CR-Sent
8-29-88

DEVELOPING A MODEL FOR PREDICTING PROBATION OUTCOME

111433

:e of Management and Information Systems
Department of Management and Budget
State of Michigan

111433

DEVELOPING A MODEL FOR

PREDICTING PROBATION OUTCOME

A REPORT TO THE
DEPARTMENT OF CORRECTIONS

Prepared by:

Department of Management and Budget
Office of Management and Information Systems
Gerald W. Williams, Director

September 1986

TABLE OF CONTENTS

FOREWORD

This development of a statistical model for predicting the probable success or failure of a convicted felon sentenced to a period of probation was conducted by the Office of Management and Information Systems, Department of Management and Budget (OMIS/DMB), at the request of the Department of Corrections.

The Office of Management and Information Systems project team for the study included:

Ram Narasimhan, Ph.D., Principal Consultant; Chairman, Department of Management, Graduate School of Business Administration, Michigan State University

Robert Landeros, Research Assistant, Graduate School of Business Administration, Michigan State University

Young Ahn, Research Assistant, Graduate School of Business Administration, Michigan State University

Walter Riley, Research Assistant, Graduate School of Business Administration, Michigan State University

Richard J. Liles, Project Manager; Senior Consultant, Human Services Group, OMIS/DMB

Bruce G. Wiley, Director, Human Services Group, OMIS/DMB

The successful completion of this research was also attributable to the contributions and assistance provided by the Department of Corrections personnel, especially:

Perry Johnson, Deputy Director, Bureau of Field Services

William Kime, Deputy Director, Bureau of Programs

Terry Murphy, Acting Chief of Research, Bureau of Programs

Carol Kierpiec, Assistant Deputy Director, Bureau of Field Services

EXECUTIVE SUMMARY

This summary describes the Probation Risk Prediction Model Development Project undertaken jointly by the Office of Management and Information Systems (OMIS), and the Michigan Department of Corrections (DOC). The purpose of the research study was to develop a prediction and classification instrument that can be used to classify probation cases into "high," "medium," and "low" risk groups. This study was conducted pursuant to the DOC's continuing effort to further improve the decision process which would lead to efficient allocation of resources in managing probationers falling into the various risk categories. The project team, working in conjunction with the DOC, designed the scope of the research, identified data needs, collected data, prepared the data for statistical analysis, analyzed the data, and developed a recommended probation risk prediction model.

The detailed discussions are in a two-part research report which contains the discussion of the research completed, and the results of statistical analysis, respectively. The latter is referred to as the "statistical output report" in the ensuing discussion.

Scope of the Research Study

The principal objective of the study was to develop a risk prediction model which had a high level of predictive accuracy, yet was easy to use and implement in practice. Since the prediction instrument was not going to be used for "testing" or "evaluating" various treatment alternatives on the success/failure of probation, treatment considerations were excluded from the research study. Rather, the purpose was to develop a predictive equation which will be useful in predicting the probability of recidivism as a function of criminological variables.

The study used "reconviction" as the measure for the criterion variable. Although the criterion for success/failure is inherently multi-dimensional, a dichotomous criterion variable was used in this study, consistent with the predominance of studies in the field of criminological prediction. The prediction equation that was developed was converted to a prediction table comprising three risk groups: "high," "medium," and "low" risks. The construction of such a table related to two major considerations: 1) the predictive power of the model could be very high for individual groups even if the overall power is not high, and 2) it relates to the intensity of surveillance as it pertains to individual probationers (a management consideration). Conversion of the prediction equation into a prediction table is common practice in criminological studies.

Variables Included in the Study for Prediction

The variables to be included in the study for analysis were jointly identified by the OMIS and DOC research teams. The DOC research staff were instrumental in identifying the critical variables and in directing the OMIS research team to pertinent criminological literature for identifying other potentially useful variables for prediction. Based on discussions with DOC research staff, a review of criminological prediction studies reported in the literature, and informal interviews with probation officers (during site visit to county probation offices), two sets of variables were identified for potential use in the analysis phase: 1) the "core" set of variables identified by the DOC research staff and 2) the "experimental" set of variables identified through literature search.

Research Methodology

The principal methodology used in this research was Multiple Linear Regression (MLR) analysis. The use of MLR was consistent with prevailing practices in criminological prediction. Although other methods were identified (for example, "loglinear" models and predictive attribute analysis), earlier discussions with DOC research staff, and comparative performance of these methods with MLR reported in criminological literature, led to the conclusion that these methods were not sufficiently superior to merit inclusion in the study. The selection of the methodology was guided by three considerations:

1) ease of use and implementability
2) predictive accuracy of the resulting equation
3) interpretability of the "weights" in the equation and ease of understanding

Accordingly, three different approaches were tried in the research: MLR, Multiple Discriminant Analysis (MDA), and preclustering (partitioning of the data set) plus MLR. Of these, MLR based equations produced the best results.

Sampling, Sample Size Selection, and Data Collection

The sample size in all of the studies reviewed as part of the literature search exceeded 1,000. The smallest sample size encountered was roughly 1,000 and the largest sample size was 6,000. Most of the studies have used a sample size close to 2,000. In this study, the sample size was 2,012 cases drawn from 1982 probation data provided by DOC. The 2,012 cases were divided into an analysis sample consisting of 1,002 cases and a validation sample consisting of 1,010 cases. The sample was drawn randomly from the 1982 probation data. Lists of names were created and then sorted by county code so that the Pre-Sentence Investigation (PSI) documents could be retrieved.

The PSI's were used to code the values for the 22 variables comprising the "core" and "experimental" variables. To test the validity and reliability of the variables, the data coders were given a random sample of 10 identical PSI's and were asked to code them independently. Their coding of the variables was examined to ensure that there was consistency among the coders. The variables were deemed acceptable since no problems were discovered at this stage.

A data coding "template" was designed on Lotus 1-2-3 and the data was first coded into the microcomputers. After data collection was completed, the data files for statistical analysis were created on the CDC 750 mainframe computer on the Michigan State University campus. The criterion variable was coded independently from the criminal history (rap) sheets provided by the Department of State Police.

Results of the Statistical Analysis

Several sets of analyses were performed in an effort to develop a satisfactory prediction instrument. Both the MLR approach and the preclustering coupled with MLR approach yielded usable results. Further statistical explorations were done to develop a parsimonious model for prediction purposes. The MDA approach was tried on the sample and initial explorations did not produce results comparable to those obtained through the use of MLR. Consequently, with the concurrence of the DOC research staff, the MDA approach was not pursued further. The details of the regression analyses done are presented in the body of the main report. The prediction equations developed were converted to prediction tables for assessing their predictive accuracy. Separate prediction tables were developed for the analysis and validation samples to estimate the "shrinkage" in predictive accuracy. The results showed that the equations were robust (i.e., minimal shrinkage) and had relatively high predictive power when compared to other studies reviewed in the criminological literature.

Recommendation

The recommended prediction equation is a linear additive model which is easy to understand, interpret, and use in practice. The following variables are used in the predictive equation for classifying cases into risk categories: age at first arrest, prior employment record, length of employment, total number of juvenile arrests, presence/absence of substance abuse problem, and outcome on prior probation. This prediction equation was converted to a prediction table for classification purposes. Prospective cases will be classified depending on their total score determined by the recommended prediction equation. The recommended equation is statistically significant with a multiple-r value of .3 (which compares well with other studies in criminological studies), and exhibits virtually no shrinkage when applied to the validation sample. The following table was developed to demonstrate a manual method for using the prediction equation.

Manual Method for Using Prediction Equation

(1) Item #	(2) Item Description	(3) Item Value	(4) Weight	(3) x (4) Score
1.	Longest time on one job including juvenile work record. If less than or equal to one year, CODE as 1; 1-4 years, CODE as 2; greater than or equal to 5 years, CODE as 3.		-49	
2.	Total number of juvenile arrests.		16	
3.	Outcome on prior probation, CODE as 0. If failure, CODE as 1.		119	
4.	Age at first arrest.		-6	
5.	Presence/absence of substance abuse. If no problem, CODE as 0. If alcohol or drug problem, CODE as 1. If alcohol and drug problem, CODE as 2.		62	
6.	Employed/unemployed at time of arrest. If employed, CODE as 1. If unemployed, CODE as 0.		-59	
7.	Constant for all cases.	440	1	440
		TOTAL SCORE		

Classification Rules

1. If total score is less than or equal to 200, classify as low risk.
2. If total score is between 201 and 499, classify as medium risk.
3. If total score is equal to or above 500, then classify as high risk.

CHAPTER I
INTRODUCTION

REPORT ORGANIZATION

The Department of Corrections (DOC), in an effort to further improve the decision process which leads to a probation officer's ability to allocate resources, requested the Office of Management and Information Systems, Department of Management and Budget, to conduct research into the feasibility of a probation risk model. The project team, working in conjunction with the DOC, designed the scope of the research, identified data needs, collected data, prepared the data for statistical analysis, analyzed the data, and developed a recommended probation risk prediction model.

This report details the efforts of the research team and provides a review of the progression of events which led to the recommended risk prediction model presented in the final section. It starts by identifying the main issues involved in the development of a probation risk prediction model. The rest of the report is divided into the following sections: variables included in the risk prediction model, sample size and sampling considerations, scope of the present study, the methodology used in developing the prediction model, and the results of statistical analyses.

RESEARCH ISSUES

The review of the literature indicated that the studies in the past have included both "objective" (i.e., criminological) variables and "qualitative/behavioral" variables in prediction models. There is considerable debate among researchers whether these two categories of variables should be used together in prediction models. The answer to this issue lies in part in the objective of the research for the use to which the prediction model will be put. In general, if the purpose is one of prediction and not the assessment of effects of criminological treatments (interventions) or their efficiencies, the inclusion of both categories of variables would be appropriate. In this research, there were no treatment related variables among those included in the prediction equations. The following discussion summarizes the kinds of variables that are generally considered for inclusion in prediction models and how the final list of variables was selected for the study.

Literature Review

The review of the literature indicated that nearly the same set of variables repeated itself in most of the studies. This project report does not present a complete review of the studies and the variables used in them. The list of variables used in a few of the other recent and/or important studies is mentioned below for the sake of completeness and also to permit a comparison with the research described in this report. An early prediction study with several interesting features is Vold's "Prediction Methods, and Parole" published in 1931. The variables that were used by Vold are shown below:

1. Previous criminal r^ecord
2. Marital status and county of residence
3. Prison punishment record
4. "Social type" - a classificatory variable
5. Previous work record
6. Occupation
7. Type of current offense

In contrast, the RAND study (1985) is a more recent criminological prediction study. Through regression analysis, the RAND study found that factors such as "income," "type of conviction crime," "number of prior juvenile and adult convictions," and "whether the defendant was living with spouse and/or children" accounted best for the rate of recidivism found in their samples. The following variables typify the kinds of variables used by the RAND study:

1. Number and type of conviction counts
2. Prior criminal record
3. Defendant social and economic characteristics
4. Victim characteristics
5. Drug and alcohol use
6. Weapon type and victim injury

It is useful to note that the RAND study developed several models to investigate the imprisonment/probation/criminal behavior aspects of felons. Consequently, the list of variables used across the studies is more inclusive than those mentioned above.

Another probation risk prediction study was the research done by the British Home Office. The research report, prepared by the British Home Office, reviews a number of studies in some detail. The following is a summary of the variables used in some of the studies reviewed:

Ohlin (1951)
- Type of current offense, current sentence, prior criminal record, a rating of home background, degree of current family interest, social type, work record, community of residence, job prospects on parole, personality rating, and number of associates in current offense.

Borstal study (1955)
- Evidence of drunkenness, prior offenses, prior probation or committal to prison, whether living with parents, whether home was in industrial area, and longest period in any one job. Drunkenness carried the greatest weight in this study.

Gottfredson and Beverly (1962)
- Nature of offense, county, previous delinquency record, age at first admission, court of most recent commitment, admission status (first or return), and age at release.

Quincy study (1963)

- This was the first comprehensive study to include clinical data (i.e., psychiatric evaluations and behavioral evaluations by clinical psychologists, and physical health data). The study showed that the inclusion of an environmental variable such as delinquency rate for the area of residence considerably improved the predictive power of the resulting predictive models.

This research report reviews other studies that have been carried out. The reviews range from methodological comparisons to summaries of the results obtained.

CHAPTER II
SCOPE AND RESEARCH METHODOLOGY

SELECTION OF VARIABLES

The process of selection of variables by the project team started with the review of the literature and discussions with the Department of Corrections (DOC) research staff. The DOC staff identified a number of variables that were to be considered for inclusion in the prediction model. These variables will be referred to as the "core" variables in the rest of the report. The DOC research staff were able to draw on their knowledge gained from a previous study pertaining to recidivism among parolees. The list of these variables is shown below:

1. Age at first arrest
2. Number of juvenile arrests (assaultive/non-assaultive/drug related)
3. Number of adult convictions (assaultive/drug related/non-assaultive)
4. Drug/alcohol problem (presence/absence/degree of severity)
5. Employment/income related variable
6. Current conviction

The general philosophy which the DOC research staff wished to reflect in the research study was a focus on criminal behavior and, in particular, felony behavior.

Following this discussion with the DOC staff, the members of the research team visited a county probation office with a view to understanding their operations, as well as to study the way in which the records were kept. As part of this site visit, probation officers were asked to reflect on the kinds of factors or "determinants" that they looked for before recommending a probation decision. This was done partly to assess the concordance between what has been reported in the literature and what the probation officers' actually used in arriving at probation decisions, and to identify additional variables (predictors) which might be of use in developing a predictive model. Interestingly, the variables that were identified by the probation officers were also behaviorally oriented thus confirming the prior expectations of the DOC staff as to their importance. The list of variables identified through the interviews is shown below:

1. Prior criminal record
   - length and type of crimes committed (violent/nonviolent)
   - number of felony convictions
   - disposition on prior crimes (prior probation outcomes)
   - frequency of crimes

2. Family and upbringing
   - parents separated/divorced
   - drug/alcohol problem in parents
   - criminal history for parents/siblings
   - victims of child abuse

3. Situational/victim impact statement
   - severe/mild

4. Employment
   - ability to keep a job
   - employed/unemployed

5. Education
   - high school graduate
   - behavioral problems while in school

6. Substance abuse related
   - alcohol or drug problem
   - length of the problem/any treatment
   - potential to develop a substance related problem

Based on review of the literature and several discussions with the DOC project staff, the following list of variables was developed for inclusion in this study. The list is divided into the two categories of a "core" group of variables and an "experimental" group of variables.

The list of variables used for developing the prediction equations is:

Core Variables

1. Age at first arrest
2. Number of juvenile arrests
   - number of assault related arrests
   - number of drug related arrests
   - number of nonassaultive (property related) arrests
3. Frequency of convictions
   - This variable was not broken out by category as was the previous variable.
4. Number of adult convictions
   - number of assaultive convictions
   - number of nonassaultive convictions
   - number of drug related convictions
5. Employment at time of conviction
6. "Predatory" versus "nonpredatory" behavior during the current crime
7. Presence or absence of drug problem

Noncore Variables

1. Prior probation history
   - did the defendant have prior probation?
   - was the prior probation successful?
2. Was the defendant living with parents?
3. Did the parents have a substance abuse problem?
4. Criminal history of family
   - did any member of immediate family have a criminal record?
5. Was the defendant a victim of child abuse?
6. Did the defendant have an employment history?
7. The longest time on one job including juvenile work record.
8. Had the defendant had behavioral or disciplinary problems while at school?

The criterion variable was a zero or one dichotomous variable representing probation success and failure, respectively. The use of a dichotomous criterion variable was fairly common in the literature reviewed. This is partly due to the use of multiple regression for developing the prediction equations. A multichotomous criterion variable would necessitate the use of other classification and prediction methods such as discriminate analysis. Further, the use of a multichotomous criterion variable would necessitate the creation of a recidivism scale that is more elaborate than that needed for a dichotomous classification. In this study, the five point recidivism scale developed by the DOC research staff for parole prediction was modified to initially define three groups:

- misdemeanors (all crimes that are not felonies), coded as a zero
- nonviolent felonies, also coded initially as a zero
- violent felonies, coded as a one

The instructions given to the coders for dealing with recidivism variables are shown in Appendix I. It also shows a partial list of misdemeanors, property felonies, and violent felonies. This condensed list was developed from a DOC policy directive dated July 1, 1984. Success or failure in probation was determined by using a follow-up period of two years after probation was granted. The use of two years for a follow-up period is in keeping with what has been used in prior prediction research. Appendix II shows the coding scheme used in the research.

SAMPLING, SAMPLE SIZE SELECTION, AND DATA COLLECTION

The sample size is related to three important factors in the construction of a predictive model: the number of predictor variables included in the model; the need for protection against shrinkage, which arises due to an intensive search for the model that best "fits" the data; and ensuring the statistical stability of the estimated coefficients. The sample size in all of the studies reviewed (i.e., the RAND study, DOC studies, and the British Home Office research report) exceeded 1,000. The smallest sample size encountered was roughly 1,000 while the largest sample size encountered was roughly 6,000. Most of the studies have employed a sample size close to 2,000 which was roughly divided evenly between the analysis or construction sample and the validation sample. The number of variables in these studies has ranged from a minimum of 7 to a maximum of approximately 65, with about 14 variables being quite common.

The size of the sample is also related to such characteristics of prediction studies as the arbitrariness of the cut-off period for follow-up. For example, a defendant who committed a crime just beyond the cut-off date would be classified as a success which, of course, distorts the information. There will always be such cases in the samples analyzed. The manner in which the predictor and criterion variables are coded also introduces a certain amount of distortion. For example, the use of a preponderance of dichotomous variables results in a loss of resolution when these variables are included in a regression equation as predictor variables. Finally, the quality of the data is less than what one will find in controlled experiments. For these reasons, the sample sizes in prediction studies need to be fairly large.

The sampling methods used also exhibit a diversity, ranging from convenience sampling to representative sampling. Some studies have used a specific year for their data base while others have constructed a representative data base (spanning more than a year) for their analysis sample. The use of random sampling is sometimes precluded by the base rate problem. If the base rate is low, there are two ways to construct the sample. The first is to increase the sample size so that the desired representation for probation successes and failures is achieved in the sample. The second way is to ensure that the sample success and failure rates are roughly equal and adjust the results of the statistical analysis for the true base rate in the population. A combination of random sampling and representative sampling can also be used.

For populations with a sufficiently high base rate to allow the analysis of a random sample, the problem of adjusting the predictor and altering its power does not arise. Although it would seem that a base rate of .50 would be the ideal, examination of many studies with base rates of between about .25 and .50 suggests that within this range, at any rate, the base rate is a very minor factor in determining the power that is actually obtained in practice. One of the most successful studies reported in the British Home Office research report had a base rate of .28. The aim was to develop a predictive instrument which will give a good separation of the failure rate.

It was important to use both a construction sample and a validation sample in the study. Even though the study was not intended to be exploratory, true assessment of predictive power cannot be made without a validation sample.

This research utilized a random sample of 2,600 cases from 1982 data provided by the DOC. These 2,600 cases were randomly drawn from the probation data for 1982 using SPSS (Statistical Package for the Social Sciences). The analysis sample consisted of 2,000 cases and the validation sample consisted of 600 cases. The samples were selected independently of each other.

After the samples were selected, the cases were sorted by county codes and printed out in the form of separate lists for each county. Each county was sent a list of cases for the validation sample and analysis sample. These lists were used by the counties to retrieve the PSI (Pre-Sentence Investigation) documents. The PSI documents received from the field were resorted to correspond to the analysis and validation samples. The data collection effort was coordinated through the DOC.

To test the validity and usefulness of the variables previously identified, the coders were given the same set of PSI's and asked to code them independently. Their coding of the PSI reports was examined to ensure that there was consistency among the coders. The variables were deemed acceptable since no problems were discovered at this stage.

A standard form (a "template") was developed to code the data. Initially, coding of the data was done on microcomputers using the LOTUS 1-2-3 program. This enabled the coding effort to be reduced considerably since the values for the variables did not have to be recorded twice (i.e., coding into a spreadsheet was equivalent to writing the values down on a sheet of paper). The worksheets containing the data were "uploaded" into the mainframe after data collection was completed. The use of LOTUS 1-2-3 also enabled easy verification of the data for missing values and the retrieval of lists of names, on request, for data retrieval purposes. The insertion of certain data items such as "frequency of crimes" was easier with the use of microcomputers since the spreadsheet had the capability to compute the "maximum" value for the elapsed time between consecutive crimes for the entire sample. This value was then automatically inserted into those cases in which the defendant had not committed more than one crime.

The criterion variable was coded from the criminal history (rap) sheets provided by the Department of State Police. The State Police were given separate lists of names for the analysis and validation samples. The retrieval of the rap sheets was also coordinated through the DOC. The rap sheets returned were sorted into the analysis and validation samples. The values for the criterion variables were then entered onto the worksheets.

Finally, the worksheets were checked to ensure that there were no missing data items for the cases. Records with a substantial number of missing data items were deleted from the data set. There were very few cases which had to be removed from the data set for this reason (less than 10). Cases which had one or two data items missing were updated by inserting the averages for the missing data items. The proportion of cases which had complete data was close to 99%.

SCOPE OF THE RESEARCH STUDY

This section briefly discusses the assumptions underlying the research that was conducted. The principal assumptions of the study were:

1. It was assumed that the scope of the present study will not include treatment considerations. That is, the prediction instrument will not be used to test various treatment alternatives on the success of probation.

2. It was assumed that the study will seek to develop a prediction instrument which will predict the probability of successful or unsuccessful completion of probation.

3. The criterion for success is inherently multidimensional. However, most studies have used "reconviction" as the measure for the criterion variable. This research was carried out by focusing on felonious behavior only.

4. The prediction instrument was to be converted to a prediction table. For example, the prediction table was to have similar groups such as those for whom the probability of success is "high," "moderate," and "low." The construction of such a table relates to two considerations: 1) often the predictive power of the model is very high for individual groups even if the overall power is not high; and 2) it relates to "surveillance levels" for the individua probationer. The relationship of this instrument to what is currently being done needs to be borne in mind. There might be potential for paperwork reduction, ease of use by probation agents, acceptance, and implementability.

The conversion of the regression equation and, in general, any predictive equation, into a prediction table is common practice in criminological studies and is used for classifying a convicted offender.

RESEARCH METHODOLOGY

The principal methodology employed by research studies thus far has been multiple linear regression analysis (presumably because of access to statistical packages and familiarity). Other methodologies which have been reported in the literature include:

- predictive attribute analysis
- configural analysis
- linear discriminate analysis
- scoring and weighting techniques

The use of linear regression for a criterion variable that is dichotomous could lead to problems. For example, it is possible for predicted values to be greater than 1.0 or less than zero, both cases being theoretical impossibilities. However, the use of linear regression has the merits of simplicity and that of a good enough approximation to a nonlinear functional form. Multiple regression offers the advantage of being easily understood by those who have to implement it in practice, and the conversion of the predicted probabilities into a prediction table is a well understood procedure.

Although it seemed advisable to explore other methods such as "logit" analysis or "loglinear" models, which take into consideration the dichotomous nature of the criterion variable, it was decided not to perform these analyses for two reasons. First, the equations resulting from these approaches are not easily interpreted as those from multiple regression analysis. For example, the coefficients cannot be interpreted as weights associated with the variables without appropriate transformations. Second, from the point of view of implementability, loglinear equations are less attractive than equations based on multiple regression.

The literature search indicated that discriminant functions may be worthy of exploration also. Discriminate analysis procedure could be potentially very useful for a classification problem such as the one studied in this research. However, the use of the discriminant procedure could only yield one discriminate function to separate the successes from the failures. It would not be possible to classify cases into three groups of "low", "medium", and "high" risk without resorting to some ad hoc procedures. To create a three group classification it is necessary to have defined the criterion variable as a trichotomous variable corresponding to these risk groups. Consequently, the discriminant procedure was tried only for the two-group classification problem. The results were not as good as the results obtained with the regression procedure.

CHAPTER III
RESULTS OF STATISTICAL ANALYSIS

This section describes the results of the statistical analyses done on the data set. It should be pointed out at the outset that the number of usable cases for the analysis or construction sample was 1,616, and for the validation sample it was 396. The reduction in the number of useful cases from the original random sample of 2,600 was due to a less than 100% response rate on both the PSI requests and the requests for rap sheets from the State Police.

ESTIMATION OF REGRESSION EQUATIONS

Before discussing the particulars of the results, a description of the chronology of the various analyses that were made should prove useful. The analysis of the data proceeded in three phases. Phase 1 consisted of several exploratory runs with the SPSS package to identify potentially useful regression equations for prediction purposes. In Phase 2, the data was analyzed after being partitioned (or clustered) into two distinct groups. In this step, separate regression equations were developed for the individual groups. In Phase 3, several discriminate functions were fitted to the data set assuming a two-group classification problem (as explained in the previous section). The discussion that follows presents only the final results of the analyses that were done. A considerable amount of exploration preceded the development of these equations. The brevity of the description in this report is not indicative of the extensive explorations that were done with the data. The details of the analyses are contained in the computer outputs on which this section is based.

The first step was to fit a regression equation using the entire set of variables (the core variables plus the experimental variables) to the analysis sample comprising 1,616 valid cases. In this run, the criterion variable was a zero/one variable with the code on "one" representing rearrest for a violent felony. In other words, the commission of a nonviolent felony (property related crime, for example) was coded as "zero" along with the misdemeanors. The regression equation resulting from this analysis had a multiple-r of .01. The multiple-r which indicates the (bi-serial) correlation between the predicted probabilities of recidivism for the cases in the sample and their actual probabilities (which, of course, are zeroes and ones) is the appropriate measure of the predictive power of the regression equation. The multiple-r values lie in the interval 0.0 to 1.0 The value of 0.0 corresponds to zero correlation between the predicted and observed values; the value of 1.0 corresponds to perfect prediction.

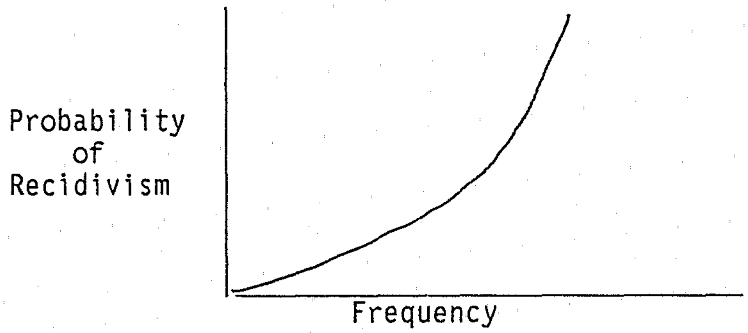
It is useful at this juncture to note that prediction studies rarely ever report a multiple-r exceeding .40. A value of multiple-r in the range of .25 to .35 should be considered good for prediction purposes. The lower the base rate in the population, and thus in the sample, the lower the multiple-r that is theoretically attainable. The problem of low base rate and the low multiple-r it produces are well documented in prior research. Typically, the studies report a multiple-r of roughly .25 with a considerable number reporting less than this value.

In view of the above discussion, a multiple-r of .01 for the first regression was unacceptably low. Several exploratory regressions with both sets of variables were run to improve the predictive power without much success. An examination of the recidivism rate in the sample revealed that it was approximately 6.5%. To overcome this extremely low base rate, the criterion variable was recoded so that a code of 1.0 reflected all felonies as opposed to violent felonies only. (This possibility had been raised before the statistical analyses were started with the DOC research staff and their concurrence had been obtained for the recoding of the criterion variable.) The base rate for the recidivism after recoding was approximately 30%. Recalling the prior discussion relating to base rate problems, the base rate of 30% is in the acceptable range for analysis to proceed without requiring adjustments for base rate. The regression was rerun and the results showed considerable improvement. The multiple-r had risen to 0.1 which was a tenfold improvement in predictive power!

After these initial results, several explorations were performed in an effort to improve predictive accuracy. The examination of regression results and prior expectations led the principal consultant to try interaction terms. The possibility of interaction terms has been alluded to in the literature. However, few studies have utilized interaction terms to improve predictive accuracy. For example, it is reasonable to expect the combined effect of alcohol use and property related crime on probability of recidivism to be more than the summation of the individual effects. Similarly, it is reasonable to expect the interaction term involving drug use and violent crimes to be significant. Pursuant to this line of reasoning, several interaction terms were introduced into the regression equation. As expected, the predictive power of the regression equation increased nearly twofold. The multiple-r increased from roughly .1 to .18 with the introduction of interaction terms.

In the next stage, several runs were made to explore the effect of nonlinear functional forms for individual variables. For example, the functional form relating the probability of recidivism to frequency of prior convictions might be a nonlinear function of the type shown in Figure 1. The reasoning being that, as one considers higher values in the horizontal axis, the propensity to commit crime is increasing.

Figure 1. Probability Versus Frequency of Crime



The functional form depicts the hypothesis that, as the propensity to commit crime increases, the probability of recidivism increases much more rapidly as opposed to near the origin where it is much flatter (reflecting the hypothesis that first time offenders are less likely to recidivate). These runs produced encouraging results. Several of the nonlinear terms proved to be statistically significant. The multiple-r values were considerably higher (in the range .22 to .28).

In an effort to continue to increase the predictive power of the regression equation, the data set was partitioned (clustered) in Phase 2 in several ways. Partitioning the data set according to whether or not the cases corresponded to those receiving prior probation proved to be useful. Accordingly, separate regression equations were developed for the two groups. The results showed considerable improvement over the results for unpartitioned data. The multiple-r increased to values in the range .35 to .43. In the ensuing discussion, the results from both Phase 1 and Phase 2 are presented.

Thus far, the equations that have been described had the full set of variables in them. Those variables that were not statistically significant were retained for two reasons. First, these equations provide a benchmark against which the other equations can be compared. Second, if interest centers on prediction only, retention of these variables in the equation does not hurt predictive accuracy. However, the length of these equations might make them difficult to implement and use. It should be noted that, if a computer is used to make the computations, these equations can be implemented just as easily as more parsimonious equations.

Next, those terms that were not statistically significant were dropped from the equations and more parsimonious models were fitted to both the entire data set and the partitioned data set. These regression equations suffered little loss in predictive power. Those variables that were significant at the 10% level or below were retained in the regression with the reduced set of variables. In the rest of this discussion these regressions are referred to as "full" regression and "reduced" regression. All the fitted regression equations are statistically significant at the 5% level. Examination of the residuals revealed no significant departures from the assumptions of a linear model.

The multiple regression analysis computer outputs from SPSS are included in the statistical output report and are marked as "regression set 1" and "regression set 2." Set 1 contains regression equations for the original analysis sample which includes the full set of variables; one for the entire data set, one for the subsample corresponding to those receiving prior probation, and one for those not receiving prior probation. Set 2 contains corresponding regression equations with the reduced set of regressors (parsimonious regressions).

PREDICTION ANALYSIS FOR REGRESSION EQUATIONS

After the estimation phase, the evaluation of the predictive accuracy of the estimated regression equations was undertaken. This step consisted of using the estimated regression equations to predict the probability of recidivism for the cases and then using the predicted score to construct a prediction table. For the purpose of constructing the prediction table, three risk groups were defined depending on the probability of recidivism score (p) as follows:

- Low risk -- $0 < p < .3$
- Medium risk -- $.3 < p < .7$
- High risk -- $.7 < p < 1.0$

The cut-off points of .3 and .7 have been used in numerous previous studies. Although these cut-off points can be changed, retaining them affords comparison to prior research.

According to this classification scheme, any case that has a predicted score of less than .3 will be classified as having a low risk of recidivism. Any case with a predicted score of .7 or higher will be classified as having a high risk of recidivism. The cases that fall in between these cut-off points will be classified as medium risk cases. The assumption is that cases classifying as low risk would be afforded minimal supervision. Cases falling into the high risk category would be afforded intensive supervision. Cases falling into the middle category will be placed under average supervision.

In converting the prediction equations to prediction tables, there will inevitably be some loss of power due to the arbitrary definition of three risk classes. However, the reduction in power is more than compensated for by the ease of use of the resulting tables. The predictive power contained in the prediction table is measured by the Ø (phi) statistic which is defined as:

$$\emptyset = \sqrt{\frac{X^2}{N}}$$

where N is the sample size; i.e., the number of cases in the sample. The value of phi, like that of multiple-r, lies between 0 and 1. The higher the value of phi, the higher the predictive power. Generally speaking, the value of phi will be close to that of multiple-r for the regression equation. The statistical significance of the value of phi is given by the chi-square statistic associated with the prediction table.

The prediction tables were constructed for both the analysis and validation samples. The results of prediction analyses are shown in Tables 1 through 4. These sets of tables correspond to full and reduced regression equations. By comparing Tables 1 and 2, it can be seen that, for the full regression, unpartitioned data set case, the phi value is .312 for the analysis sample and .224 for the validation sample. The reduction in the power is to be expected and normal, considering the size of the validation sample (396) and that of the analysis sample (1,616). The prediction table performs exceptionally well in the analysis sample. For example, in Table 1, the overall failure rate of 30% is split into individual failure rates of 17.2% for the low risk group and 82.6% for the high risk group, thus attaining excellent separation between these two groups. The smaller the failure rate for the low risk group and higher the failure rate for the high risk group, the higher the predictive power of the table. The middle group has a failure rate of 43% which is consistent with the definition and expectation for this group. Overall, Table 1 does very well on the analysis sample. The performance of prediction tables for the portioned data sets can be interpreted similarly. As can be seen from

TABLE 1

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	705	423	4	1132
FAILURE	146	319	19	484
TOTAL	851	742	23	1616
FAILURE RATE	17.2%	43.0%	82.6%	30.0%

CHI SQ = 156.946 PHI = 0.312

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	550	162	5	717
FAILURE	119	118	17	254
TOTAL	669	280	22	971
FAILURE RATE	17.8%	42.1%	77.3%	26.2%

CHI SQ = 91.062 PHI = 0.306

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	214	198	3	415
FAILURE	31	174	25	230
TOTAL	245	372	28	645
FAILURE RATE	12.7%	46.8%	89.3%	35.7%

CHI SQ = 111.647 PHI = 0.416

TABLE 2

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	171	118	6	295
FAILURE	33	63	5	101
TOTAL	204	181	11	396
FAILURE RATE	16.2%	34.8%	45.5%	25.5%

CHI SQ = 19.890 PHI = 0.224

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE - NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	137	47	2	186
FAILURE	23	20	3	46
TOTAL	160	67	5	232
FAILURE RATE	14.4%	29.9%	60.0%	19.8%

CHI SQ = 12.303 PHI = 0.230

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	39	63	7	109
FAILURE	18	33	4	55
TOTAL	57	96	11	164
FAILURE RATE	31.6%	34.4%	36.4%	33.5%

CHI SQ = 0.168 PHI = 0.032

Table 1, these prediction tables indicate a high degree of power in separating the low risk and the high risk groups. The chi-square values are highly significant compared to the critical value of 5.99 at the 5% level.

Table 2 shows the same results for the validation sample. It can be seen that predictive accuracy for the entire data set is still good. However, the predictive power drops to near zero for the partitioned data set (the chi-squared values are less than the critical value). This prediction loss is probably due to the imbalance in the sizes of the analysis sample and validation sample. It should also be remembered that not all cases retrieved to be part of these samples were returned during the data collection phase of the study. This could have affected the validation sample more than the analysis sample. That is, the validation sample does not accord well with the analysis sample. The results for the reduced regressions show similar results (see Tables 3 and 4). The full regression performs better than the reduced regression on both samples although not overwhelmingly so. The computer outputs from which these prediction tables were constructed as shown in the statistical output report and are marked: "Table 1 in text," etc., so that the printouts can be matched with the tables in the text.

The effect of changing the cut-off point for high risk is shown in Tables 5 through 8. These tables contain the same information as Tables 1 through 4 except for the new cut-off point of .6 instead of .7. The observations pertaining to Tables 1 through 4 are equally applicable to these tables. The full regression performs exceedingly well and suffers some shrinkage when applied to the validation sample. The chi-square values are significant for all but one of the cases--the one corresponding to the prior probation partition in the validation sample. The full regression once again outperforms the reduced regression but not significantly so.

In summary, at this stage, the results would suggest the use of the full regression if a microcomputer or a programmable calculator can be assumed to be available. If not, the reduced regressions can be used without much loss of power.

FURTHER ESTIMATION AND PREDICTION ANALYSIS

Since the analysis described in the previous section indicated that much of the shrinkage could be due to the small validation sample, a different approach was tried for estimating the prediction equations. To make the sample sizes approximately equal, 614 cases were randomly selected from the analysis sample and merged with the initial validation sample, thus bringing the totals for the analysis and validation samples to 1,002 and 1,010, respectively. This data set is referred to as the "reconstructed" sample from now on in the report. The analyses performed on the original sample were also performed on the reconstructed sample. The computer printouts showing the results of full and reduced regressions for the reconstructed sample are shown in the statistical output report and are marked "regression sets 3 and 4" for identification purposes.

TABLE 3

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	683	448	1	1132
FAILURE	154	327	3	484
TOTAL	837	775	4	1616
FAILURE RATE	18.4%	42.2%	75.0%	30.0%

CHI SQ = 112.473 PHI = 0.264

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	526	189	2	717
FAILURE	122	123	9	254
TOTAL	648	312	11	971
FAILURE RATE	18.8%	39.4%	81.8%	26.2%

CHI SQ = 64.094 PHI = 0.257

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	196	218	1	415
FAILURE	36	185	9	230
TOTAL	232	403	10	645
FAILURE RATE	15.5%	45.9%	90.0%	35.7%

CHI SQ = 72.336 PHI = 0.335

TABLE 4

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	169	124	2	295
FAILURE	37	63	1	101
TOTAL	206	187	3	396
FAILURE RATE	18.0%	33.7%	33.3%	25.5%

CHI SQ = 12.860 PHI = 0.180

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	130	55	1	186
FAILURE	28	18	0	46
TOTAL	158	73	1	232
FAILURE RATE	17.7%	24.7%	0.0%	19.8%

CHI SQ = 1.759 PHI = 0.087

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	40	64	5	109
FAILURE	13	41	1	55
TOTAL	53	105	6	164
FAILURE RATE	24.5%	39.0%	16.7%	33.5%

CHI SQ = 4.126 PHI = 0.159

TABLE 5

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	705	407	20	1132
FAILURE	146	282	56	484
TOTAL	851	689	76	1616
FAILURE RATE	17.2%	40.9%	73.7%	30.0%

CHI SQ = 175.263   PHI = 0.329

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	550	159	8	717
FAILURE	119	108	27	254
TOTAL	669	267	35	971
FAILURE RATE	17.8%	40.4%	77.1%	26.2%

CHI SQ = 99.600   PHI = 0.320

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	142	115	8	265
FAILURE	19	87	27	133
TOTAL	161	202	35	398
FAILURE RATE	11.8%	43.1%	77.1%	33.4%

CHI SQ = 72.343   PHI = 0.426

TABLE 6

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	171	110	14	295
FAILURE	33	60	8	101
TOTAL	204	170	22	396
FAILURE RATE	16.2%	35.3%	36.4%	25.5%

CHI SQ = 19.283 PHI = 0.221

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	137	45	4	186
FAILURE	23	19	4	46
TOTAL	160	64	8	232
FAILURE RATE	14.4%	29.7%	50.0%	19.8%

CHI SQ = 11.488 PHI = 0.223

PREDICTION TABLE:
FULL VARIABLE SET
ORIGINAL VALIDATION SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	39	53	17	109
FAILURE	18	28	9	55
TOTAL	57	81	26	164
FAILURE RATE	31.6%	34.6%	34.6%	33.5%

CHI SQ = 0.150 PHI = 0.030

TABLE 7

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	683	435	14	1132
FAILURE	154	305	25	484
TOTAL	837	740	39	1616
FAILURE RATE	18.4%	41.2%	64.1%	30.0%

CHI SQ = 119.681 PHI = 0.272

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	526	184	7	717
FAILURE	122	113	19	254
TOTAL	648	297	26	971
FAILURE RATE	18.8%	38.0%	73.1%	26.2%

CHI SQ = 69.395 PHI = 0.267

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL ANALYSIS SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	196	212	7	415
FAILURE	36	163	31	230
TOTAL	232	375	38	645
FAILURE RATE	15.5%	43.5%	81.6%	35.7%

CHI SQ = 85.911 PHI = 0.365

TABLE 8

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	169	117	9	295
FAILURE	37	61	3	101
TOTAL	206	178	12	396
FAILURE RATE	18.0%	34.3%	25.0%	25.5%

CHI SQ = 13.369 PHI = 0.184

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	130	55	1	186
FAILURE	28	17	1	46
TOTAL	158	72	2	232
FAILURE RATE	17.7%	23.6%	50.0%	19.8%

CHI SQ = 2.235 PHI = 0.098

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
ORIGINAL VALIDATION SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	40	57	12	109
FAILURE	13	37	5	55
TOTAL	53	94	17	164
FAILURE RATE	24.5%	39.4%	29.4%	33.5%

CHI SQ = 3.490 PHI = 0.146

The prediction tables are shown in Tables 9 through 12 for the case where the cut-off point for high risk is .7. In contrast to the previous set of tables, it can be seen from these tables that there is very little shrinkage when the regression equations are applied to the validation samples. All the chi-square values are highly significant when compared to the critical value of 5.99. It can be seen from these tables that the reduced regression suffers little, if any, shrinkage when applied to the validation sample. The prediction performance on the reconstructed analysis sample for the reduced regression is exceedingly good as indicated by the failure rates for the low and high risk groups. The prediction performance is acceptable when viewed in the context of the validation sample. The prediction performance on the reconstructed analysis sample for the reduced regression is exceedingly good as indicated by the failure rates for the low and high risk groups. The prediction performance is acceptable when viewed in the context of the validation sample. The regression equations are nearly identical to those obtained with the original data set. In view of the above comments, the recommendation would be to use the reduced regression derived based on the reconstructed data set.

The prediction tables were recomputed by setting the cut-off point for the high risk group at .6 instead of .7. These are shown in Tables 13 through 20. The conclusions reached in the discussion above remain unaltered when these tables are compared. The shrinkage is minimal and the chi-square values are all highly significant. The computer printouts corresponding to these tables are included in the statistical output report and are marked, "Table 13 in text," etc., for identification purposes.

In summary, the results indicate that the prediction tables developed by using multiple regression methods are statistically significant. The phi values are near the upper end of the spectrum of values reported in the literature. The shrinkage when applied to the validation sample appears to be less than what was generally reported in the literature reviewed. The recommendation is to use the reduced regression equations developed with the reconstructed sample and to use a cut-off point of .7 for the high risk category. From an implementation point of view, it might be desirable to convert the prediction equations to more manageable form by multiplying the coefficients in the equations by an appropriately large constant so that computationally they become easier to apply. It needs to be kept in mind that converting the equations for ease of application would degrade their predictive power further. It would, of course, be desirable to automate the computational scheme, in which case it will not be necessary to adjust the equations.

TABLE 9

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	463	245	3	711
FAILURE	97	175	19	291
TOTAL	560	420	22	1002
FAILURE RATE	17.3%	41.7%	86.4%	29.0%

CHI SQ = 104.891 PHI = 0.324

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	351	98	1	450
FAILURE	70	71	13	154
TOTAL	421	169	14	604
FAILURE RATE	16.6%	42.0%	92.9%	25.5%

CHI SQ = 75.142 PHI = 0.353

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	141	120	0	261
FAILURE	27	93	17	137
TOTAL	168	213	17	398
FAILURE RATE	16.1%	43.7%	100.0%	34.4%

CHI SQ = 65.505 PHI = 0.406

TABLE 10

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	453	251	12	716
FAILURE	108	172	14	294
TOTAL	561	423	26	1010
FAILURE RATE	19.3%	40.7%	53.8%	29.1%

CHI SQ = 61.487 PHI = 0.247

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	338	108	7	453
FAILURE	80	59	7	146
TOTAL	418	167	14	599
FAILURE RATE	19.1%	35.3%	50.0%	24.4%

CHI SQ = 22.076 PHI = 0.192

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	136	130	14	280
FAILURE	44	89	15	148
TOTAL	180	219	29	428
FAILURE RATE	24.4%	40.6%	51.7%	34.6%

CHI SQ = 15.496 PHI = 0.190

TABLE 11

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	448	262	1	711
FAILURE	100	190	1	291
TOTAL	548	452	2	1002
FAILURE RATE	18.2%	42.0%	50.0%	29.0%

CHI SQ = 68.438 PHI = 0.261

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	358	91	1	450
FAILURE	86	65	3	154
TOTAL	444	156	4	604
FAILURE RATE	19.4%	41.7%	75.0%	25.5%

CHI SQ = 35.408 PHI = 0.242

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.7	HIGH P >= .7	TOTAL
SUCCESS	123	137	1	261
FAILURE	30	102	5	137
TOTAL	153	239	6	398
FAILURE RATE	19.6%	42.7%	83.3%	34.4%

CHI SQ = 28.450 PHI = 0.267

TABLE 12

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	440	274	2	716
FAILURE	109	183	2	294
TOTAL	549	457	4	1010
FAILURE RATE	19.9%	40.0%	50.0%	29.1%

CHI SQ = 50.113 PHI = 0.223

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	350	101	2	453
FAILURE	79	66	1	146
TOTAL	429	167	3	599
FAILURE RATE	18.4%	39.5%	33.3%	24.4%

CHI SQ = 29.181 PHI = 0.221

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	123	134	6	263
FAILURE	30	112	6	148
TOTAL	153	246	12	411
FAILURE RATE	19.6%	45.5%	50.0%	36.0%

CHI SQ = 28.555 PHI = 0.264

TABLE 13

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	463	234	14	711
FAILURE	97	154	40	291
TOTAL	560	388	54	1002
FAILURE RATE	17.3%	39.7%	74.1%	29.0%

CHI SQ = 111.819 PHI = 0.334

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	351	95	4	450
FAILURE	70	67	17	154
TOTAL	421	162	21	604
FAILURE RATE	16.6%	41.4%	81.0%	25.5%

CHI SQ = 72.889 PHI = 0.347

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - PRIOR PROBATION

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.6	P >= .6	
SUCCESS	141	111	9	261
FAILURE	27	70	40	137
TOTAL	168	181	49	398
FAILURE RATE	16.1%	38.7%	81.6%	34.4%

CHI SQ = 74.893 PHI = 0.434

TABLE 14

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	453	241	22	716
FAILURE	108	157	29	294
TOTAL	561	398	51	1010
FAILURE RATE	19.3%	39.4%	56.9%	29.1%

CHI SQ = 66.068 PHI = 0.256

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	338	104	11	453
FAILURE	80	56	10	146
TOTAL	418	160	21	599
FAILURE RATE	19.1%	35.0%	47.6%	24.4%

CHI SQ = 22.172 PHI = 0.192

PREDICTION TABLE:
FULL VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	136	99	28	263
FAILURE	44	77	27	148
TOTAL	180	176	55	411
FAILURE RATE	24.4%	43.8%	49.1%	36.0%

CHI SQ = 19.109 PHI = 0.216

TABLE 15

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	448	256	7	711
FAILURE	100	172	19	291
TOTAL	548	428	26	1002
FAILURE RATE	18.2%	40.2%	73.1%	29.0%

CHI SQ = 81.243 PHI = 0.285

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	358	89	3	450
FAILURE	86	59	9	154
TOTAL	444	148	12	604
FAILURE RATE	19.4%	39.9%	75.0%	25.5%

CHI SQ = 40.340 PHI = 0.258

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	123	133	5	261
FAILURE	30	90	17	137
TOTAL	153	223	22	398
FAILURE RATE	19.6%	40.4%	77.3%	34.4%

CHI SQ = 36.252 PHI = 0.302

TABLE 16

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	440	266	10	716
FAILURE	109	176	9	294
TOTAL	549	442	19	1010
FAILURE RATE	19.9%	39.8%	47.4%	29.1%

CHI SQ = 50.425 PHI = 0.223

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE -NO PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	350	99	4	453
FAILURE	79	62	5	146
TOTAL	429	161	9	599
FAILURE RATE	18.4%	38.5%	55.6%	24.4%

CHI SQ = 30.464 PHI = 0.226

PREDICTION TABLE:
PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - PRIOR PROBATION

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	123	127	13	263
FAILURE	30	108	10	148
TOTAL	153	235	23	411
FAILURE RATE	19.6%	46.0%	43.5%	36.0%

CHI SQ = 28.511 PHI = 0.263

TABLE 17

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - SQUARED TERMS
RECONSTITUTED ANALYSIS SAMPLE

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	454	257	0	711
FAILURE	116	173	2	291
TOTAL	570	430	2	1002
FAILURE RATE	20.4%	40.2%	100.0%	29.0%

CHI SQ = 51.910 PHI = 0.228

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - SQUARED TERMS
RECONSTITUTED VALIDATION SAMPLE

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	440	276	0	716
FAILURE	102	188	4	294
TOTAL	542	464	4	1010
FAILURE RATE	18.8%	40.5%	100.0%	29.1%

CHI SQ = 66.815 PHI = 0.257

TABLE 18

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - NONSQUARED TERMS
RECONSTITUTED ANALYSIS SAMPLE

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	453	257	1	711
FAILURE	112	177	2	291
TOTAL	565	434	3	1002
FAILURE RATE	19.8%	40.8%	66.7%	29.0%

CHI SQ = 54.396 PHI = 0.233

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - NONSQUARED TERMS
RECONSTITUTED VALIDATION SAMPLE

	LOW	MEDIUM	HIGH	TOTAL
	P <=.3	.3< P <.7	P >= .7	
SUCCESS	440	275	1	716
FAILURE	100	190	4	294
TOTAL	540	465	5	1010
FAILURE RATE	18.5%	40.9%	80.0%	29.1%

CHI SQ = 66.742 PHI = 0.257

TABLE 19

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - SQUARED TERMS
RECONSTITUTED ANALYSIS SAMPLE

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	454	256	1	711
FAILURE	116	168	7	291
TOTAL	570	424	8	1002
FAILURE RATE	20.4%	39.6%	87.5%	29.0%

CHI SQ = 57.193 PHI = 0.239

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - SQUARED TERMS
RECONSTITUTED VALIDATION SAMPLE

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	440	272	4	716
FAILURE	102	188	4	294
TOTAL	542	460	8	1010
FAILURE RATE	18.8%	40.9%	50.0%	29.1%

CHI SQ = 60.333 PHI = 0.244

TABLE 20

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - NONSQUARED TERMS
RECONSTITUTED ANALYSIS SAMPLE

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	453	254	4	711
FAILURE	112	164	15	291
TOTAL	565	418	19	1002
FAILURE RATE	19.8%	39.2%	78.9%	29.0%

CHI SQ = 67.336 PHI = 0.259

PREDICTION TABLE:
SIMPLIFIED VARIABLE SET - NONSQUARED TERMS
RECONSTITUTED VALIDATION SAMPLE

	LOW P <=.3	MEDIUM .3< P <.6	HIGH P >= .6	TOTAL
SUCCESS	440	268	8	716
FAILURE	100	185	9	294
TOTAL	540	453	17	1010
FAILURE RATE	18.5%	40.8%	52.9%	29.1%

CHI SQ = 64.233 PHI = 0.252

FURTHER ANALYSIS OF DATA

The data was also analyzed using the discriminant analysis procedure. The computer outputs from these analyses are shown in the statistical output report. In general, the performance is less than that of the multiple regression counterparts. For example, the discriminant function corresponding to the full set of variables, but without interaction and nonlinear terms, correctly classified only 27% of the failure cases in the analysis sample. This number, when compared with the failure rates for the high risk groups in excess of 80% with multiple regression, is not very impressive. However, it should be noted that discriminant analysis did show a higher accuracy rate in classifying successful cases in the analysis sample. The discriminant functions which included interaction and nonlinear terms did not do considerably better. Although the discriminant analysis procedure was not investigated as thoroughly as the multiple regression procedure, it is less likely to be of practical use if three classification categories are desired and a dichotomous criterion variable is used. This is because the middle category will have to be arbitrarily established based on the classification functions derived from discriminate procedures. However, if the criterion variable is appropriately defined, discriminant analysis might produce good results.

CHAPTER IV
A RECOMMENDED RISK MODEL FOR PROSPECTIVE PROBATIONERS

CLASSIFICATION EQUATION

Initially, the following was the recommended regression equation for classification purposes:

$$Y = .43893 - .02240\ X_{20} + .09237\ X_{14}$$
$$+ .09952\ X_{11} + .02236\ X_4 - .05408\ X_9$$
$$- .00417\ X_1$$

Where:

Y = Predicted probability of failure on probation

X_{20} = Longest time on one job including juvenile work record

X_4 = Number of nonassaultive (property related) juvenile arrests

X_{11} = Presence/absence of drug problem

X_9 = Employed/unemployed at the time of conviction

X_1 = Age at first arrest

X_{14} = Outcome on prior probation

The regression equation above is somewhat cumbersome to use. Consequently, at the risk of losing some predictive power, it was prudent to convert the recommended equation to a more readily usable form. This is achieved by multiplying the coefficients in the regression equation by 1,000 and rounding the resulting coefficients to the nearest integer. Of course, the cut-off points for classifying prospective cases have to be revised to reflect this transformation of the predictive model. The revised classification scheme would be as follows:

If the Computed Score is:	Risk Group
0 - 300	Low
301 - 699	Medium
700 and above	High

CLASSIFICATION FORM

The predictive equation, when converted to tabular form, is presented in Table 21.

TABLE 21

ITEM #	ITEM DESCRIPTION	ITEM VALUE	WEIGHT	SCORE
1	2	3	4	5 = 3 x 4

1.	Longest time on one job including juvenile work record. If,
	less than 1 year, Code as 1;
	1-2 years Code as 2;
	2-3 years Code as 3;
	3-4 years Code as 4;
	4-5 years Code as 5;
	greater than 6 Code as 6.
			-22	

2.	Outcome on prior probation. If success or not given probation, Code as 0. If failure, Code as 1.		92	
3.	Presence/absence of drug problem. If no, Code as 0. If yes, Code as 1.		100	
4.	Number of non-assaultive (property) juvenile arrests.		22	
5.	Employed/unemployed at the time of conviction. If employed, Code as 1. If unemployed, Code as 0.		-54	
6.	Age at first arrest.		-4	
7.	Constant for all cases.	439	1	439
		Total Score		

Classification Rules

1. If total score is less than or equal to 300, classify as low risk
2. If total score is between 301 and 699, classify as medium risk.
3. If total score is equal to or above 700, then classify as high risk.

COMPUTATION ILLUSTRATION

To illustrate the use of the form, consider the following profile for a hypothetical case: current offense code is 5 (robbery), number of nonassaultive juvenile arrests is 7, less than 1 year on any job, number of juvenile assault related arrests is 3, has been identified as having both an alcohol and drug abuse problem, total number of adult assaultive convictions is 1, the defendant had experienced behavioral and disciplinary problems at school, was not employed at the time of current offense, age at first arrest was 15, and had failed prior probation. The computation of the score for this profile is presented in Table 22.

The computed total score (sum of the item scores in column 5) is: 703. Consequently, this defendant will be classified as high risk. The specified cut-off point for high risk could be adjusted downward to reflect a more stringent classification policy (e.g., a cut-off score of 600 would correspond to .6 in the original prediction table).

TABLE 22

The example of use for the hypothetical case follows:

Item No.	Item Description	Item Value	Weight	Score
1.	Less that 1 year on job	1	-22	-22
2.	Outcome on prior probation	1	92	92
3.	Presence of drug problems	1	100	100
4.	No. of juvenile property related arrests	7	22	154
5.	Unemployed at the time of arrest	0	-54	0
6.	Age at first arrest	15	-4	-60
7.	Constant term	439	1	439
	Total Score:			703

Result: Classify as high risk, since total score exceeds 700.

CHAPTER V
ALTERNATE COMPUTATION FOR HIGH RISK CATEGORY

The results discussed in Chapter IV were presented to the DOC research staff. Although the research staff was satisfied with the results, some concerns were raised regarding the low number of cases classified as "high risk." The recommended equation classifies only four cases on the average (between analysis and validation samples) into the high risk group. It was felt by the DOC research staff that, from a management point of view, the number of cases placed in the high risk category should be higher for the regression equation to be of practical value. The following suggestions were made for additional analysis:

- Use total number of juvenile arrests instead of property related juvenile arrests only.

- Create a new variable called "substance abuse" which captures both alcohol and drug related problems. (Note that the recommended equation did not have alcohol abuse as a prediction.)

- Recode the length of employment variable so that the number of categories are minimized.

The computer programs that were used to produce the previously discussed results were rewritten to accommodate the suggested changes. A frequency distribution was run on the length of employment variable; the results of which are as follows:

Length of Employment (X_{20})

Code	Absolute Frequency	Relative Frequency	Cumulative Frequency	Recoded Categories
1	825	51.1	51.1	1
2	216	13.4	64.4	
3	203	12.6	77.0	2
4	100	6.2	83.2	
5	55	3.4	86.6	
6	217	13.4	100.0	3

Based on the above frequency distribution, the length of employment was recoded into three categories instead of six as was originally done. The new variable, substance abuse, was defined as the sum of the variables pertaining to alcohol abuse and drug abuse, both of which were 0/1 (dichotomous) variables. Thus, the substance abuse variable was coded on a three point scale with "0" defining absence of both problems, "1" defining either alcohol or drug problem and "2" defining the presence of both problems. In addition, two new variables were tried in the equation. These were "total adult convictions" and the interaction effect of this variable with total juvenile arrests.

Regression analysis was performed with these modifications on the reconstructed analysis sample. The computer output pertaining to this equation is shown in the output marked C in the statistical output report. The regression equation was significant with a multiple-r of 0.3. Prediction analysis was done for this equation on the analysis and validation samples. Several cut-off points (.7, .6, and .55) were tried for high risk in an effort to increase the number of cases classified as high risk. The final results are shown in Table 23. It should be noted that the cut-off point for high risk is set at .55.

Although the results shown in Table 23 are comparable to those in Table 18, it can be seen that the total number of cases classified into the high risk group is approximately 3%, which is below the acceptable level of 5%. Additionally, the sign of the interaction term involving total number of juvenile arrests and adult convictions is not easily interpretable. The next stage in the analysis involved performing new regression analyses that excluded the interaction term and the variable on total number of adult convictions (which was only marginally significant in the previous regression run). All the terms in the newly developed regression equation were statistically significant at the 5% level. The signs of the coefficients in the equation were consistent with prior expectations. Prediction tables were developed for this equation using various cut-off points for high risk and low risk groups (see Tables 24 and 25). The final choice of cut-off points were .2 for the low risk group and .5 for the high risk group. The prediction tables corresponding to these analyses are shown in Table 26. It can be seen that the chi-square and phi values compare favorably with those for Table 18 (that corresponding to the initially recommended equation). The shrinkage associated with this new equation is minimal when applied to the validation sample (i.e., the phi value decreases from .226 to .213). The separation among the three groups as measured by the failure rates was deemed acceptable by the DOC research staff. The prediction analysis showed that the number of cases placed in the high risk category was roughly 6%. In contrast to the previously presented prediction tables, the results in Table 26 also show that the failure rate for the medium risk group is close to the base recidivism rate of approximately 29%.

MODIFIED CLASSIFICATION EQUATION

The recommended modified equation for classification purposes is:

$$Y = .44039 + .01582 * \text{JUVARR} - .04884 * \text{EMP}$$
$$+ .11878 * X_{14} - .00573 * X_1$$
$$+ .06199 * \text{SUBBAB} - .05924 * X_9$$

Where:

Y = Predicted probability of failure on probation

JUVARR = Total number of juvenile arrests

EMP = Longest time on one job

TABLE 23

PREDICTION TABLE:
MODIFIED PARSIMONIOUS VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.55	HIGH p>=.55	TOTAL
SUCCESS	461	235	15	711
FAILURE	118	155	18	291
TOTAL	579	390	33	1002
FAILURE RATE	20.38%	39.74%	54.5%	29%

CHI. SQ. = 53.12 PHI = .230

PREDICTION TABLE:
MODIFIED PARSIMONIOUS VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.55	HIGH p>=.55	TOTAL
SUCCESS	464	235	17	716
FAILURE	119	157	18	294
TOTAL	583	392	35	1010
FAILURE RATE	20.4%	40%	51.42%	29.1%

CHI SQ. = 52.18 PHI = .227

TABLE 24

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.55	HIGH p>=.55	TOTAL
SUCCESS	461	235	15	711
FAILURE	118	155	18	291
TOTAL	579	390	33	1002
FAILURE RATE	20.4%	39.7%	54.5%	29.0%

CHI. SQ. = 53.170 PHI = .230

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.55	HIGH p>=.55	TOTAL
SUCCESS	460	238	18	716
FAILURE	122	159	13	294
TOTAL	582	397	31	1010
FAILURE RATE	21.0%	40.1%	41.9%	29.1%

CHI SQ. = 44.222 PHI = .209

TABLE 25

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.5	HIGH p>=.5	TOTAL
SUCCESS	461	224	26	711
FAILURE	118	142	31	291
TOTAL	579	366	57	1002
FAILURE RATE	20.4%	38.8%	54.4%	29.0%

CHI. SQ. = 55.751 PHI = .236

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL CASES

	LOW p<=.3	MEDIUM .3<p<.5	HIGH p>=.5	TOTAL
SUCCESS	460	219	37	716
FAILURE	122	143	29	294
TOTAL	582	362	66	1010
FAILURE RATE	21.0%	39.5%	43.9%	29.1%

CHI SQ. = 44.705 PHI = .210

TABLE 26

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED ANALYSIS SAMPLE -- ALL CASES

	LOW p<=.2	MEDIUM .2<P<.5	HIGH p>=.5	TOTAL
SUCCESS	200	485	26	711
FAILURE	29	231	31	291
TOTAL	229	716	57	1002
FAILURE RATE	12.7%	32.3%	54.4%	29%

CHI SQ. = 51.179 PHI = .226

PREDICTION TABLE:
REDUCED - AGGREGATED VARIABLE SET
RECONSTITUTED VALIDATION SAMPLE - ALL COSTS

	LOW p<=.2	MEDIUM .2<P<.5	HIGH p>=.5	TOTAL
SUCCESS	201	478	37	716
FAILURE	27	238	29	294
TOTAL	228	716	66	1010
FAILURE RATE	11.8%	33.2%	43.9%	29.1%

CHI SQ. = 45.98 PHI = .213

X_{14} = Success/failure on prior probation

X_1 = Age at first arrest

SUBBAB = Presence/absence of substance abuse problem

X_9 = Employed/unemployed at the time of arrest

This regression equation can be converted to tabular form, for ease of use, as shown in Table 27. The revised classification scheme would be as follows:

If the Computed Score is	Risk Group
0 - 200	Low
201 - 499	Medium
500 and above	High

At the request of DOC research staff, additional analysis was performed by recording the length of employment variable into two categories instead of three as in the above equation. The results showed that the length of employment variable (EMP) was no longer statistically significant. As can be seen from the frequency distribution for EMP, recording EMP into two classes reduces the variance associated with this variable and thus causes it to be of no value as a predictor in the regression equation. Additionally, variable X_{10}, the presence/absence of predatory behavior, was included in the regression equation both by itself and as an interaction with other variables in the equation. The results showed that X_{10} was not useful as a predictor either singly or jointly with other variables. The results of these exploratory analyses are shown in computer outputs marked (a) through (h) in the statistical output report.

It is useful to note that in Table 27 the signs of all the coefficients conform to prior expectations. An illustrative computation is presented in Table 28.

TABLE 27

(1)	(2)	(3)	(4)	(3) x (4)
Item #	Item Description	Item Value	Weight	Score
1.	Longest time on one job including juvenile work record. If less than or equal to one year, CODE as 1; 1-4 years, CODE as 2; greater than or equal to 5 years, CODE as 3.		-49	
2.	Total number of juvenile arrests.		16	
3.	Outcome on prior probation, CODE as 0. If failure, CODE as 1.		119	
4.	Age at first arrest.		-6	
5.	Presence/absence of substance abuse. If no problem, CODE as 0. If alcohol or drug problem, CODE as 1. If alcohol and drug problem, CODE as 2.		62	
6.	Employed/unemployed at time of arrest. If employed, CODE as 1. If unemployed, CODE as 0.		-59	
7.	Constant for all cases.	440	1	440
		TOTAL SCORE		

Classification Rules

1. If total score is less than or equal to 200, classify as low risk.
2. If total score is between 201 and 499, classify as medium risk.
3. If total score is equal to or above 500, then classify as high risk.

TABLE 28

Using the tabular format for the hypothetical case follows:

Item No.	Item Description	Item Value	Weight	Score
1.	Less than 1 year on job	1	-49	-49
2.	Total juvenile arrests	10	16	160
4.	Outcome on prior probation	1	119	119
5.	Age at first arrest	15	-6	-90
6.	Both drug and alcohol problem	2	62	124
7.	Unemployed at the time of arrest	0	-59	0
9.	Constant term	440	1	440
		TOTAL SCORE		704

Result: Classify as high risk since total score exceeds 500.

BIBLIOGRAPHY

1. Simon, F. H., Prediction Methods in Criminology, A Home Office Research Unit Report, Great Britain, 1971.

2. Petersilia, J., Turner, S., Kahan, J. and J. Patterson, "Granting Felons Probation-Public Risks and Alternatives," RAND Research Report, January 1985.

3. Vold, G. B., Prediction Methods applied to problems of classification within institutions, Journal of Criminal Law and Criminology, Chicago. Vol. XXVI, page 205, 1935.

4. Ballard, K. B., Association analysis in prediction studies of parolees. Research Report, Department of Institutions, State of Washington.

5. Tarling, R., and J. A. Perry. "Statistical methods in criminological prediction," in Prediction in Criminology, Farrington, D. P., and Tarling, R., Editors, pages 211-231.

6. Babst, D. V., Inciardi, J. A., and D. R. Jayman. "The use of configural analysis in parole prediction research," Canadian Journal of Criminology and Corrections, Vol. 13, No. 3 (July 1971).

7. Brown, L. D., "The development of a parole classification system using discriminant analysis," Journal of Research in Crime and Delinquency, January 1978, pages 92-105.

8. Murphy, T. H., Prediction of Minimum Security Walkaways, Research Report, Department of Corrections, State of Michigan, January 1984.

APPENDIX I

CODING INSTRUCTIONS

RECIDIVISM INSTRUCTIONS

1.) Recidivism is defined as:

0 = No illegal activity or misdemeanor arrest

1 = Nonviolent felony arrest

2 = Violent felony arrest

2.) Establish Date of Probation, add two years to determine follow-up period.
Example: P.D. = 2/19/81-------- 2/19/83

3.) Determine Crimes Committed During Follow-Up Period.

4.) Determine Most Serious Felony Arrest (if applicable).

Example 1.	3/19/82	Aggravated Assault	Code	0
	5/10/82	Drunk and Disorderly	Code	0

Since both are misdemeanors, the recidivism code is 0.

Example 2.	3/19/82	Assault and Battery	Code	0
	5/19/82	Larceny Under $100	Code	0
	7/2/82	Uttering and Publishing		1
	9/10/82	Felonious Assault	Code	2

Recidivism Score = 2

NOTE: (1) If no offenses are listed after the probation, the recidivism score is assumed to be "0".

(2) For assistance in determining offense severity, please refer to the attached lists of misdemeanors and felonies.

(3) Use arrest charge and not final charge.

MOST COMMON TYPES OF MISDEMEANORS

Assault and Battery (A & B)

Aggravated Assault

Resisting Officer

Larceny Under $50, $100, etc.

Anything Under (Key word is under (misd.) vs. over (felony)

Shoplifting

Petty Theft

Petty Larceny

Simple Larceny

Joyriding

Disorderly

Illegal Entry

Checks NSF under $50

Motor Vehicle Tampering

PROPERTY OFFENSE

The study used the following list of property/nonviolent offenses for coding purposes.

Arson
- (All except dwelling)

Burglary
- Breaking and entering
- Entering without breaking
- Breaking and entering; or entering without breaking; buildings, tents, boats, railroad cars; entering public buildings when expressly denied.
- Burglar's tools, possession

Larceny
- Larceny
- Larceny from motor vehicles or trailers
- Breaking and entering coin operated telephone, penalty
- Larceny from vacant dwelling
- Larceny from building
- Larceny by conversion, etc.
- Larceny by false personation
- Larceny from libraries
- Receiving or concealing stolen property. Note: May be referred to as RCSP.

Auto Theft
- UDAA (Unlawfully driving away an automobile)
- UDAA without intent to steal

Forgery - Uttering and Publishing Note: May be referred to as U&P.
- Forgery of records and other instruments
- Uttering and publishing
- Forgery of notes
- Forgery of bank bills and notes
- Possession of counterfeit notes, etc., with intent to utter same
- Uttering counterfeit notes, etc.
- Possession of counterfeit bank, state or municipal bills or notes
- Affixing fictitious signature
- Counterfeiting and possession of coins
- Certifying checks/insufficient funds
- Checks without accounts or insufficient funds, usually over a certain amount.

Embezzlement

- All forms except when noted as under a certain amount.

Fraud

- Building contr. funds-fraud, use
- False pretenses with intent to defraud
- Personal property, fraudulent disposition

Malicious Destruction

- All forms except when noted as under a certain amount

Weapons

- Carrying concealed weapons
- Carry weapon with unlawful intent
- Weapons manufacture

Drugs

Because of the State Police reporting format, it is sometimes difficult to distinguish between misdemeanors and felonies. The general rule is that illegal use or possession with intent to use is a misdemeanor, and the sale or possession with intent to sell is a felony. Unfortunately, the State Police may only list Dangerous Drugs or Violation of Drug Law (VDL). The following procedures should minimize any coding difficulties:

1) Dangerous Drugs or Violation of Drug Law with the designation of use is considered a misdemeanor. Illegal use and possession of drug paraphernalia are also misdemeanors.

2) Dangerous Drugs or Violation of Drug Law with the designation of sale or manufacture is considered a felony.

3) When the only information available is Dangerous Drugs, use the disposition (if listed) to determine seriousness. A disposition of greater than 1 year is considered a felony (e.g., 2 years probation, 6 months jail and 5 years probation). All prison sentences are felonies (e.g., 6 months-2 years, 10 years-20 years). Sentences of jail terms only are misdemeanors (e.g., 6 months jail, 30 days).

4) When no disposition is available and a coder cannot determine use or sale, then assume a felony when only designated as Dangerous Drugs or VDL.

VIOLENT OFFENSES

Homicide
- First Degree Murder
- Second Degree Murder
- Manslaughter
- Attempted Murder

Rape/Criminal Sexual Conduct
- Rape, Forcible
- (Does not include statutory)
- Assault with intent to rape
- Criminal Sexual Conduct 1st, 2nd and 3rd
- Attempt or Assault to Commit CSC

Kidnapping
- Kidnapping (all forms)

Assault
- Felonious assault
- Assault with intent to commit murder
- Assault with intent to do great bodily harm less murder
- Assault with intent to maim
- Assault with intent to commit felony
- Extortion

Robbery
- Robbery armed - any weapon or indication thereof
- Robbery unarmed
- Bank, safe, and vault robbery
- Assault to commit robbery-armed
- Assault to commit robbery-unarmed
- Attempted robbery
- Larceny from person

Children
- Child exposing with intent to injure
- Cruelty to children
- Torturing of children

Sex
- Sodomy
- Gross indecency between males
- Gross indecency between females
- Males under 15, debauching by females
- Males under 15, debauching by males
- Female patient in institution for insane, ravish, abuse
- Female ward, carnal knowledge

Other violent

- Arson of a dwelling
- Place explosives to damage or injure
- Possession of bomb
- Explosive device
- Careless use of firearms to kill

APPENDIX II

Coding for Predictor Variables

Guidelines to Coding:

"Child Abuse" Coded 0/1. If not mentioned in PSI, then code as zero.

Core set of Variables

0 = No (not a felony)
1 = Yes (it is a felony)   Criterion Variable

a. "Current conviction": Coded 1 through 17 (see attachment)
b. Age at first arrest
c. Number of juvenile arrests
   Number of assault related arrests
   Number of drug related arrests
   Number of nonassaultive arrests (property)
d. "Frequency" of convictions
   Average inter incident time for all categories of crimes not broken out by type of conviction (months).
e. Number of adult convictions
   Assaultive
   Drug related
   Nonassaultive
f. Employed at the time of conviction? [0 = no, 1 = part time, 2 = full time] Preference: 0/1
g. Predatory/nonpredatory behavior in current conviction. Code: 0/1 No/Yes.
h. Drug problem? 0 = No problem
   1 = Yes (there is a problem)

"Noncore" or Auxiliary Variables

a. Did the person have prior probation? (0 = no, 1 = yes)
b. Outcome on prior probation (0 = success, 1 = failure)
c. Living with both parents? (0 = no, 1 = yes)
d. Substance abuse problem with either parent? (0 = no, 1 = yes)
e. Criminal history of family (parents and/or siblings)
f. Child abuse? (0 - no, 1 = yes)
g. Previous employment history? (0 = no, 1 = yes)

OFFENSE CODES

01	Homicide
02	Rape/CSC
03	Kidnapping
04	Assault
05	Robbery
06	Children
07	Sex (other)
08	Other Violent
09	Burglary
10	Larceny
11	Auto Theft
12	Forgery - Uttering & Publishing
13	Fraud & Embezzlement
14	Malicious Destruction
15	Arson
16	Drugs
17	Other

APPENDIX III

List of Variables

LIST OF VARIABLES

X_1 Age of first arrest (any type)

X_2 Number of juvenile assault arrests

X_3 Number of juvenile drug arrests NOTE: $JUVARR = X_2 + X_3 + X_4$

X_4 Number of juvenile property arrests

X_5 Frequency of adult convictions

X_6 Number of adult assaultive convictions

X_7 Number of adult drug convictions

X_8 Number of adult property convictions

X_9 Employed at time of conviction 0-No, 1-Yes

X_{10} Predatory behavior 0-No, 1-Yes

X_{11} Drug problem 0-No, 1-Yes

X_{12} Current conviction

X_{13} Prior probation 0-No, 1-Yes

X_{14} Success of probation 0-Success, 1-Failure

X_{15} Living with both parents 0-No, 1-Yes

X_{16} Substance abuse parents 0-No, 1-Yes

X_{17} Criminal history parents/siblings 0-No, 1--Yes

X_{18} Child abuse 0-No, 1-Yes

X_{19} Previous employment history 0-No, 1-Yes

X_{20} Longest time on one job

1= 1 year	4= 3-4 years
2= 1-2 years	5= 4-5 years
3= 2-3 years	6= 5 years

X_{21} Behavioral/disciplinary problems in school 0-No, 1-Yes

X_{22} Alcohol abuse 0-No, 1-Yes

$SUBBAB = X_{11} + X_{22}$