and various measures of character, such as job and residence stability. Furthermore, the overall profitability of the customer relationship is often considered as well as the general riskiness of the loan terms being negotiated. For example, credit applicants having deposit relationships with the lending institution may possibly be treated more favorably, while loans with longer maturities and those generated through indirect dealership arrangements are typically viewed as higher risk loans.

Explanatory Variables

As previously discussed, predictor (discriminant) variables for the analysis were taken directly from the standard credit applications required of all borrowers. Lending officers of the subject bank were asked to identify the most important credit information collected during the lending process. The following six variables were selected:[1]

Credit rating. The first and perhaps most obvious variable included was past credit rating. Ratings were available to loan officers through both the local credit bureau and the bank's own records. The subject bank uses an alphabetical rating scale with a wide range of possible grades: AA (as agreed), A (good), AC (good to fair), AD (good to slow), C (fair), CD (fair to slow), D (slow), DD (very slow) and DDX (very, very slow; no further credit to be extended). No previous experience is graded NE, which includes new accounts with less than six monthly installments. To insure maximum consistency, all payback grades and classifications were

---

[1]A seventh variable designed to measure the importance of the savings deposit relationship was included in the initial model but proved to be of little significance and was deleted from the model.

made by the same individual.

To simplify the analysis, the credit rating variable in this model is based on a scale of 1 to 8 (8-AA, 7-A, etc). Since ratings AD and CD both suggest a significantly deteriorating credit situation, the two ratings were combined in the analysis and assigned a value of 4 (CD-fair to slow). No previous credit was similarly assigned the value of 4 as the majority of loan officers stated that lack of past credit would normally be treated with the same precautions as a somewhat questionable rating.

Relative debt load. The cash flow position of the applicant was broken down into the following three component parts: (1) monthly gross income, (2) existing monthly debt load, and (3) the requested monthly loan payment amount. All three variables were identified as potentially important factors since they are frequently employed to give the loan officer a realistic picture of the extent to which the applicant's income is being absorbed by fixed debt obligations, plus the amount of money remaining for variable living costs (food, utilities, etc.) and unexpected expenses. For example, high income alone has little meaning if most of it is paid out in fixed charges. On the other hand, a high loan payment may be easily met if a customer has no other debts. The final measure of relative debt burden (RELDEBT) is the sum of the existing monthly debt expense plus the requested loan payment divided by monthly gross income, all multiplied by 100 percent [(existing debt expense + requested loan payment)/total monthly income]X 100%.

Job/Age and Residence/Age. Two measures of borrower character and stability were identified: time on the job and time at present residence.

Both values are important to loan officers since many applications are rejected on the basis of insufficient employment experience or failure to meet a minimum residency requirement (usually a six-month minimum). In earlier formulations of the model, borrower age and years at present job and residence were simultaneously considered. Undesirably high correlations between these variables produced illogical results and led to their respecification. Both of these variables were subsequently divided by age to yield two new variables--Job/Age and Residence/Age. This formulation has the advantage of reducing the high correlations and provides a relative measure of borrower character and stability. When job and residence were entered in actual years, older borrowers were unduly favored while younger borrowers were unfairly penalized. After adjusting for age, both the young and older borrower start on a more equal footing.

Indirect-direct lending (IORD). Since personal interviews with loan officers are generally considered important when judging credit risk, a variable was included to distinguish between direct and indirect loans. Indirect loans represent dealer paper purchased by the bank in which the loan officer frequently reviews a less than complete application and is forced to make a loan decision without the benefit of a personal interview. This lack of personal contact, plus the possibility that the dealer may be more interested in making a sale than a high quality loan, suggests that indirect loans in general would be more risky than direct loans. For computational purposes, a direct loan was assigned a value of + 1, while indirect loans were assigned a value of 0.

Number of payments in contract (NCONT). This variable takes into consideration the uncertainties of life and payback capacity by measuring the number of payments in the loan contract. It seems logical that the longer the original contract maturity, the greater the likelihood that unforeseen events could prevent the loan from being paid back as originally agreed. A full range of installment maturities is included in the sample.

Dependent or classification variable. The precise distinction between a good and bad loan is somewhat subjective. A loan where the borrower paid every stipulated payment in full and on time throughout the term of the contract would undoubtedly be classified as a good loan and the customer would certainly qualify as a creditworthy borrower. On the other hand, one can legitimately question whether one or two late payments during a three- or four-year loan indicates a bad or unprofitable lending decision. Hence, the actual designation of a loan as either good or bad is a somewhat arbitrary decision. When a consumer loan at the subject bank was closed out, its complete payment history was reviewed and a credit rating was assigned according to the schedule previously discussed. To insure consistency, all loans included in the sample were rated by the same individual and loans with grades AA (as agreed), A (good), and AC (good to fair) were deemed to be acceptably profitable and hence classified as good loans. The remaining loans with credit ratings less than AC were considered to be unsuitable and designated as bad loans.

Section IV

Conceptual Problems and Issues

Proper Group Definitions

One of the most common errors in the early development and application of statistical credit scoring models was the exclusion of major segments of the applicant population from the analysis. Since the crux of the credit scoring effort is to be able to distinguish between creditworthy and noncreditworthy applicants, excluding a major segment of the population is wasteful of potentially useful applicant information and, more important, it can lead to biased results. Joy and Tollefson [17,18] and Eisenbeis [13] emphasize the importance of insuring that the sampling procedure employed and the purpose to which the discriminant model is put are consistent. For example, previous studies frequently employed a model based upon samples of both good and bad accepted loans to estimate the likelihood that a given applicant would ultimately prove to be a good or bad credit customer. More appropriately, a sample of all "through-the-door" applicants, which would include a sizable number of both accepted and rejected loan applicants, should be used to estimate a three-group discriminant model designed to serve as a general screening device for all potential bank customers. A two-group, "bad" versus "good" accepted loan model would more appropriately function as an internal loan review system designed to monitor the bank's portfolio of existing loans. Eisenbeis [13], for example, reports that researchers have found that using a model based solely upon the truncated population of accepted applicants can frequently generate misleading results. Use of truncated samples can lead the researcher to incorrectly adopt the quadratic form of the classification function when the individual group's variance-covariance matrices are in fact similar. The final result of using truncated samples is to generate biased estimates of 1) the group means (centroids) and dispersion matrices, 2) the actual group cut-off point, and 3) the true misclassification error rates.

As discussed in Section III, the model developed here is a three-group model designed to identify important differences between rejected loan applicants and both good and bad borrowers. Thus the model is designed to make predictions regarding the entire population of "through-the-door" applicants. In order to illustrate the importance of proper group specification, the results of a two-group model are presented in Table I and contrasted with the results obtained from the more general three-group model.

The standardized discriminant coefficients for the two-group (good vs. bad) model and the three-group model are presented in Table I-A. All coefficients carry the appropriate sign, and the relative importance of the discriminatory variables are similar except for JOBAGE and RESAGE.[2] A comparison of their predictive capabilities in Table I-B indicates that the two models are comparable in terms of overall classification accuracy. The model is also reasonably accurate in identifying good loans, correctly classifying approximately two-thirds of the sample, while the remaining one-third of the good loans were equally misclassified into the bad loan and turndown groups. The model correctly classified approximately three-fourths of the bad loans in the sample. It will be noted that the addition of the turndown category had very little effect upon the model's good loan prediction results. Only three out of the 103 loans initially predicted

---

[2]The variable IORD was initially deleted because of its dichotomous nature but was included in a later section.

Table I. Two- and Three Group Models

A. Standardized Discriminant Coefficients

Variable	Two-Group (Good vs. Bad)	Three-Group[a] (Good, Bad, and Turndown)
CREDIT	-.994	-1.485
JOBAGE	-.008	- .161
RESAGE	-.266	- .008
TOTDEBT	.263	.187
NCONT	.114	.163

B. Classification Results[b]

Two-Group (Good vs. Bad)

Predicted Group

Actual Group	#	Bad	Good
Bad	51	38 (74.5)	13 (25.5)
Good	148	45 (30.4)	103 (69.6)

Overall Accuracy[c] = 70.9%

Three-Group (Good, Bad, Turndown)

Predicted Group

Actual Group	#	Bad	Good	Turndn
Bad	51	13 (25.5)	13 (25.5)	25 (49.0)
Good	148	23 (15.5)	100 (67.6)	25 (16.9)
Turndn	126	6 (4.8)	3 (2.4)	117 (92.9)

Overall Accuracy = 70.8%

[a]Second discriminant function is not statistically significant.

[b]The classification results reflect a 50 percent holdout testing procedure and the assumption of equal group prior probabilities. Table elements represent both the number and percent of loans in the actual groups classified into each prediction group (percent in parentheses).

[c]Overall accuracy is measured down the diagonal of the classification matrix.

to be good were reclassified as turndowns. On the other hand, 25 out of the 38 loans initially predicted to be bad were reclassified as turndowns. Furthermore, the model was quite accurate in predicting turndowns, correctly classifying 93 percent of the sample.

The logic behind these results is made clear when one examines Figure 3 where the classification scores and group centroids are displayed. The classification scores provide substantial separation between the good loan centroid (-1.01) and the turndown centroid (1.47), with the bad loan centroid (-.26) falling in between. It seems plain that while the bad loan group possesses characteristics similar to both groups, on balance it appears to be more closely related to the good loan group. This is, of course, the same conclusion reached by the loan officer since credit was in fact extended to both the good and bad loan groups. Thus, the loan officer apparently has the same difficulty in distinguishing between the two groups as does the model. Given the unique characteristics associated with the sample of rejected applicants, the introduction of the turndown group has not improved the ability of the model to identify good loans nor has it reduced the number of bad loans predicted to be good. Instead it has simply split the bad loan group in two--those predicted to be bad and those predicted to be turned down. In fact, when one isolates the bad versus good prediction matrix within the three-group model (i.e., the elements within the dashed boundary in Table I-B, the model's overall accuracy declines to approximately 57 percent. The significantly higher overall accuracy rate of 70.8 percent simply reflects the introduction of the highly predictable turndown group.

The problem in distinguishing between good borrowers and bad borrowers

DISCRIMINANT SCORE 1 (HORIZONTAL) VS. DISCRIMINANT SCORE 2 (VERTICAL). * INDICATES A GROUP CENTROID.

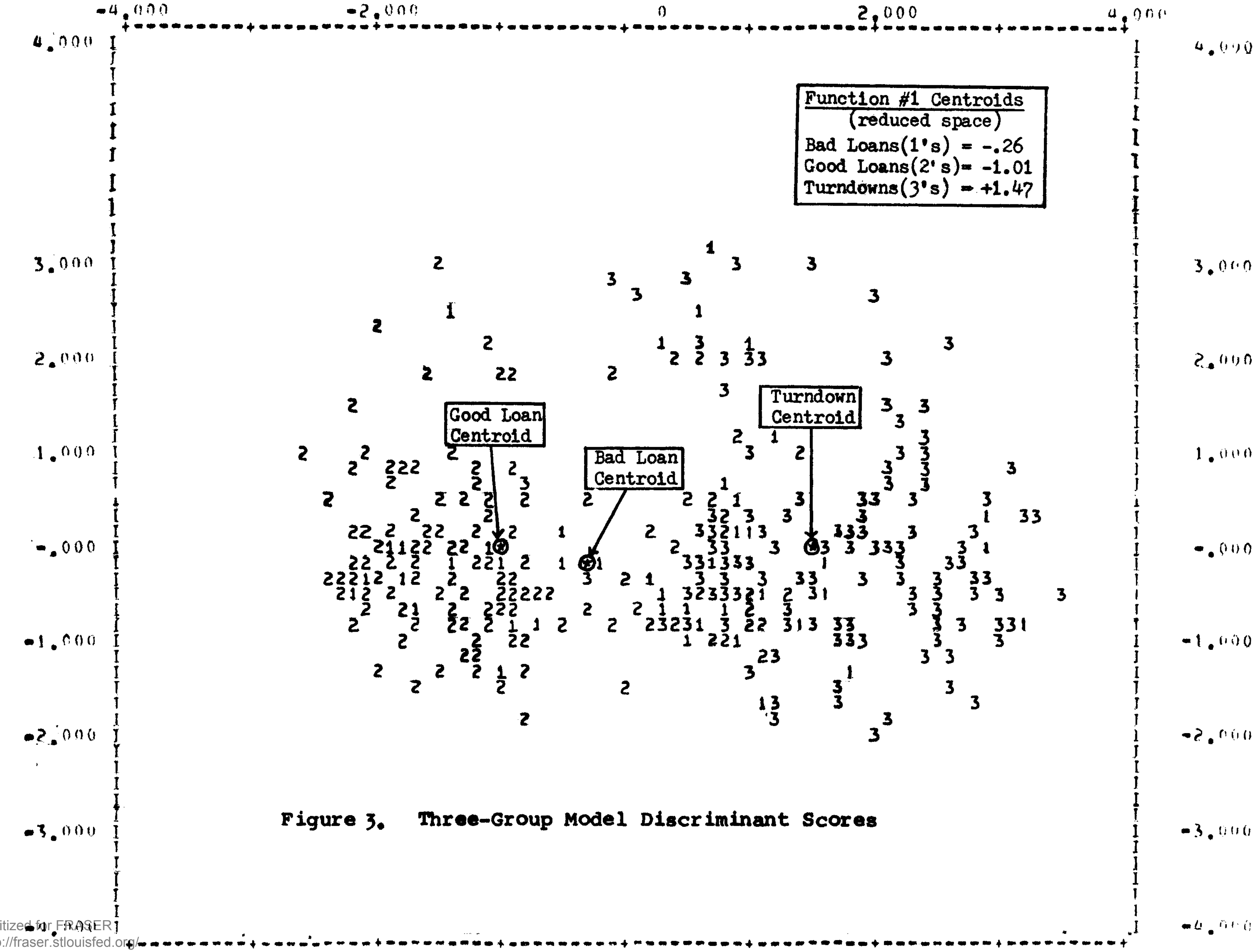

Figure 3. Three-Group Model Discriminant Scores

is visually illustrated in Figure 3 by the extensive overlapping of the discriminant scores for good and bad loans. The introduction of the centrally clustered turndown group appears to provide little additional information that might be of use in distinguishing between potentially good and bad borrowers--the crux of the credit problem.

Distribution of Data

Standard discriminant programs assume that the discriminating variables within each group are distributed in a multivariate normal fashion. While functional forms other than multivariate normal are theoretically acceptable, the practical problems of calculating the joint probability density functions are substantial. Most discriminant programs presuppose the underlying group distribution to be multivariate normal and proceed to assign classification probabilities accordingly. There is evidence to suggest that discriminant results can be quite sensitive to significant departures from multivariate normality [13]. When such violations are thought to exist, the common procedure is to transform the variables to conform to a more normal distribution. The natural and standard log transformations are most frequently employed in this regard.

The widespread use of categorical variables in statistical credit scoring models represents a clear violation of the normality assumption. When dichotomous variables are present, the correct procedure is to split the sample accordingly and estimate separate discriminant functions. For example, one might elect to estimate separate models for direct and indirect lending arrangements.

The studentized range test was used as a rough test for normality for each of the five continuous discriminant variables (CREDIT, TOTDEBT, NCONT, JOBAGE, AND RESAGE).[3] Only TOTDEBT and NCONT in the good loan group and CREDIT in the turndown group appeared to be drawn from a normal population. Both natural and standard log transformations were made in an attempt to normalize the data. Since both transformations yielded similar results, only the natural log transformation results are presented. To summarize, the effect of the natural log transformation was to (1) normalize CREDIT, NCONT, RESAGE in the bad loan group, (2) normalize TOTDEBT, NCONT, JOBAGE, and RESAGE in the good loan group, and (3) normalize TOTDEBT, JOBAGE, and RESAGE in the turndown group. Thus, prior to the natural log transformation, only three of the possible 15 grouped variables were distributed in a manner consistent with the assumption of normality. Following the transformation ten of the 15 grouped variables appear to have been normalized.

The three-group standardized discriminant coefficients presented in Table I of the previous section indicate the important role that the credit rating variable plays in the model. Frequency distributions for the credit rating variable for each of the three groups are presented in Figure 4. As indicated earlier, the distribution of the CREDIT variable for the turndown group appears to be consistent with the assumption of a normal distribution while the distributions for CREDIT in the bad and good loan groups are not. Attempts to transform the individual distributions using both standard and natural log transformations had the effect of "normalizing" CREDIT in the bad loan group but served to "denormalize"

---

[3]A 95 percent level of confidence was employed. The studentized range is the quotient of the data range divided by its standard deviation. (See Fama for the associated probability distribution [15]).

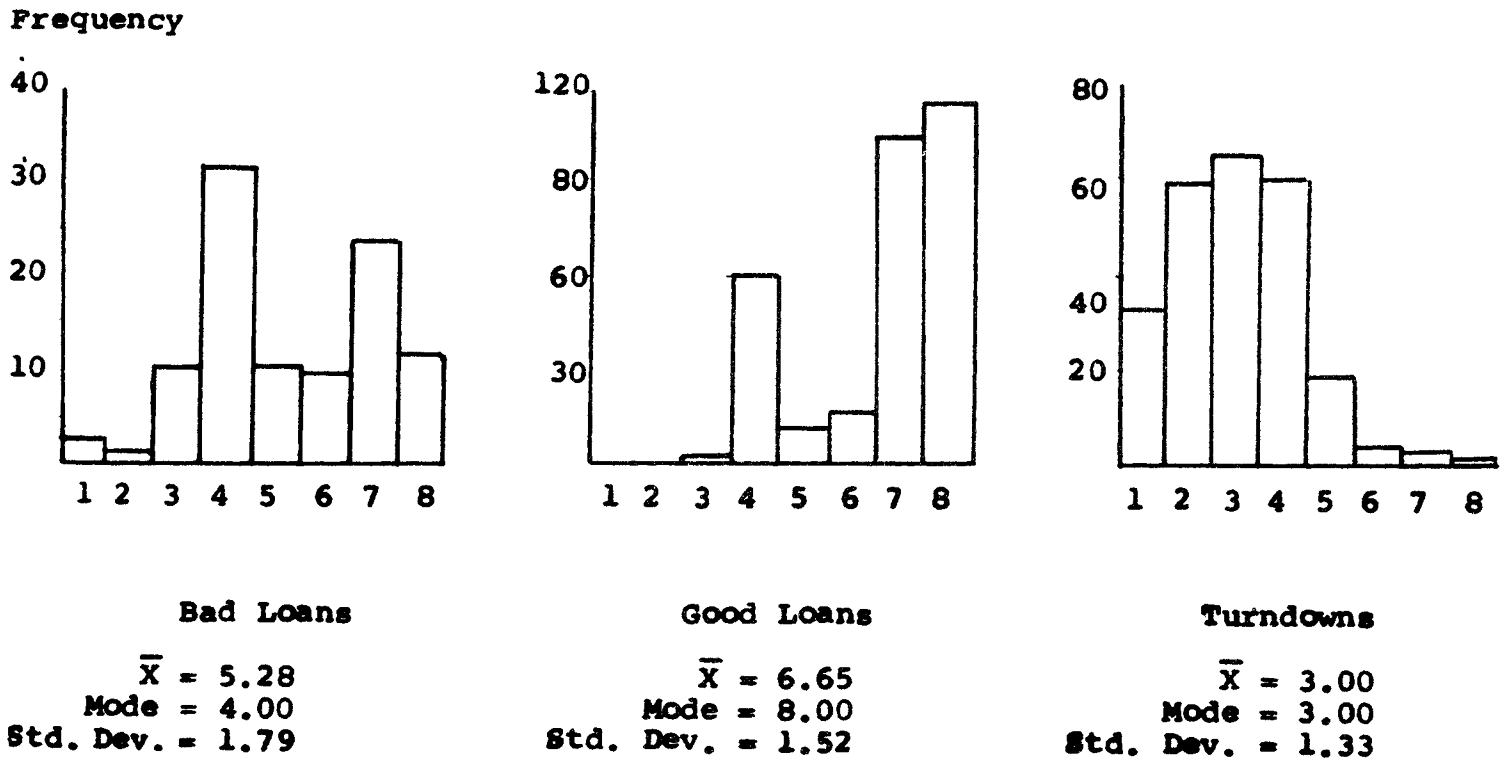

Figure 4. Distribution of Credit Variable

CREDIT for the turndown group. In either case, CREDIT for the good loan group was not normally distributed. For the remaining four variables the transformations generally had the desired effect of normalizing the distribution of each group. The results of the discriminant analysis using both transformed and untransformed data are represented in Table II.

All variables in both models carried the appropriate signs, and the sizes of their standardized discriminant coefficients relative to the respective coefficient on CREDIT were quite similar with the exception of

Table II. Log Transformation Results

A. Standardized Discriminant Coefficients[a]

Variable	Untransformed Results	Natural Log Transformation Results
CREDIT	-1.485	-1.438
JOBAGE	- .161	- .289
RESAGE	- .008	- .138
TOTDEBT	.187	.185
NCONT	.163	.151

B. Classification Results

Untransformed Results

Actual Group	#	Predicted Group: Bad	Good	Turndn
Bad	51	13 (25.5)	13 (25.5)	25 (49.0)
Good	148	23 (15.5)	100 (67.6)	25 (16.9)
Turndn	126	6 (4.8)	3 (2.4)	117 (92.9)

Overall Accuracy = 70.8%

Transformed Results

Actual Group	#	Predicted Group: Bad	Good	Turndn
Bad	51	24 (47.1)	15 (29.4)	12 (23.5)
Good	148	43 (29.1)	97 (65.5)	8 (5.4)
Turndn	126	25 (19.8)	2 (1.6)	99 (78.6)

Overall Accuracy = 67.7%

[a]Second discriminant function is not statistically significant.

[b]The classification results reflect a 50 percent holdout testing procedure and the assumption of equal group prior probabilities. Table elements represent both the number and percent of loans in the actual groups classified into each prediction group (percent in parentheses).

[c]Overall accuracy is measured down the diagonal of the classification matrix.

RESAGE, which appears to be the least significant variable in the model. The classification results presented Table II-B summarize the effect of the transformation on the model's ability to predict. The accuracy rate for the transformed model is slightly lower than that reported for the untransformed data. The transformation had virtually no impact upon the model's good loan prediction rate. On the other hand, the transformation dramatically affected the prediction rates for both bad loans and turndowns, since it classified a much higher percentage of loans as bad rather than turndown.[4] This result seems logical when one examines the distribution of CREDIT for each of the two groups (Figure 4). The natural log transformation would serve to shift both distributions to the left and condense the bimodal distribution for bad loans into a more compact normal shape. Since the classification procedure assigns group probabilities upon the basis of an assumed multivariate normal distribution, it would seem logical that the classification results would be more heavily weighted towards the now "normalized" bad loan group.

The results indicate the difficulty associated with attempting to transform credit data into a normalized distribution. One would logically expect the CREDIT distribution for good loans to be strongly skewed to the left. Attempts to normalize such a distribution would likely prove to be quite difficult and somewhat artificial. In general, the findings are consistent with previous research [21] which indicates that while standard

---

[4] A separate model was estimated transforming all variables <u>except</u> CREDIT. This hybrid model yielded virtually the same results as the untransformed model presented above. Thus, it seems clear that the shifts observed above are mainly attributable to the effect of transforming the CREDIT variable.

linear classification procedures may be quite sensitive to a lack of multivariate normality, the severity of the problem is substantially reduced when dealing with bounded distributions. Furthermore, the between group classification errors are more strongly affected than the overall classification error rate.

Dichotomous Variables

The problems encountered with nonnormal multivariate distributions are further compounded when dichotomous variables (i.e., variables which may assume only one of two possible values) are present in the discriminant model since such variables are inherently not normally distributed. While some effort has been made to develop optimal classification procedures for handling cases which consist of both continuous and dichotomous variables [7] the generally accepted procedure is to split the sample based upon individual values for the dichotomous variable and estimate distinct models for each subgroup.

To test the effect of introducing a dichotomous variable into the model, the sample was divided into two groups, one composed of 406 direct loans, while the other included 242 indirect loans. An indirect loan represents dealer paper purchased by the bank, whereas a direct loan indicates the extension of credit based upon information gathered during a formal interview with bank lending personnel. When purchasing indirect paper, the bank relies upon the credit information supplied on the loan application provided by the dealer. It seems likely that the underlying populations of direct and indirect borrowers might possess distinctive characteristics since potential borrowers somewhat skeptical of their present ability to obtain bank credit directly may be more inclined to seek dealer financing. To the extent that such a self-imposed screening process does take place, the population of direct and indirect borrowers would likely differ in terms of basic credit characteristics. Furthermore, the bank in purchasing dealer paper will likely consider the same basic credit factors they deemed to be important when evaluating direct loans--although the relative weights associated with specific variables may possibly vary.

The variable IORD was introduced into the model to signify whether or not the loan or application was an indirect or a direct credit arrangement. The standardized discriminant coefficients and the classification results are presented in Table III and compared with the previous model, which ignored this credit characteristic. Table III also reports the discriminant coefficients and classification results for the split samples of direct and indirect loans.

An examination of the standardized coefficients in Table III-A indicates that the type of lending arrangement is in fact an important consideration. The coefficient on IORD (-.235) is relatively large in comparison with the remaining coefficients and carries the expected sign. On the other hand, a comparison of the classification results in Table III-B indicates that the introduction of IORD in the model had very little impact upon the classification results. The results indicate a slightly greater tendency to classify a questionable borrower as a bad loan rather than a turndown. Hence, the introduction of this dichotomous variable had virtually no impact upon the model's overall prediction accuracy and had only a slight effect upon the between-group classification results.

Table III. Dichotomous and Continuous Variable Models

A. Standardized Discriminant Coefficients[a]

Variable	Continuous Variable Model[b] (Full Sample Coeff.)	Continuous & Dichotomous Variable Model (Full Sample Coeff.)	Split Sample: Direct Loans	Split Sample: Indirect Loans
CREDIT	-1.485	-1.421	1.259	-1.752
JOBAGE	- .161	- .188	.115	- .286
RESAGE	- .008	- .008	.004	.015[c]
TOTDEBT	.187	.187	- .351	.087
NCONT	.163	.092	.095[c]	.254
IORD	N.A.	- .235	N.A.	N.A.

B. Classification Results[d]

Continuous Variable Model

Prediction Group

Actual Group	#	Bad	Good	Turndn
Bad	51	13	13	25
		(25.5)	(25.5)	(49.0)
Good	148	23	100	25
		(15.5)	(67.6)	(16.9)
Turndn	126	6	3	117
		(4.8)	(2.4)	(92.9)

Overall Accuracy[e] = 70.8%

Continuous and Dichotomous Variable Model

Prediction Group

Actual Group	#	Bad	Good	Turndn
Bad	51	14	13	24
		(27.5)	(25.5)	(47.1)
Good	148	28	100	20
		(18.9)	(67.0)	(13.5)
Turndn	126	10	3	113
		(7.9)	(2.4)	(89.7)

Overall Accuracy = 69.8%

---

[a]Second discriminant function is not statistically significant.

[b]Model excludes dichotomous variable IORD.

[c]Incorrect sign.

[d]The classification results reflect a 50 percent holdout testing procedure and the assumption of equal group prior probabilities. Table elements represent both the number and percent of loans in the actual groups classified into each prediction group (percent in parentheses).

[e]Overall accuracy is measured down the diagonal of the classification matrix.

Table III - B continued

Direct Loan Model

Predicted Group

Actual Group	#	Bad	Good	Turndn
Bad	37	9	18	10
		(24.3)	(48.6)	(27.0)
Good	105	16	77	12
		(15.2)	(73.3)	(11.4)
Turndn	57	11	2	44
		(19.3)	(3.5)	(77.2)

Overall Accuracy = 65.3%

Indirect Loan Model

Predicted Group

Actual Group	#	Bad	Good	Turndn
Bad	9	3	3	3
		(33.3)	(33.3)	(33.3)
Good	39	4	29	6
		(10.3)	(74.4)	(15.4)
Turndn	67	4	0	63
		(6.0)	(0)	(94.0)

Overall Accuracy = 82.6%

C. Pooled Classification Results

Actual Group	#	Bad	Good	Turndn
Bad	46	12	21	13
		(26.1)	(45.7)	(28.8)
Good	144	20	106	18
		(13.9)	(73.6)	(12.5)
Turndn	124	15	2	107
		(12.1)	(1.6)	(86.3)

Overall Accuracy = 71.7%

Analysis of the standardized discriminant coefficients and the classification results for the split sample suggests another conclusion. While coefficients for two of the variables, RESAGE in the indirect loan model and NCONT in the direct loan model, had inappropriate signs, these two variables were not particularly significant in their respective models. On the other hand, the variables JOBAGE and NCONT have substantially greater influence in the indirect loan model than in the direct loan model. These differences are not surprising given the greater risk typically associated with indirect lending. For example, one would logically expect borrower character, as measured by JOBAGE, and length of contract period (NCONT) to be relatively more important in credit decisions involving dealer paper.

A comparison of classification results indicates a substantial difference in the two models' abilities to predict. The overall accuracy rate for the indirect loan model was 82.6 percent versus 65.3 percent for the direct loan model. In addition to the fact that the direct loan results were based upon a sample roughly twice the size as the indirect loan model, the holdout classification sample for the direct loan model had a significantly greater percentage of bad loans, while the indirect loan holdout model was more heavily weighted towards turndowns. Since the model has consistently been more accurate in predicting turndowns than either of the other two groups, the greater influence of the turndown group undoubtedly has affected the overall accuracy of the indirect loan model. One way to eliminate this influence is to combine the classification results achieved by the direct and indirect loan models and to compare these "pooled" prediction results with the full sample results. The pooled classification results are presented in Table III-C.

A comparison of the pooled classification results and the results of the continuous variable model in Table III-B indicates very little improvement in overall prediction accuracy. The only major difference occurs in the percentage of bad loans predicted to be good (45.7 percent for the pooled model versus 25.5 percent for the initial full-sample model). This same tendency towards good loan prediction is seen in the direct loan model classification results and undoubtedly reflects the greater percentage of bad loans in the direct loan model's holdout sample and the problems associated with identifying bad loans in general.

Thus, it appears that introducing the single dichotomous variable IORD directly into the model has had no measurable destabilizing effect upon the linear classification function's ability to predict. At the same time, given the differences observed in the abilities of the direct and indirect loan models to predict, it seems desirable to split the sample and estimate a separate function for each sample of direct and indirect loan applicants.

Classification Testing Procedures

It is generally recognized that employing the same sample used in deriving the discriminant function to test the model's overall classification accuracy typically leads to biased and overly optimistic classification results.[5] Eisenbeis evaluates ten alternative classification testing procedures and in general recommends either the "holdout" or Lachenbruch method, depending upon sample size and data availability [13].

---

[5]Referred to henceforth as the reclassification approach. In the holdout approach, the data sample is split in two with one subset being used to estimate the discriminant functions and the second or holdout subset being used to evaluate the model's predictive accuracy. The Lachenbruch method is a special case of the general holdout approach and will be discussed later.

This section examines the degree of bias inherent in the reclassification approach and the stability of the discriminant coefficients and classification results for a limited number of models. Furthermore, the impact of various holdout sample sizes upon classification accuracy is examined.

As can be seen in Table IV-A, the standardized discriminant coefficients obtained using the reclassification approach and a 50 percent random holdout procedure are quite similar. The classification results presented in Table IV-B reflect the reclassification bias previously described. Several 50 percent random holdout models were estimated. The typical model generated results that reflected an overall accuracy rate of 66 to 67 percent. Hence the reclassification bias appears to lie in the 2 to 3 percent range. While this bias does not appear to be large, the bias can readily be eliminated using appropriate holdout techniques.

In an effort to determine the stability of the standardized discriminant coefficients, twenty-four randomly selected 50 percent holdout models were estimated. Summary statistics for the absolute and relative values of these coefficients as well as the model's overall accuracy rate are presented in Table V.

Using the coefficient of variation as a measure of the relative stability of the standardized discriminant coefficients, the overriding importance of the prior credit rating variable (CREDIT) in the model is evident. The absolute mean size of the coefficients on CREDIT indicate its high degree of discriminating power relative to the other variables, and its low coefficient of variation (3.6 percent) indicates that a strong stable influence is being exerted. The next most important variable in

Table IV. Reclassification vs. Holdout Testing Procedures

A. Standardized Discriminant Coefficients

Variable	Reclassification	Holdout
CREDIT	-1.477	-1.504
JOBAGE	- .184	- .224
RESAGE	.013[b]	- .011
TOTDEBT	.126	.149
NCONT	.171	.154

B. Classification Results[c]

Reclassification Method

Predicted Group

Actual Group	#	Bad	Good	Turndn
Bad	99	31	37	31
		(31.3)	(37.4)	(31.3)
Good	306	48	226	32
		(15.7)	(73.9)	(10.5)
Turndn	243	48	4	191
		(19.8)	(1.0)	(78.6)

Overall Accuracy = 69.1%

Holdout Method[d]

Predicted Group

Actual Group	#	Bad	Good	Turndn
Bad	53	13	22	18
		(24.5)	(41.5)	(34.0)
Good	148	27	109	12
		(18.2)	(73.6)	(8.1)
Turndn	122	28	1	93
		(23.0)	(0.8)	(76.2)

Overall Accuracy = 66.6%

[a]Second discriminant function is not statistically significant.

[b]Incorrect sign.

[c]The classification results reflect a 50 percent holdout testing procedure and the assumption of equal group prior probabilities. Table elements represent both the number and percent of loans in the actual groups classified into each prediction group (percent in parentheses).

[d]Average or typical random holdout sample results.

Table V. Stability of Standardized Discriminant Coefficients[a]

Variable	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation (%)	Percent of Coefficients with Correct Sign
Actual Value						
CREDIT	-1.581	-1.392	-1.477	.053	3.6	100
NCONT	.090	.284	.169	.052	30.6	100
TOTDEBT	.023	.210	.099	.045	45.1	100
JOBAGE	- .321	- .104	- .193	.046	23.8	100
RESAGE	- .177	.162	.015	.075	498.8	42
Relative Values						
NCONT/CREDIT	- .195	- .059	- .115	.036	31.7	--
TOTDEBT/CREDIT	- .151	- .016	- .068	.032	46.7	--
JOBAGE/CREDIT	.068	.222	.131	.034	25.6	--
RESAGE/CREDIT	- .105	.115	- .011	.051	470.6	--
Overall Accuracy Rate	64.6	75.3	68.2	2.49	3.7	

[a]Based on twenty-four random 50 percent holdout samples.

terms of size and stability appears to be JOBAGE, followed by length of contract period (NCONT), total debt ratio (TOTDEBT), and RESAGE. Since JOBAGE and RESAGE are both designed to measure borrower character and stability, the strong influence of JOBAGE and the weak influence of RESAGE is not unexpected.

Eisenbeis points out that while the absolute sizes of the discriminant coefficients are not unique, the ratios between coefficients are unique up to a constant factor of proportionality. In recognition of this constant proportionality between coefficients, the coefficients for each of the variables was divided by the coefficient for CREDIT. The summary statistics for these relative coefficients are presented in the lower portion of Table V. By inspection one notices that the actual coefficients and the relative coefficients possess virtually the same degree of stability as measured by their respective coefficients of variation. Hence, it would appear that the holdout procedure adopted throughout the analysis generates reasonably consistent and reliable coefficients for those variables which possess a significant degree of discriminating power.

Of the two methods previously mentioned--holdout and Lachenbruch--the later procedure is frequently put forth as being superior when total sample size is small [19].[6] Using the Lachenbruch approach, a single observation is withheld from an estimation sample of size n and individually classified using the remaining n-1 cases to estimate the appropriate discriminant function(s). This initial observation is then

---

[6]To be precise, it is the sample size relative to the number of variables that is important, rather than the absolute size of the sample.

replaced and a second observation withheld and classified based upon the remaining cases, and so forth, until all n cases are classified one at a time. For small sample sizes, this procedure is employed to avoid splitting the sample in two and substantially reducing the size of both the estimation and classification samples. While this approach has significant advantages for relatively small samples, it is computationally more complex, and only a few standard computer programs are designed to implement this procedure.[7] On the other hand, many discriminant programs include some type of random or fixed holdout procedure.

The question as to the appropriate size of the holdout sample is important since the larger the holdout sample, the smaller the estimation sample and, possibly, the less reliable the discriminant coefficients. The usual procedure cited in the literature is to split the initial sample in half and test the classification accuracy of the model using a 50 percent random test sample. In an effort to determine the stability of the model's results for various holdout sample sizes, ten randomly selected samples were withheld and classified for each of the following holdout percentages: 10, 20, 30, 40, and 50 percent. The mean accuracy rates and average Type I and Type II error rates for each of these holdout percentages are defined and depicted in Figure 5.

As can be seen, the mean accuracy levels and error rates are remarkably stable over the indicated range of holdout percentages. It should be pointed out that the variation in observed error and accuracy rates for the 10 percent holdout group was substantially larger than that

[7]Furthermore, as sample size increases, the Lachenbruch and the reclassification procedures yield results which are similar for all practical purposes.

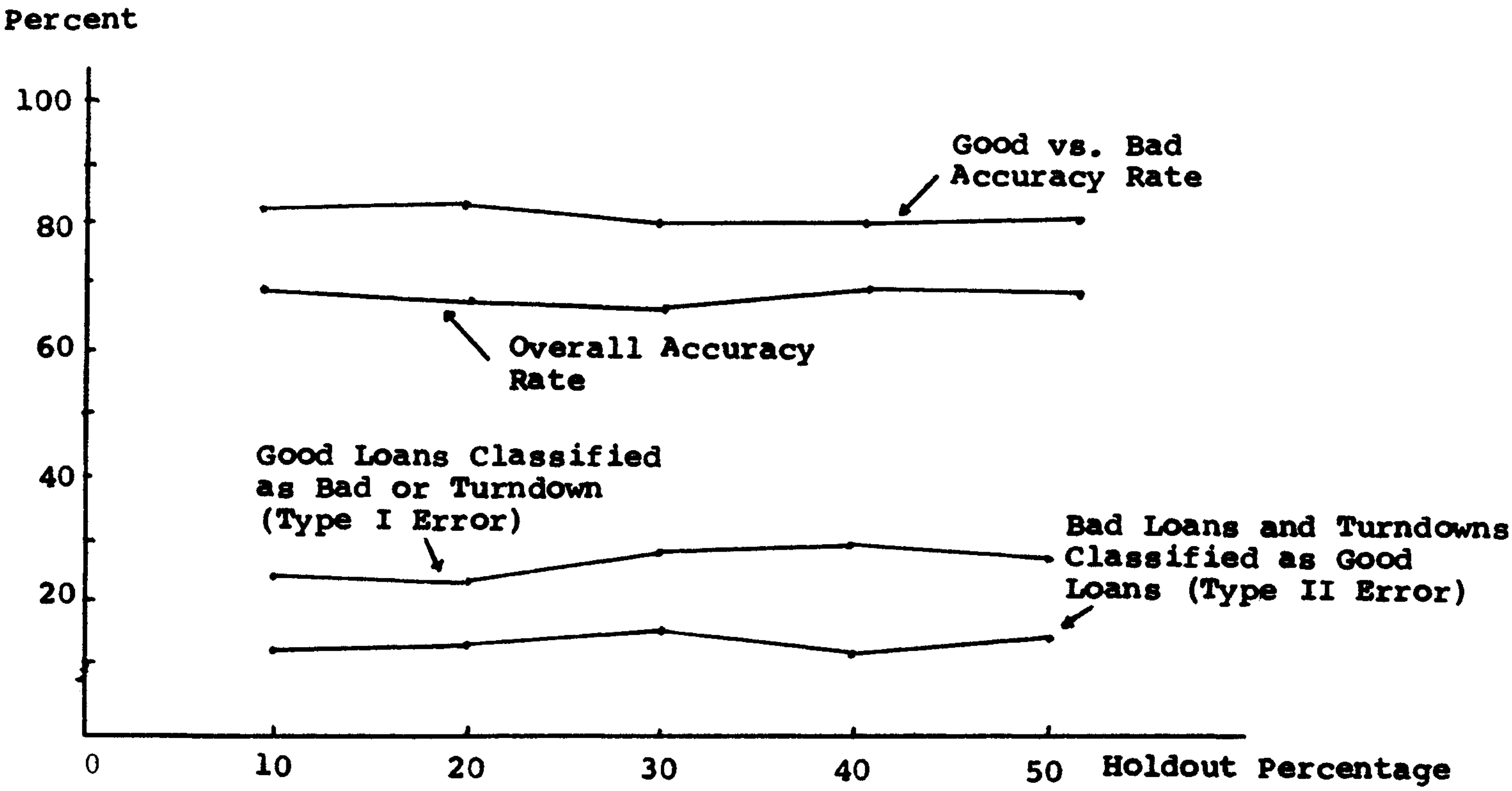

Figure 5. Accuracy Levels and Specific Error Rates for Various Holdout Percentages.

obtained for the larger holdout percentages. Hence, while one should avoid using extremely small test samples, reliable results can be achieved using relatively small holdout samples in the range of 20 to 30 percent when necessary. Thus, unless the initial sample size is extremely small, the standard holdout procedure can prove to be a suitable testing procedure.

Use of A Priori Probabilities

Discriminant procedures require the user to specify, either explicitly or implicitly, both the a priori probabilities of group membership in the general population and the relative cost of misclassification. While recognizing that Type I errors (rejecting a good loan) and Type II errors (accepting a bad loan) will likely involve different types and levels of cost to the lender, this section concentrates upon the estimation and appropriate use of prior group probabilities. As Eisenbeis [12] points out, use of the appropriate a priori (or prior) probabilities is necessary to insure minimization of the misclassification error rate when the model is applied to the true population of potential borrowers. Use of inappropriate priors will not minimize classification errors and may generate overly optimistic classification results. In the past, researchers have frequently assumed equal group priors or have employed sample group proportions as the relevant prior probabilities. In either case misclassification errors will not be minimized unless the groups are found to appear with equal frequency in the general population or unless the data represent a truly random sample. Both of these assumptions seem unreasonable given the general nature of the lending/borrowing relationship and the data limitations inherent in most credit scoring models. In light of these issues, it seems more appropriate to actually calculate the a priori group probabilities and then weight the classification function accordingly.

In this study, prior group probabilities were estimated by sampling 1001 loans closed during the first three months of 1978, yielding proportions of 80 percent for good loans and 20 percent for bad loans. To obtain a prior probability for turndowns, 3160 loan applications submitted during June and July 1977 were examined. Of this number 483 applications were rejected yielding a turndown prior of 15 percent. The proportions of good and bad loans determined above were then applied to the percentage of granted loans (85 percent of all applicants) yielding the prior probabilities for each group as follows:

Group				Prior Probability
Good loans	=	(0.80 x 0.85)	=	0.68
Bad loans	=	(0.20 x 0.85)	=	0.17
Turndowns	=			0.15
				1.00

Table VI indicates three alternative sets of prior probability assumptions under examination. Compared to the estimated population prior probabilities, the sample is overrepresented by turndowns and underrepresented by good loans.

Table VI. Alternative Prior Probabilities

Group	Equal Priors	Sample Priors[a]	Population Priors
Bad Loans	.33	.16	.17
Good Loans	.33	.46	.68
Turndowns	.33	.38	.15

[a]Based upon actual group proportions in the estimation sample.

Since prior probabilities are employed as weights in the classification function, the standardized discriminant coefficients for each set of assumptions are identical to those presented in Table I. When discussing the most appropriate classification approach, Anderson [1] states that, "In constructing a procedure of classification, it is desired to minimize the probability of misclassification, or more specifically, it is desired to minimize on the average the bad effects of misclassification." Concentrating on the role of prior probabilities Eisenbeis [14] develops the following classification function for the two-group case which minimizes the expected probability of misclassification (M).

$$M = P(1|2)\pi_2 + P(2|1)\pi_1$$

where $P(1|2)$ is the probability of assigning an observation to group one when it belongs to group two and $P(2|1)$ is the probability of assigning an observation to group two when it belongs to group one. π_1 and π_2 are the proportions of the population associated with groups and 1 and 2, respectively. This objective function is then minimized when the following classification rule is employed.

Classify X into group 1 when

$$\frac{f_1(X)}{f_2(X)} \geq \frac{\pi_2}{\pi_1}$$

In the expression above, the functions f_1 and f_2 represent the multivariate normal probability density functions for groups 1 and 2, respectively. In the case of equal priors ($\pi_1 = \pi_2$), any given observation x will be assigned to group 1 as long as the probability of group 1 membership is at least equal to the probability of group 2 membership. On the other hand, if π_2 exceeds π_1 by a substantial margin, say, three times as large, then its probability of group 1 membership must be at least three times as large as its probability of group 2 membership in order for the observation to be classified into group 1. Thus, the greater the population prior associated with one group relative to another, the more "difficult" it is to classify a given observation into the group with the smaller prior probability. When estimating the discriminant model using sample priors, the appropriate sample group proportions are substituted for π_1 and π_2 in the expression above. For example, using the priors given in Table VII, the sample priors associated with good loans and turndowns exceed the bad loan prior by a factor between two and three. Looking at the estimated population priors, the good loan probability exceeds both the bad loan and turndown probabilities by a factor of four to one. Hence, in comparison to the equal priors model, the sample priors model will tend to "favor" good loans and turndowns at the expense of bad loans, while the population priors model will tend to "favor" good loan predictions relative to both bad loans and turndowns.

The results presented in Table VII document these shifts. Comparing the sample prior results to the outcome of the equal prior model, one sees an increase from 36 percent to 43 percent in the percentage of total loans predicted to be good and a similar increase from 51 percent to 57 percent in the proportion of loans predicted to be turned down. The difficulty in obtaining a bad loan prediction under these conditions is dramatically illustrated here since the bad loan prediction percent declines from 13 percent to zero. In analyzing the results of the population priors model one clearly sees the effect of the dominant good loan prior probability as the good loan prediction rate rises to 60 percent at the expense of a decline in the turndown rate. Once again no loans are predicted to be bad.

In terms of predictive accuracy, the use of population priors appears to have minimized the number of misclassifications. This shift in prior probabilities had increased the frequency of Type II errors (i.e., bad loans and turndowns predicted to be good) from 9 percent to 33 percent and has reduced the number of Type I errors (i.e., good loans classified as bad loans or turndowns) from 32 percent to 7 percent.[8] Whether this represents a favorable trade-off or not depends, of course, upon the relative cost of misclassification.

The population priors model, more accurately reflecting the distribution of potential borrowers in the general population, does an excellent job in identifying good borrowers, a reasonably good job in predicting turndowns, and has little or no success in identifying potentially bad borrowers. This, of course, parallels actual loan officer experience

[8]Specific calculations are:
(3 + 13)/177 = .09; (25 + 33)/177 = .33; (23 + 25)/148 = .32; and 11/148 = .07

Table VII. Assessment of Equal, Sample, and Population Prior Probabilities

Classification Results[a]

Equal Priors

Prediction Groups

Actual Group	#	Bad	Good	Turndn
Bad	51	13	13	25
		(25.5)	(25.5)	(49.0)
Good	148	23	100	25
		(15.5)	(67.6)	(16.9)
Turndn	126	6	3	117
		(4.8)	(2.4)	(92.9)

Overall Accuracy = 70.8%[b]

Sample Priors

Prediction Groups

Actual Group	#	Bad	Good	Turndn
Bad	51	0	21	30
		(0)	(41.2)	(58.8)
Good	148	0	116	32
		(0)	(78.4)	(21.6)
Turndn	126	0	4	122
		(0)	(3.2)	(96.8)

Overall Accuracy = 73.2%

Population Priors

Prediction Groups

Actual Group	#	Bad	Good	Turndn
Bad	51	0	33	18
		(0)	(64.7)	(35.3)
Good	148	0	137	11
		(0)	(92.6)	(67.4)
Turndn	126	0	25	101
		(0)	(19.8)	(80.2)

Overall Accuracy = 73.2%

[a]The classification results reflect a 50 percent holdout testing procedure and the assumption of equal group prior probabilities. Table elements represent both the number and percent of loans in the actual groups classified into each prediction group (percent in parentheses).

[b]Overall accuracy is measured down the diagonal of the classification matrix.

since both the model and the loan officer are working with the same source of potential borrowers and, at least intuitively, similar group probabilities. The model which incorporates equal group priors assumes that a relatively large percentage of the population will be bad borrowers and becomes highly selective in its lending decisions. When applied to the true population of potential borrowers of which bad borrowers constitute a relatively small percentage, the equal priors model's "conservatism" becomes a liability as the model misclassifies a large portion of potentially good customers.

It is important to note that while it predicted no bad loans, the population priors model did predict one-third of the bad loans to be turndowns. The loan officer, of course, failed to predict any bad loans as either bad or turndown since credit was extended in every case. At the same time, approximately 20 percent of the turndown group was classified as good borrowers using the population priors model. Furthermore, since no information is available regarding the true credit worthiness of the turndown group, it seems conceivable that a certain number of those refused credit would ultimately prove to be acceptable credit risks and hence should have been classified into the good loan group initially.

Which set of prior probabilities is more appropriate depends upon the design objectives of the model. Eisenbeis [14] points out that the assumption of equal priors is valid when the model is intended to be used for "descriptive classification" purposes where the basic objective is to examine the relative position and shapes of the various group distributions. But when the model is to be used for "predictive classification" purposes, as would be the case in most credit scoring applications, the actual population probabilities are required. To the extent that the relative costs of misclassification (Type I and Type II errors) are significantly different, the proper procedure is not to adjust the population priors accordingly but instead to weight the classification function according to the relative misclassification costs.[9] The result would be to minimize the total cost of misclassification rather than the total number of misclassifications per se.

Equal Versus Unequal Group Dispersion

An important requirement of the linear discriminant model is equality between the variance-covariance matrices across all groups [13]. If these dispersion matrices are unequal, a quadratic classification rule should be substituted for the pooled linear classification function.[10] In general, several factors explain observed differences in prediction results when comparing linear and the quadratic classification functions. These

---

[9] The new two-group objective function (L) becomes: $L = C(1|2)P(1|2)\pi_2 + C(2|1)P(2|1)\pi_1$ and the associated classification rule becomes:

$$\frac{f_1(X)}{f_2(X)} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}$$

where $C(1|2)$ represents the cost of classifying a group two observation in group one and $C(2|1)$ represents the cost of classifying a group one observation into group two.

[10] See Eisenbeis and Avery [14] for the precise form of the quadratic classification rule.

differences are directly related to the degree of inequality among the dispersion matrices and the distance between groups. Classification differences increase as the number of variables and the uniqueness of their dispersion matrices increase. Furthermore, the classification differences tend to diminish as the distance between the groups increases.

For example, Figure 6 illustrates a situation where the dispersion matrices are equal and the distance between groups is small. Figure 7 depicts a situation where significant differences in the dispersion matrices exist and the distance between the two groups is substantial. In Figure 6 equality between group dispersions mitigates the potentially disruptive influence of relatively small group separation, whereas, in Figure 7 the disruptive influence of very different dispersion matrices is reduced by the wide separation between groups. In both cases use of the linear or quadratic classification functions would likely yield similar results. On the other hand, in situations where the differences in dispersion matrices are substantial and group separation is minimal, the two classification rules will frequently generate significantly different results, especially when relatively few variables are employed. In this case (see Figure 8), one would want to adopt the quadratic classification rule since the quadratic function takes advantage of the additional information inherent in the two heavily overlapped distributions.

Table VIII-A presents the results of tests for equality of group means for the two- and three-group models.[11] In each model the null

---

[11] It should be noted that sample size and the level of significance are related in any such measure of group separation. Thus, the tests were conducted using the appropriate degrees of freedom which takes into consideration both sample size and the number of variables in the model.

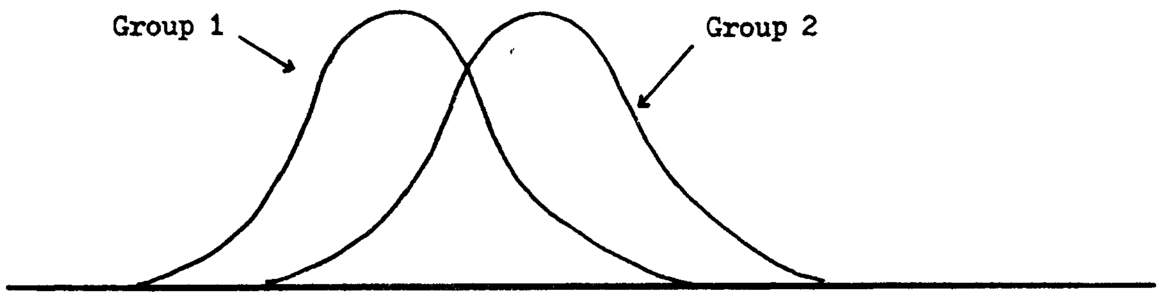

Figure 6. Equal Dispersion Matrices for Closely Grouped Distributions

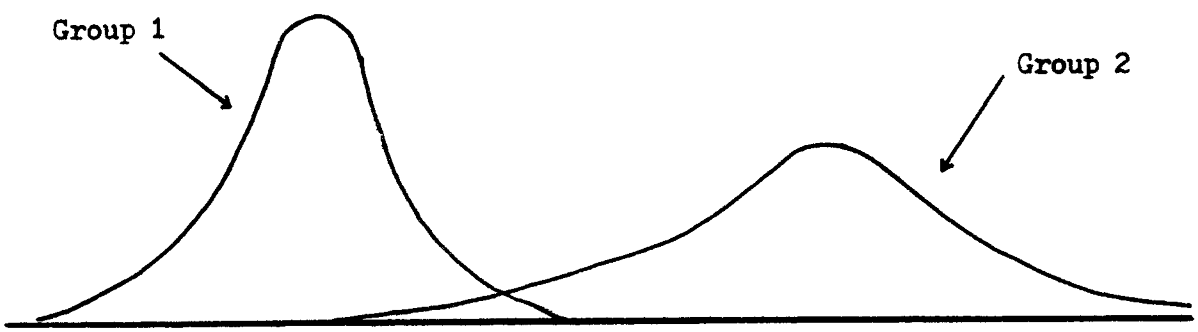

Figure 7. Unequal Dispersion Matrices for Widely Separated Distributions

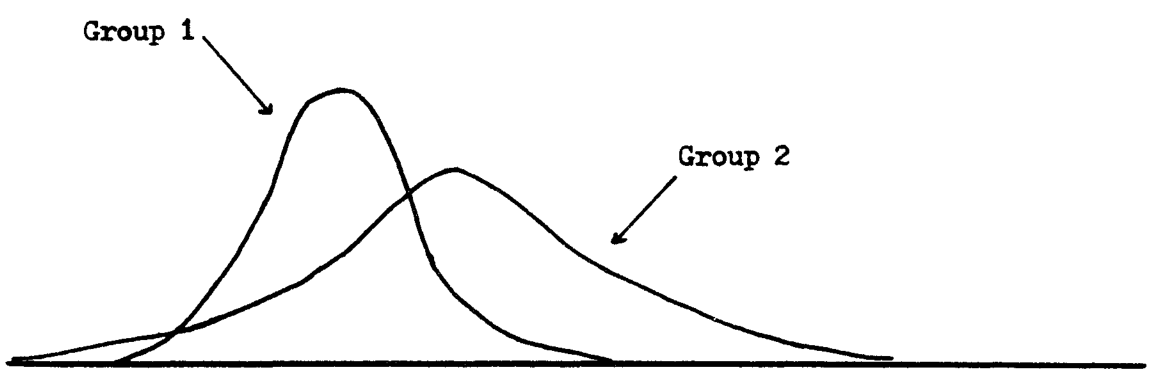

Figure 8. Unequal Dispersion Matrices for Closely Grouped Distributions

Table VIII. Tests for Equality of Group Means and Covariance Matrices

A. F-Test for Equality of Group Means[a]

Two Group (Good vs. Bad)

Group:	Bad
Good	3.94 (.002)

Degrees of Freedom: 5/205

Three Group (Good, Bad, Turndown)

Group:	Bad	Good
Good	4.47 (.001)	-
Turndown	20.49 (.0001)	81.85 (.000)

Degrees of Freedom: 5/316

B. Results of Test for Equality of Group Covariance Matrices

Model	Box's M	Approx. F	Degrees of Freedom (Numerator/ Denominator)	Level of Statistical Significance
Test of Covariance Matrices				
Linear Two-Group Model (Good vs. Bad)	14.16	.905	15/30638	.558
Linear Three-Group Model (Good vs. Bad vs. Turndown)	58.43	1.89	30/73781	.003

[a]The F-statistic and the associated degrees of freedom appear in the body of the table along with the level of statistical significance in parentheses immediately below. A careful examination of the distribution of discriminant scores in Figures 9 and 10 plus the pairwise results presented above strongly suggests that the null hypothesis be rejected. While a more appropriate approach would be to simultaneously test for equality between the three-group means a direct test of all three-group means was not readily available.

hypothesis of equal group means is rejected for all meaningful levels of statistical significance. In the three-group model maximum group separation is obtained when comparing the good and turndown groups. These differences in group separation are graphically illustrated in Figures 9 and 10.

Table VIII-B presents the results of tests for equality of the group dispersion matrices and the quadratic discriminant functions for both the two-group and the three-group models. The Box's M statistic and the F-distribution are employed to test for statistically significant differences [4]. Using a 95 percent confidence level, the null hypothesis of equal dispersion matrices is accepted for the two-group, bad versus good loan model. On the other hand, the assumption of equality is clearly rejected in the case of the three-group model. Thus, the two-group, bad versus good loan model is analogous to the situation depicted in Figure 6 where equal dispersion matrices are present and group separation is modest. The three-group model behaves in a fashion comparable to the one illustrated in Figure 7 where significant differences exist between the dispersion matrices but the groups appear to be widely separated. In both cases use of either the linear or quadratic classification rules should yield quite similar results.

Table IX presents the classification results for both the linear and quadratic rules for both models. The classification results were in fact identical for the two-group model. In the three-group model, the classification results are similar, but a greater tendency towards bad loan prediction is noted using the quadratic model. Three bad loans classified as good under the linear rule were correctly classified as bad using the quadratic model, while ten good loans originally predicted to

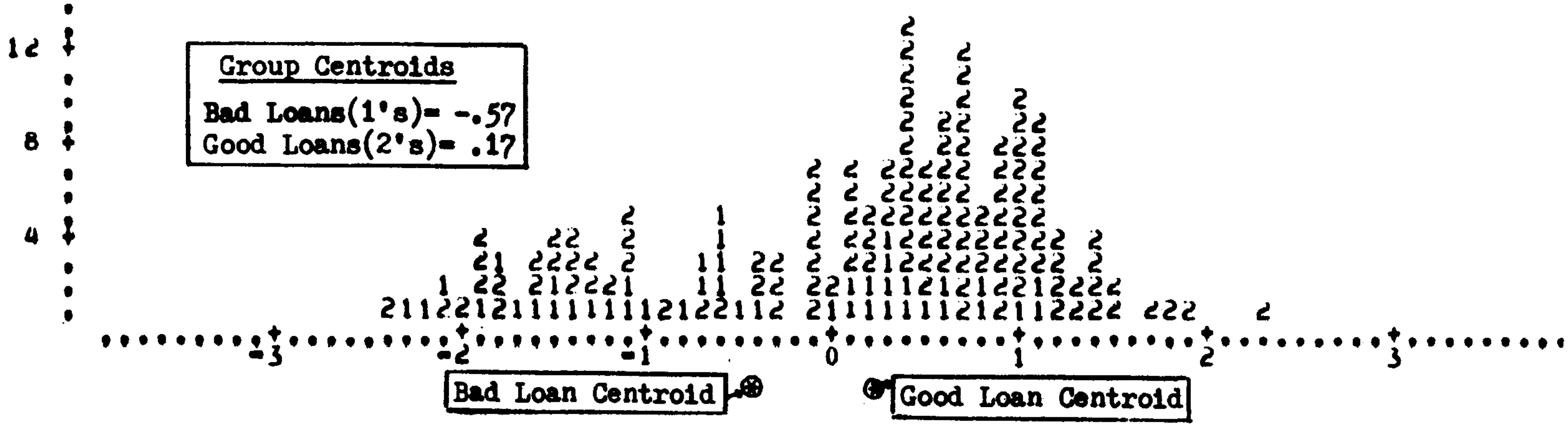

Figure 9. Distribution of Classification Scores for the Two-Group, Good Versus Bad Model.

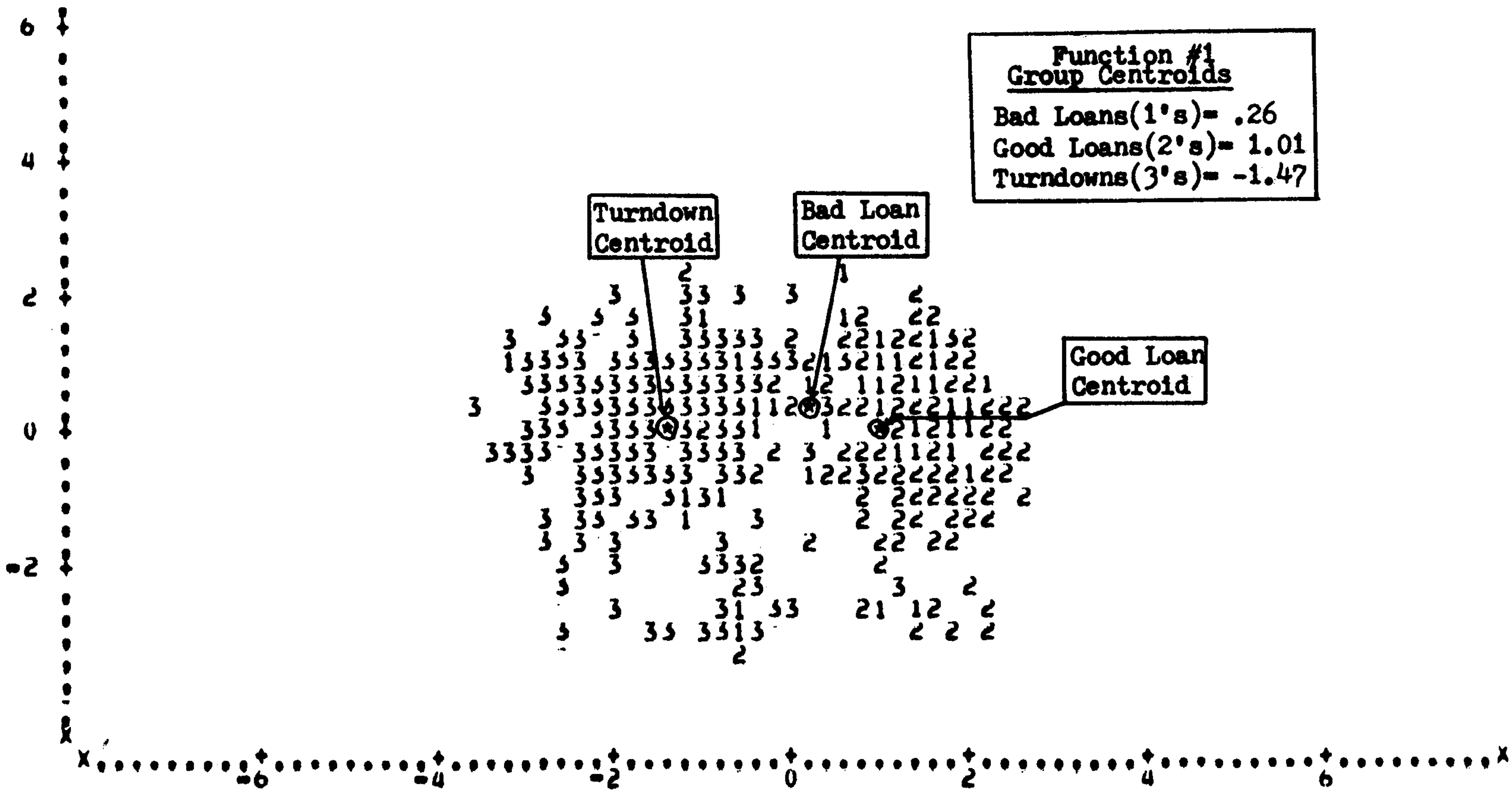

Figure 10. Distribution of Classification Scores for the Three-Group, Good, Bad, and Turndown Model.

Table IX. Linear and Quadratic Classification Results

Two Group (Good vs. Bad)
Linear and Quadratic
Classification Results

Prediction Group

Actual Group	#	Bad	Good
Bad	48	29	19
		(60.4)	(39.6)
Good	158	41	117
		(25.9)	(74.1)

Overall Accuracy Rate = 70.9%

Three Group (Good, Bad, Turndown)
Linear Classification Results

Prediction Group

Actual Group	#	Bad	Good	Turndn
Bad	48	13	19	16
		(27.1)	(39.6)	(33.3)
Good	158	14	118	26
		(8.9)	(74.7)	(16.5)
Turndn	117	14	2	101
		(12.0)	(1.7)	(86.3)

Overall Accuracy Rate = 71.8%

Three Group (Good, Bad, Turndown)
Quadratic Classification Results

Prediction Group

Actual Group	#	Bad	Good	Turndown
Bad	48	17	16	15
		(35.4)	(33.3)	(31.3)
Good	158	26	108	24
		(16.5)	(68.4)	(15.2)
Turndn	117	11	3	103
		(9.4)	(2.6)	(88.0)

Overall Accuracy Rate = 70.6%

be good were incorrectly reclassified as bad using the same quadratic rule. A similar analysis was conducted using the natural log transformed data which appeared to be more normally distributed. The results once again seem to be independent of the classification rule selected.

Section V

Summary and Conclusion

The rapid growth in consumer credit over the past 20 years has created new lending opportunities for an increasingly wide range of financial and retail organizations. Dramatic increases in the rate of inflation plus greater volatility in interest rates and business cycles have reduced earnings, encouraging firms to develop new credit management techniques. Concurrently, public policy has become increasingly sensitive to possible inequities inherent in the lending process. These concerns have prompted the Congress to enact major new legislation in the consumer lending field. As a result, the use of statistically based credit scoring models has grown dramatically over the past decade as commercial banks, consumer finance companies, and large national retailers have developed a variety of analytical techniques designed to improve productivity and efficiency.

The most common statistical credit scoring models employ multiple discriminant analysis (MDA) techniques to identify creditworthy borrowers. But MDA is not without its limitations and a careful assessment of the conceptual issues associated with the development of MDA credit scoring models is in order. The objective of this study has been to examine several of the model's most important statistical assumptions in the context of a realistic lending situation.

A word of caution seems appropriate. While the model employed in this paper is typical of the types of credit scoring models commonly adopted and was estimated using actual consumer credit information, the reader is cautioned against over-generalizing the results for a variety of reasons. First of all, the model is estimated from data collected at a single point in time and no attempt has been made to determine the stability of the findings over time. Second, several variables frequently employed in credit scoring systems were not incorporated into the model because of data limitations. Third, in actual practice credit scoring models are normally estimated using data samples substantially larger than the sample employed here. Thus, while the model is squarely based on sound credit theory the results should be viewed primarily as suggestive in nature.

The general conclusion suggests that it is possible to develop a model which satisfies at least the majority of the theoretical assumptions that underlie the MDA model. At the same time, the findings indicate that the model's ability to predict is relatively independent of a number of the assumptions examined. Furthermore, in several cases the assumptions themselves appear to introduce certain trade-offs and inconsistencies when applied in a realistic decision-making environment. Specifically, the evidence presented in this paper suggests the following conclusions:

(1) The inclusion of the group of rejected applicants appears to have added little information that is useful in classifying marginal credit risks. At the same time, the classification results appear to be somewhat sensitive to the manner in which this turndown group is included. In further analysis not reported in this study, the turndown group was combined with the sample of bad loans and a two-group model was estimated

which generated marginally better classification results.[12]

(2) The fact that a significant portion of most credit information is not normally distributed may not be a critical limitation since it appears to have little impact upon the model's classification results. Efforts to transform the data in a fashion more consistent with the assumption of normality will likely meet with only partial success and may provide little or no improvement in classification results. The inclusion of a limited number of dichotomous variables in the model doesn't appear to have a disruptive or destabilizing effect on the prediction results; although when a dichotomous variable is determined to be statistically important, the most appropriate procedure is to estimate a separate discriminant model for each distinct group.

(3) Regarding appropriate classification procedures, the reclassification approach that employs the initial estimation sample as the test sample provides prediction results that may involve a relatively small degree of bias. If sample size is sufficiently large and computational procedures permit, this bias can be eliminated by employing a holdout testing method based upon holdout samples as small as 20-30 percent of the total sample size. On the other hand, if sample size is a constraint and more sophisticated procedures are not available, the reclassification method appears to be a realistic alternative.

---

[12]For example, the proportion of good loans incorrectly classified in the two group bad versus good loan model presented in Table I was 30.4 percent. The comparable misclassification percentage for a modified two group model where bad loans are combined with turndowns was reduced to 25.0 percent. While it is appealing to combine bad loans and turndowns into a single group since management presumably would treat an applicant classified into either category in an identical fashion, the legal and statistical grounds for combining them are open to dispute. Given these problems and the relatively small improvement in predictive accuracy, the modified two group model was rejected in favor of the three group model.

(4) The use of population priors is necessary to minimize the total number of misclassifications and is conceptually more appropriate when employed in conjunction with relative misclassification costs. But in many realistic management applications where the identification of potentially bad borrowers receives top priority and estimates of misclassification costs are not readily available, the use of equal prior probabilities in the classification function may provide useful results.

(5) The basic nature of credit related data is such that the stability of the classification results appears to be insensitive to the form of the classification rule adopted. That is, in many lending situations the linear and quadratic forms for the classification function may generate quite similar classification results.

While it is not uncommon in empirical research for the theoretical requirements of a statistical model to be violated, the importance of the credit granting decision and the strict requirements of the MDA model make any major violation a cause for concern. The general findings that the model's results are relatively insensitive to a number of important assumptions plus the model's inability to adequately forecast marginal credit risks suggest that statistical models other than MDA might possibly be more appropriate. For example, Winginton [22] feels that the general nature of credit data is more consistent with using maximum likelihood techniques for estimating the parameters of a logit probability function. The author conducts a direct comparison of linear discriminant and logit consumer credit scoring models using a common data base and found the linear discriminant function "to be no better than chance in correctly classifying individuals. The logit function performed considerably better than chance, but nevertheless does not appear to make a significantly high proportion of correct classifications to warrant use for unaided decision-making."

Other techniques that might be explored include regression analysis, factor analysis, and multidimensional contingency analysis. It seems likely that each of these techniques, employed singly or in combination, can be useful in understanding the basics of the credit granting process, while human judgement and past experience are necessary for a more complete analysis. In effect this is what Winginton concludes when he states "While it is tempting to conclude that we might have done better with 'better' data, it seems more plausible that models of this kind, by their nature, tend to use data which are not very suitable to capturing the process whereby people decide to pay or not to pay their credit accounts, and thus (these models) are not particularly suited to aiding decisions regarding granting of credit."

This is not to say that numerical credit scoring techniques are not useful, but it does illustrate the importance of a careful assessment of the relative strengths and weaknesses of such systems. For example, there is little doubt that credit scoring systems can be developed that will reduce the number of bad borrowers to any predetermined level. For certain applications, such as the evaluation of retail credit card requests where the issuer typically has no personal contact with the applicant and processes a large volume of applications, numerical scoring systems provide an objective and efficient method of evaluation. In such systems credit cards are typically issued only to those applicants with a relatively high degree of creditworthiness. On the other hand, the essence of credit evaluation is the ability to effectively sort out creditworthy individuals from the group of marginal applicants. It is here where numerical, as well as judgemental systems, encounter their greatest difficulty.

Various approaches for implementing numerical credit scoring models exist. One approach would be as an "initial screening" device designed to identify clearly good and bad credit risks. Marginal credit applicants would then be separated for additional evaluation and analysis. This approach would allow highly trained credit personnel to concentrate on those applicants for whom their credit expertise is likely to yield the greatest return. A second procedure would be to employ the credit scoring model as a "second check" device providing the credit analyst with an "independent" evaluation of each loan request under review. Discrepancies between the analyst and the model could then be identified for further consideration. This second approach has considerable appeal in situations where new credit personnel are being trained since the model provides a tangible, quantitative assessment of the various factors that presumably define creditworthiness.

In conclusion, it is all too easy to exaggerate the predictive capabilities of credit scoring models. Their real benefit may relate not to any superiority in predictive power but to the highly consistent, objective, and efficient manner in which such predictions are made. On the other hand, a potential danger lies in the possible failure of management to clearly recognize the limitations inherent in the forecasting process and the problems associated with adopting an overly simplistic or mechanistic approach.

Bibliography

1. Anderson, T. W. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, 1958.

2. Biborosch, Rudolph. "Numerical Credit Scoring," The Credit World, June (Part 1), July (Part 2), 1965.

3. _______. "Credit Scoring Systems Have Built-In Bonuses," Bankers Monthly Magazine, March 15, 1967.

4. Box, G. E. P. "Problems in the Analyses of Growth and Linear Curves," Biometrics, Volume 6, 1950.

5. Brandez, Roland. "New Dangers in Credit Scoring--But You Can't Afford to be Without It," Banking, Journal of the American Banking Association, March, 1976.

6. Chandler, Gary and John Coffman. "A Comparative Analysis of Empirical vs. Judgmental Credit Evaluation," Working Paper, Georgia State University, 1979.

7. Chang, P. C. and A. A. Afifi. "Classification Based on Dichotomous and Continuous Variables," Journal of the American Statistical Association, June, 1974.

8. Churchill, Gilbert. "The Role of Credit Scoring in the Loan Decision," The Credit World, March 1977.

9. _______. "Developing a Credit Scoring System," The Credit World, April, 1977.

10. Cloakley, William. "Points for Small Loan Credit Scoring," Burroughs Clearing House, December, 1970.

11. Durand, David. "Risk Elements in Consumer Installment Financing," (Study #8) National Bureau of Economic Research, New York, 1941.

12. Eisenbeis, Robert and Edward Altman. "Financial Applications of Discriminant Analysis: A Clarification," Journal of Financial and Quantitative Analysis, March, 1978.

13. Eisenbeis, Robert. "Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics," Journal of Finance, June, 1977.

14. Eisenbeis, Robert and Robert Avery. Discriminant Analysis and Classification Procedures: Theory and Applications. Heath Company, publisher, 1972.

15. Fama, Eugene F. Foundations of Finance: Portfolio Decisions and Security Prices. Basic Books, 1976.

16. Hsia, David. "Credit Scoring and the Equal Credit Opportunity Act." The Hastings Law Journal, November, 1978.

17. Joy, Maurice and John Tollefson. "Some Clarifying Comments on Discriminant Analysis," Journal of Financial and Quantitative Analysis, March, 1978.

18. _______. "On the Financial Application of Discriminant Analysis," Journal of Financial and Quantitative Analysis, December 1975.

19. Lachenbruch, Peter A. Discriminant Analysis. Hafner Press, 1975.

20. Orgler, Yair. Analytical Methods in Loan Evaluation. Lexington Books, D. C. Heath and Company, 1975.

21. Scott, Elton. "On the Financial Application of Discriminant Analysis." Journal of Financial and Quantitative Analysis, March, 1978.

22. Winginton, John. "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior," Journal of Financial and Quantitative Analysis, September, 1980.

PUBLISHED STAFF MEMORANDA

Staff Memoranda -- A series of occasional papers in draft form prepared by members of the Research Department for review and comment.

75-1	Thomas A. Gittings	"Autoregressive Form of Time-Polynomial Extrapolation," 1975
75-2	Robert D. Laurent	"Reserve Requirements--Are They Lagged in the Wrong Direction?" 1975
75-3	Thomas A. Gittings	"Lissajous Figures and Cyclical Fluctuations in Unemployment and Inflation," 1975
76-1	Thomas A. Gittings	"Money, Prices and Interest Rates in Stable Monetary Growth Models," 1976
76-2	Chayim Herzig-Marx	"Advertising for Demand Deposits," 1976
76-3	Chayim Herzig-Marx	"Long-Run Deposit Variability," 1976
76-4	Chayim Herzig-Marx	"Financial Disclosure and Market Evaluations of Bank Debt Securities," 1976
77-1	Thomas J. Sargent	"The Persistence of Aggregate Employment and the Neutrality of Money," 1977
78-1	Anne S. Weaver Chayim Herzig-Marx	"A Comparative Study of the Effect of Leverage on Risk Premiums for Debt Issues of Banks and Bank Holding Companies," 1978
78-2	David R. Allardice	"Convenience and Needs Considerations: A Post-Audit Survey," 1978
78-3	Dale S. Drum	"Relative Costs of Affiliated Versus Independent Banks," 1978
79-1	Chayim Herzig-Marx Anne S. Weaver	"Bank Soundness and the Market for Large Negotiable Certificates of Deposit," 1979
79-2	Dale S. Drum	"The Effect of Holding Company Affiliation Upon the Scale Economies of Banks," 1979
79-3	Chayim Herzig-Marx	"Large Bank Failures and Investor Perceptions of the Riskiness of Banking," 1979

PUBLISHED STAFF MEMORANDA CON'T.

79-4	Chayim Herzig-Marx	"Comparing Market and Regulatory Assessments of Bank Soundness," 1979
79-5	Chayim Herzig-Marx	"Modeling the Market for Bank Debt Capital," 1979
79-6	Thomas A. Gittings	"A Linear Model of the Long-Run Neutrality of Money," 1979
79-7	Robert D. Laurent	"Reserve Requirements, Deposit Insurance and Monetary Control," 1979
81-1	Robert D. Laurent	"A Comprehensive Proposal for Accurate Monetary Control through Reserves" (Available 2/82)
81-2	George G. Kaufman	"Impact of Deregulation on the Mortgage Market," 1981
81-3	Alan K. Reichert	"An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models in the Consumer Lending Field," 1981
81-4	Robert D. Laurent	"A Critique of the Federal Reserve's New Operating Procedure," 1981
81-5	George G. Kaufman	"Banking as a Line of Commerce: The Changing Competitive Environment," 1981