the prediction of juvenile offending: a new zealand study

D. M. Fergusson, A. A. Donnell,
S. W. Slater, J. K. Fifield

research report no. 3

research unit,
joint committee on young offenders,
new zealand.

THE PREDICTION OF JUVENILE OFFENDING: A NEW ZEALAND STUDY

D. M. FERGUSSON (Senior Research Officer, J.C.Y.O.)

ANNE DONNELL (Research Officer, J.C.Y.O.)

S. W. SLATER (Senior Lecturer in Psychology, Victoria University of Wellington)

JUNE FIFIELD (Assistant Research Officer, J.C.Y.O.)

Research Report No. 3

Research Unit,
Joint Committee on Young Offenders,
NEW ZEALAND.

A. R. Shearer, Government Printer, Wellington, New Zealand—1975

BRN 9/89

PREFACE

This report is the third in a series describing the results of a major longitudinal study into young offending in New Zealand. In this study we attempt to answer the question, "To what extent is it possible to identify potential young offenders at 10 years of age?". The results of our analysis suggest that some prediction is possible but that the practical utility of the findings is low. At best, the results can be used to provide broad guidelines concerning those children who have high or low risks of offending.

The paper is necessarily highly statistical; it has also provided us with the opportunity to present some new statistical approaches to the problem of criminological prediction. We hope that these theoretical contributions will assist the development of an adequate methodology in this area. At this point we would like to acknowledge a debt of intellectual gratitude to Frances Simon. Mrs Simon's work in The Prediction of Probation Success laid much of the theoretical groundwork for the research reported here and we have borrowed freely from this research both for results and for the presentation of material. In many ways, the theoretical developments we suggest are extensions to the basic framework which has been laid down by Mrs Simon.

D.M.F.
A.A.D.
S.W.S.
J.K.F.

CONTENTS

CHAPTER			Page
1	INTRODUCTION, PREVIOUS RESEARCH AND PREDICTION METHODS		
	1.1	Introduction	1
	1.2	Review of Selected Prediction Studies	3
	1.3	The Prediction Problem and Models	14
	1.4	The Need for Validation	23
	1.5	The Base Rate Problem	24
	1.6	The Measurement of Predictive Power	27
2	DATA COLLECTION AND PREPARATION		
	2.1	Sampling and Data Preparation	44
	2.2	Analysis Sample and Variable Definitions	49
3	PREDICTION RESULTS		
	3.1	Introduction	50
	3.2	Delinquency Prediction Scores	52
	3.3	A Revised Delinquency Prediction Instrument	55
	3.4	AID Analysis	63
	3.5	The Effects of Other Variables	75
	3.6	Evaluating the Results	88
4	CONCLUDING COMMENTS		
	4.1	The Predictive Power of the BSAG	99
	4.2	The Use of the BSAG	103
	4.3	Reasons for the Low Predictive Validity of the BSAG	105
	4.4	In Defence of Prediction Research	108
REFERENCES			114
APPENDICES			121

CHAPTER 1

INTRODUCTION, PREVIOUS RESEARCH AND PREDICTION METHODS

Section 1.1 Introduction

A previous report (Fergusson, Donnell and Slater (1975b)) presented a detailed analysis of the structure and content of the Bristol Social Adjustment Guide (BSAG) as applied to a sample of 5,472 ten year old New Zealand boys. In the present report we extend this analysis to a consideration of the extent to which BSAG and other data collected at age ten years predict juvenile offending by the age of 17 years.

In 1967, every boy born in 1957 attending a New Zealand State school was the subject of a questionnaire completed by his class teacher. This questionnaire contained a copy of the 1956 version of the BSAG and a number of questions on the boy's background, school performance, personal characteristics and health (cf. Fergusson, Donnell and Slater (1975b)). The sample was then followed up until the end of 1973 to determine the frequency with which its members came to the attention of the Children's Court for various offences and other forms of misbehaviour. The purpose of this process was to establish a body of data to form the basis of an analysis of the relationship between the information collected at age ten and subsequent juvenile offending.

Previous research (Stott 1959) has indicated that BSAG scores discriminate between delinquent and non-delinquent boys. However, the research was conducted using a cross-sectional comparison of known delinquents and non-delinquents. As Simon (1971) has pointed out, this design is likely to be contaminated by the fact that teachers may rate known delinquents more adversely than non-delinquents. In view of this, the apparent predictive power of the BSAG may merely reflect biases in the way in which the instrument was completed by teachers. These problems can be overcome by a longitudinal design in which a sample of subjects is measured on the BSAG and then followed up for some fixed time period to determine the extent to which BSAG scores are related to future offending.

An analysis of the predictive capacity of the BSAG offers the following advantages:

(1) A suggestion made frequently in criminological literature is that by the time young offenders come to Court it is often too late to change the conditions which gave rise to their offending and that, because of this, early detection of potential young offenders is an essential step in the prevention and reduction of juvenile crime (Glueck and Glueck 1950, 1959; Herzog 1960; Stott 1960a; Venezia 1971).

(2) The second use of the findings is more abstract. The development of predictive devices may extend or modify existing theory on the causes or nature of juvenile offending.

The remainder of this chapter is devoted to developing an appropriate theoretical background for the analysis in the body of the report; the following issues will be considered:

(1) An examination of a number of selected studies which have attempted to identify factors associated with potential young offending. It must be stressed that this is not intended as an extensive survey of the literature on prediction studies. Rather, the examples are illustrative of some of the more important studies in this area.

(2) A semi-formal treatment of the statistical problems and methods involved in constructing, validating and evaluating prediction instruments.

Section 1.2 Review of Selected Prediction Studies

The Gluecks

The research reported by Sheldon and Eleanor Glueck (1950) is one of the largest and most controversial studies in criminological prediction. In this study the Gluecks compared a matched sample of 500 delinquent and 500 non-delinquent boys on a large number of variables. This information was then refined by a series of contrasts of the variables in the delinquent and non-delinquent populations. The Gluecks concluded that five variables, or as they describe them, factors - discipline by father; supervision by mother; affection of father for boy; affection of mother for boy and cohesiveness of family - best discriminated between the delinquent and the non-delinquent boys. On the basis of this finding the Gluecks constructed a five factor prediction table which assigned a score to any subject by summing the five factors weighted according to the relative frequency of occurrence of the factor in the delinquent and non-delinquent populations. The Gluecks claimed to be able to predict delinquency from this score.

The results of the Gluecks' research have been extremely controversial. A complete summary of the issues involved is beyond the scope of this review but the following are the major points at issue:

(1) A criticism that has been levelled by a number of authors (Reiss 1951; Stott 1960a; Duncan 1960; West and Farrington 1973) is that the original sample on which the Gluecks constructed their prediction instrument had an artificial base rate of offending of 50% whereas in the population at large the incidence of offending is of the order of 10%. It is well known that the level of prediction achieved for a sample in which the base rate is 50% will tend to be greater than that for a sample in which the base rate of offending is 10% (Meehl and Rosen 1955). The result of the use of the 50% base rate is that the figures presented by the Gluecks tend to provide over-optimistic estimates of the power of the five factor table if it were to be applied to the general

population. This point is clearly illustrated by Reiss (1951) who recomputed the risk estimates for the Glueck table for a population containing only 10% delinquents: the reduction in predictive power that followed this adjustment was quite dramatic.

(2) A second criticism is that the results were based on a cross-sectional comparison and not on a longitudinal study (West and Farrington 1973). In general, the results of cross-sectional comparisons indicate those factors which discriminate between delinquents and non-delinquents; they do not necessarily indicate the factors which predict delinquency (Thurston et al 1971). A variable can be established as a predictor only if it is measured prior to the predicted outcome.

(3) The study has also been criticised on the grounds of failure to consider sociological factors (Taft 1951; Reiss 1951); the selection of delinquents from penal institutions (Rubin 1951; West and Farrington 1973); the unrepresentativeness of the sample (Rubin 1951; Shaplin and Tiedman 1951); the possible unreliability of the ratings and the fact that the Gluecks capitalised rather heavily on chance in weighting their categories (Prigmore 1963).

The Gluecks' response to these criticisms has not always been entirely satisfactory and frequently they have asserted that "the proof of the pudding is in the eating". By this they mean that the efficacy of their prediction tables should be tested by further validation studies. A resume of attempts to validate the Glueck tables is presented in Identification of Predelinquents, (Glueck and Glueck 1972). On the whole, the findings in this volume support the idea that both the Glueck five factor table and a revised three factor table (Glueck 1960), are able to identify children with high and low risks of juvenile offending at an early age. Unfortunately, the standard of analysis in the validation studies is not high and it is not always possible to gain a complete indication of the extent to which accurate predictions could be made. Marshall (1973), in reviewing the Identification of

Predelinquents, has this to say about the contents of the volume:

> " The first part of the book consists of "validation" studies, none of which will do anything to convert the already teeming ranks of critics of the Gluecks' method. The first paper, by Elmering, is in fact solely concerned with "retrospective" validations, and the failure to give sufficient details of the (mostly unpublished) studies mentioned by which to judge their value is typical of the cavalier manner in which methodology is treated in this book, despite the fact that some of the papers indicate problems with the subjectivity of the ratings. The papers by Glick and by Tait and Hodges try to tackle some of the problems and are perhaps the best represented here, but both employ selected samples, in the latter case selected on the basis of school misbehaviour. The paper by LaBrie on "Verification of Glueck Prediction Table by Mathematical Statistics following Computerised Procedure of Discriminant Function Analysis" provides a mystical superstructure which avoids the fundamental problems of sample and variable selection and of prediction rates in a real community situation with a normal delinquent/non-delinquent ratio" (p.410).

We cannot disagree with these criticisms.

Hathaway and Monachesi

The results of pioneer studies in the late 1940s (Capwell 1945; Monachesi 1948, 1950) suggested that the Minnesota Multiphasic Personality Inventory (M.M.P.I.) might have some predictive potential in delinquency research. The M.M.P.I. is a widely used psychometric instrument designed to provide measures of the more clinically important aspects of personality. The content of the test forms a series of sub-scales; the early studies showed that most of these scales reliably differentiate known delinquent and non-delinquent groups. The largest score differences have been consistently found to be on scale 4 (Psychopathic deviate), scale 6 (Paranoia), scale 7 (Psychasthenia), scale 8 (Schizophrenia) and scale 9 (Hypomania).

These preliminary results led to the undertaking of a large longitudinal study of the predictive power of the M.M.P.I. (Hathaway and Monachesi 1953). The population considered in the study was all ninth-grade public school registrations in Minneapolis in 1947 - 1948. This comprised a total of 4,572 children from which a testing programme yielded 4,048 completed M.M.P.I. forms. Two years after testing, follow-up information on offending was obtained by means of a search through the local official probation and police records. The results of the research were summarised in a prediction table which showed risks of offending for 17 groups defined by M.M.P.I. scores. These risks ranged from 9% to 49%.

Although the authors stress that the results are in preliminary form, at least two criticisms of the research can be made. First, the prediction table was not cross-validated with the result that the findings probably give an overly optimistic impression of the predictive power of the M.M.P.I. Second, the criterion of offending used included subjects who offended both before and after testing, with the result that the study gave estimates of the predictive power of the M.M.P.I. for both prospective and retrospective data. The interpretation of an offending criterion based on such data is not entirely clear.

Rempel (1958) drew on the data collected by Hathaway and Monachesi in a prediction study using multivariate statistical techniques. In this study, Linear Discriminant Function analysis was used to determine the extent to which boys could be correctly classified as potential delinquents or non-delinquents from M.M.P.I. profiles. The sample used was 351 delinquent and 350 non-delinquent boys drawn from the sample of ninth graders tested by Hathaway and Monachesi in 1948. The choice of M.M.P.I. variables for inclusion in the classification formula was made on the basis of their contributions to Rao's Generalised Distance Function (Rao 1947). Rempel states the results of the analysis as follows:

> " The techniques employed proved to be effective to the extent that 62.3 per cent of the non-delinquents and 69.5 per cent of the delinquents were correctly identified by the use of multiphasic data alone" (p.22).

Rempel achieved moderate predictive accuracy. However:

(1) No attempt was made to correct for the 50% offending rate in the sample, therefore the results provide inflated estimates of the predictive power of the M.M.P.I. as applied to the general population.

(2) Rempel discarded from his analysis all those offenders in the less serious delinquency groups and this would have the effect of increasing the level of prediction obtained.

The weight of the available evidence seems to indicate that the M.M.P.I. has some validity as a predictor of future delinquency, especially when extreme offending criteria are adopted. Its success is probably limited by the fact that it was not designed for use with juvenile populations.

Mannheim and Wilkins : Borstal Success Prediction

Mannheim and Wilkins (1955) measured a random sample of 720 youths, who entered Borstal between August 1946 and July 1947, on 61 variables obtained from their records. The sample was then followed up for four years to determine the extent to which it was possible to predict from the data collected which boys would re-offend. Each of the 61 variables was correlated with future offending and the ten variables which showed the highest correlation with offending were selected. These variables included such factors as total number of convictions prior to entry to Borstal, age at first offence, number of misdemeanours in Borstal, etc. The selected variables were combined in a multiple regression equation to predict further offending. The equation was then transformed to a table which showed the risk of future offending corresponding to any particular score. The results produced by Mannheim and Wilkins indicated that prediction of Borstal success or failure was to some extent possible: the group of boys with the lowest scores on the multiple regression equation had only a 13% chance of re-offending within four years while those with the highest scores had an 87% chance of re-offending. These results were validated on a new sample which was followed up for a period of three years. By and large, the validation process supported

the idea that the table had predictive utility and the risk estimates for the validation table appear to be similar to those derived in the construction table. The main conclusions from the validated risk table were as follows: it was possible to place the boys in three groups: a group of "successes" who had a 25% chance of re-offending; a group of "unpredictables" who had a 50% chance of re-offending and a group of "failures" who had about a 70% chance of re-offending.

The work of Mannheim and Wilkins is extremely thorough and few methodological criticisms can be levelled at the results. Perhaps the major comment that can be made is that in the construction sample nearly 50% of the observations were discarded because of missing data. However, since the table was validated on a fresh sample of observations, the worst effect that could have followed was that the equation derived on the construction sample could have been a less than optimal predictor.

Stott (1962) has criticised the Mannheim and Wilkins score on the grounds that the information on which the score was based was drawn from official records and was necessarily limited in its predictive power. Nevertheless, the level of prediction achieved by Mannheim and Wilkins was high in comparison to other similar prediction studies. Stott further points out that although the Mannheim and Wilkins' table may be useful as a predictor of Borstal success, it is of little use for prediction with the general population of juveniles.

The Mannheim and Wilkins' results have not been upheld in some later studies. For example, Hood (1965) applied the tables to a sample of 200 boys released from Borstal in 1953 and 1957: the predictive capacity of the tables was extremely poor when applied to this sample.

Despite these comments, the Mannheim and Wilkins research stands out in the field of criminological prediction studies as being one of the most thorough and systematic pieces of work.

Stott : The Bristol Social Adjustment Guides

Stott and Sykes (1956) developed a method for measuring the social adjustment of children. In this method the child's commonly occurring behaviours are described on a standard check list which is completed by his class teacher. This method of measuring maladjustment resulted in the development of the Bristol Social-Adjustment Guide (BSAG).

To examine the predictive utility of the BSAG, Stott (1959, 1960a, 1960b) compared the BSAG scores of a sample of 415 Glasgow boys who were on probation with those of a matched sample of 404 non-delinquent controls. Stott concluded that a weighted sum of 54 items selected from the BSAG was an effective predictor of delinquency. This weighted sum of items is described as the delinquency prediction instrument (DPI). The predictive power of the DPI, as described in the 1963 manual for the BSAG, appears to be impressive: the instrument divides the population into a series of groups which differ in their risk of delinquency from 4% to 100%.

While Stott's findings are promising there are several criticisms that can be levelled at the research. First, Stott's work is based on the cross-sectional comparison of selected samples of known delinquents and non-delinquents. There is no guarantee that such comparison will yield appropriate estimates of the predictive efficacy of the BSAG, as applied to the general population, for at least two reasons. As Simon (1971) has pointed out, the BSAG ratings for the delinquent group may have been contaminated because these children were known to be delinquents and hence their higher DPI scores may merely reflect bias in the BSAG ratings. Further, Stott's groups were selected and matched samples of the population; estimates based on such selected samples do not necessarily provide good risk estimates for the general population.

Marsh (1969) has criticised Stott for his failure to take account of the base rate problem. On this matter Marsh writes:

> " Recalling, however, that the present rate of court appearances is more like one in 10 than one in two, it becomes obvious that to give a proper perspective to these results the sizes of the above groups should be weighted in the ratio

of about nine non-delinquents to every delinquent ...." (p.280).

Marsh then reweights Stott's data using a Bayesian weighting procedure and allegedly demonstrates that the predictive utility of the BSAG is low. Marsh's criticism of Stott is slightly puzzling in that Stott is well aware of the base rate problem and in fact has criticised the Gluecks' research on these very grounds (Stott 1960c). Further, in the introduction to the DPI in the 1963 manual for the BSAG he states:

> " .... the incidences of maladjusted items found among the non-delinquent controls were multiplied some twenty times so as to make them proportionate to the boy-population as a whole. Thus the efficiency of the present prediction instrument is calculated on the assumption that potential delinquents have to be discovered from an unselected boy-population" (p.61).

This suggests that Stott has, in fact, made the adjustment to the base rate and that the risk estimates are given for a population in which the ratio of delinquents to non-delinquents is 1:20. If so Marsh's criticisms are without foundation. However, Stott does not make clear the way in which the base rate adjustment was carried out. One gains the impression that he is of the belief that simply multiplying the frequency of the DPI items for the non-delinquent group by a factor of 20 will automatically overcome the problem. This is not the case; such weighting will give the distribution of DPI scores for a population in which the ratio of delinquents is 1:20. However, unless the actual sample ratio of delinquents to non-delinquents is also weighted in this way, the risk estimates attached to each score range of this distribution will not be appropriate estimates of the risks for the general boy-population. Thus, if Stott has merely used the 1:20 ratio to weight his DPI scores without similarly adjusting his base frequency of delinquency, Marsh's criticism is justified.

Stott's approach to the prediction of delinquency offers the following advantages:

(1) The BSAG is a simple and standard measure of maladjustment which shows an acceptable level of reliability. In contrast, other studies have tended to use less standard or reliable measures. For example, the Gluecks' five factors have been repeatedly criticised on the grounds of their vagueness, unreliability and subjectivity.

(2) The rationale for using the BSAG as a predictor of future delinquency is fairly clear: one would expect that the degree of social adjustment displayed by a child at an early age would bear some close relationship to his subsequent behaviour and hence it is an eminently sensible idea to use such data to predict future delinquency.

However, before the BSAG can be used as a basis for delinquency prediction it is essential that further validation work on the DPI is conducted. In particular, it would seem necessary for the predictive power of the instrument to be assessed from the results of a longitudinal (prospective) study based on a random sample of the child population. This is the purpose of the present research.

West : Cambridge Study in Delinquent Development

An extensive study into the early concomitants of juvenile delinquency is at present being carried out under the guidance of Professor D.J. West. This study is known as the Cambridge Study in Delinquent Development. The sample for the research is 411 boys, selected from six London schools, who have been followed up from the age of eight years to the age of 18 years. The findings of the research have been presented in Present Conduct and Future Delinquency (West 1969) and Who Becomes Delinquent? (West and Farrington 1973).

The focus of the study is not on the development of formal statistical prediction devices but on the more general issue of the identification of the early symptoms and aetiology of delinquent behaviour. However, West and Farrington (1973) give some attention to the problem of predicting delinquency. These authors examined the relationship between a number of measures

taken at age eight years and subsequent juvenile offending. The results of this examination may be summarised as follows:

(1) Five variables measured at age eight years were found to be the best predictors of future delinquency. These variables were: (a) teachers ratings of the child's conduct; (b) family income; (c) social handicaps - measures of such things as poor housing, low income, etc; (d) acting out - a measure based on peer ratings, conduct ratings and neuroticism scores; (e) Troublesomeness - a measure based on the number of adverse ratings given to the boy by his teachers and his peers. The Troublesomeness measure is similar in design and content to the Stott DPI and in fact showed the highest degree of association with future delinquency.

(2) A series of background factors - (a) criminal offending by parents; (b) low family income; (c) large family size; (d) poor parental behaviour; (e) low intelligence - were also found to be associated with future delinquency.

West and Farrington then experimented with a number of scores created by combining the variables described above into unweighted sums. None of these combined scores was markedly better than the single measure of Troublesomeness. On the basis of this finding West and Farrington conclude ".... for the purpose of predicting delinquency, there is little point in measuring anything other than pre-delinquent behaviour" (p.131).

West and Farrington do not show how the risk of delinquency varies with the Troublesomeness index, although such data as are presented indicate that this index has a reasonable degree of predictive power. However, the measure still leaves a large degree of indeterminancy in the prediction of juvenile offending. A comparison is also made between the predictive power of the Gluecks' table and the results; the authors argue that when the Gluecks' work is adjusted for base rate effects the level of prediction achieved by the Gluecks' table is no greater than that evident for their results.

The work of West and Farrington is not subject to the criticisms that were levelled at the selected and matched samples used by Stott and the Gluecks. Further, the research design involves the longitudinal study of subjects and those factors identified as predictors were measured prior to the occurrence of delinquent behaviour. The research design thus does not suffer the deficiencies of a cross-sectional comparison. One criticism that can be levelled at the study is that the variables identified as predictors were selected from a number of potential predictors and it is possible that in this process, chance factors may have inflated the level of prediction achieved. Further, the relatively small sample size used in the study did not allow the authors to cross-validate their results. However, since the emphasis of the research is not on the development of formal statistical prediction devices these criticisms are scarcely justified.

Section 1.3 The Prediction Problem and Models

The discussion in the preceding section indicates, in an informal way, the general scope and nature of prediction studies. In this section we develop the idea of prediction in a more formal way.

Consider some group of N subjects measured on a set of m variables X1, X2 ..... Xm at some point in time t1. This set of variables is postulated to be predictive of some future outcome measured by a set of criterion variables Y1, Y2 ..... Yk. The sample is then measured on these criterion variables at time t2 subsequent to t1. To ensure that the values of the criterion variables are not contaminated by the effects of time, the interval t1 - t2 is a constant for all subjects.

The outcome of this procedure can be represented by two data matrices: the N x m matrix X of subjects measured on predictor variables and the N x k matrix Y of subjects measured on criterion variables. The aim of prediction research is to establish systematic relationships between the matrices X and Y, or selected subsets of these matrices, such that the score of any subject on an element of Y can be estimated from knowledge of his score distribution on X.

In most criminological research, Y is a N x 1 vector of subjects measured on one criterion variable and X is a matrix of N subjects measured on m predictor variables. It is usual to call the variable Y the criterion variable or the dependent variable; the elements of X are called predictor variables or independent variables.

We next consider some of the ways by which combinations of predictor variables can be constructed to predict values of a single criterion variable and the means by which the efficiency of such prediction can be evaluated. This discussion is not intended to be a thorough statistical analysis of prediction models; its purpose is to indicate the general features of various methods.

Perhaps the most straightforward approach to dealing with the prediction problem described above is to assume that the relationship between the criterion variable Y and the predictor variables X1, X2 .... Xm is linear. This assumption gives rise to a general data model of the form:

$$Y'j = \sum_{i=1}^{m} BiXij + Bo \quad \ldots\ldots\ldots\ldots \text{(Eq. 1.3.1);}$$

where Y'j is the estimated score of the jth subject on the criterion variable Y; Bi is the weight attached to the ith predictor variable Xi and Bo is some constant for all subjects. This model can be relaxed somewhat by allowing the relationship between the predictor variables to remain additive but not imposing the requirement of a linear relationship between criterion variables and the composite. In this case we have the model:

$$Y'j = f\left(\sum BiXij\right) \quad \ldots\ldots\ldots\ldots \text{(Eq. 1.3.2).}$$

Equation 1.3.2 asserts that the estimated score of the jth subject on the criterion variable is some (as yet unspecified) function of a sum of weighted predictor variables.

Both of these models have been used in prediction studies. Their most sophisticated application is multiple linear regression. This method provides an explicit analytic solution to obtaining the weights B1, B2 .... Bm in equation 1.3.1. This is done through the minimisation of the sum of the squared deviations of the estimated scores Y'j around the observed scores Yj. Given the constraints of a linear relationship between the predictor and criterion variables, multiple linear regression is the most efficient means of obtaining estimates of the criterion values. Further, as we show below, some of the methods that have been used for constructing prediction rules are in fact weaker versions of this model.

A system which is frequently used for constructing prediction scores, when the predictor variables are in dichotomous form, is the "points" or Burgess system of scoring (Simon 1971). In this system, a composite score of the sum of the values of the dichotomous (0,1) variables is constructed.

This composite may be used in one of two ways. The first, and by far the most common, method is to arrange the scores in a series of class intervals and to tabulate these class intervals against the criterion variable. This results in an additive prediction model in which the functional relationship between the predictors and the criterion is specified by a table. While this approach is statistically unsophisticated it offers the advantages of being robust and simple to apply in practice. Further, it makes no assumptions about the mathematical relationship between the predictors and the criterion, which is simply specified empirically.

An alternative way of using a points score is to assume that the relationship between the score and the criterion is linear. This leads to the model:

$$Y'j = B \sum Xij + C \quad \ldots\ldots\ldots\ldots\ldots \text{(Eq. 1.3.3)};$$

where B and C are constants estimated to minimise the sum of squares of Y'j around the observed criterion values. This equation is merely a special case of equation 1.3.1, in which the weights are all set equal to each other. However, it will be recalled that the multiple linear regression model produces weights which minimise the amount of predictive error (i.e. the sum of squared deviations between Y'j and the observed values) and as a consequence it follows that the model in equation 1.3.3 will produce results that are no better than those given by multiple regression but it may produce results which are worse.

The points system can only be used with any degree of confidence when all predictor variables are scored on the same scale. When predictor variables are scored on different scales, the points system may result in some variables being given undue weight. One way in which this problem can be overcome is to reduce all predictor variables to a common scale by normalising these variables. The normalised variables may then be used to form a composite score in the same way in which the points score is constructed. If a linear relationship between predictor and criterion variables is assumed it can be shown that this normalised score will have similar properties to the points score: the resulting prediction equation will be no more efficient than the corresponding multiple regression equation but it may be less efficient.

While the points score and multiple linear regression are the most common means of constructing additive or linear prediction systems some writers have used more unorthodox methods. The Gluecks (1950) and Stott (1960c) have constructed prediction scores by weighting the values of predictor variables by the relative frequency of occurrence of the variables amongst delinquent and non-delinquent boys. Doubtless the idea here is to weight the predictor variables in accordance with their predictive contribution. There is, however, no guarantee that such weighting will achieve this. In particular, when the variables are highly correlated the weighting system may assign large weights to a series of predictor variables whose predictive power collectively is no greater than the power of the best single variable. The variables thus weighted will be given undue importance in the prediction equation. Further, if a linear relationship between the predictor variables and the criterion is assumed, then it can be shown that the weights obtained by this method will do no better and may do worse than the weights obtained by multiple linear regression.

In summary, if linearity is assumed, multiple regression gives optimal prediction.

However, there are a number of factors which reduce the apparent theoretical superiority of multiple regression. In general, when a sample of observations for a prediction equation is fitted, the process of estimating the parameters or constants for the equation tends to capitalise on chance variation in the data. This results in a situation in which the prediction equation works better for the sample of observations on which it was constructed than for other samples. This effect is related to the number of parameters in the prediction equation: the more parameters estimated the greater will be the shrinkage in the equation.

It can be seen from the above that the multiple regression model has the most parameters and hence is more likely to produce shrinkage effects. On the other hand, the other approaches described above involve the estimation of few parameters and hence are relatively robust and resistant to shrinkage. This results in the theoretical superiority of the multiple regression method being reduced in the practical situation. For example,

Simon (1971), after reviewing the results of a series of prediction methods applied to the prediction of probation success, concludes that the simple points system gives results as good as more sophisticated methods of prediction.

A somewhat different use of linear prediction models is linear discriminant function analysis. This model is, conventionally, applied in the case where the criterion variable Y is a series of categories rather than a continuous variable or an approximation to a continuous variable. In discriminant function analysis a weighted composite W of the predictor variables is constructed i.e.:

$$W = B1X1 + B2X2 \ldots\ldots\ldots\ldots Bm\ Xm \quad \text{(Eq. 1.3.4)}.$$

The weights B1, B2 ..... Bm are selected so as to maximise the differences between the mean values of W for the k groups of the criterion variable. The mean score then may be used for assigning subjects to various groups. When the criterion variable is dichotomous the linear discriminant function model reduces to a multiple regression analysis in which the criterion variable assumes the values of either 0 or 1 (Tatsuoka 1971).

A number of objections can be raised to the use of linear or additive models. Often there are no sound theoretical reasons for assuming that the criterion variable is linearly related to the predictors or that an additive combination of predictor variables will necessarily produce the optimum prediction model. In short, additive and linear prediction models are not necessarily appropriate for all prediction problems.[1] While such an argument can be sustained on theoretical grounds, in practical terms it does not seem to matter which prediction method is applied in criminological research: most methods appear to produce about the same general level of prediction (cf. Simon 1971; Schumacher 1974; Challinger 1974).

---

1. A point that should be made here is that the use of a linear model does not necessarily entail the assumption that the relationship between the criterion and the predictors is linear. It is possible to introduce curvilinear relationships into linear models by applying various transformations on the predictor variables (cf. Ezekiel and Fox 1966).

A more serious problem in the use of linear and additive models is that of interpreting the results of the analysis. If all that is desired is some form of "black box" prediction system this problem is not great. However, if one wishes to place theoretical interpretations on the findings concerning the variables which are the most important or significant predictors, the models described above may be quite difficult to interpret. This is particularly the case with the points system in which the variables are simply added up without weights. The most that can be said of this data model is that it implies that the more adverse conditions the individual suffers the greater is his likelihood of offending; there is no way of singling out the contributions of individual variables or their relative importance.

At first sight, multiple linear regression would appear to overcome this problem: one might expect that the size of the coefficients attached to the predictor variables provides some indication of the importance of the contribution of the variables. However, there are two reasons why this interpretation is not correct. The first is that in the multiple regression equation, the size of the weight attached to a variable reflects the scale on which the variable is measured as much as its predictive contribution: variables measured on scales with small absolute units will, ceteris paribus, receive greater coefficients than those measured on scales with large units. The way of overcoming this problem is simply to transform the regression equation into normalised form thus placing all measurements on a common scale. However, even after such a transformation has been made the coefficients in the equation still do not necessarily reflect the importance of the contribution of the variables. This is because the size of the weights is determined to a considerable extent by the pattern of intercorrelations between predictor variables. In particular, a group of highly correlated predictor variables may all receive low weights in the equation, even though each of them correlates substantially with the criterion variable. Although there is no entirely satisfactory means of assessing the contribution of a particular variable to a multiple regression equation, there are a number of techniques which are customarily used for assessing the importance of variables in the

equation.[1] These techniques are reviewed in an article by Darlington (1970) to which the interested reader is referred.

The difficulties associated with linear prediction models have led to a number of attempts to develop non-linear prediction models suitable for handling predictor variables in categorical form. Two such models have been applied to criminological data: MacNaughton-Smith's (1963) Predictive Attribute Analysis (PAA) and Sonquist and Morgan's (1964) Automatic Detection of Interaction Effects (AID). Both models work on the same principle: the sample of observations is sequentially split into a series of binary partitions defined on the predictor variables so that the within groups variability of the criterion variable Y becomes smaller.

AID requires that the criterion variable Y is measured on at least an interval scale with the minimum condition that Y can be expressed in dichotomous (0,1) form. The predictor variables X1, X2 ..... Xm may be on nominal, ordinal, interval or ratio scales. The requirements for PAA are more constrained: all variables, including the criterion, must be expressed in dichotomous form. However, it can be shown that PAA is merely a special case of the more general AID model. This proof is given in Appendix 1 to this paper. Thus, to display the logic of both methods we will simply describe the general basis of AID.

First consider the total sum of squares around the mean of a specified criterion variable Y for a sample of N observations:

$$TSS_t = \sum_{j=1}^{N} Y_j^2 - \frac{\left(\sum_{j=1}^{N} Y_j\right)^2}{N} \quad \text{..... (Eq. 1.3.5).}$$

---

1. Perhaps the best means of making such an assessment is to use a technique known as causal path analysis (Blalock 1971). This procedure examines both the direct and indirect contributions of variables to a given outcome.

The aim of AID is to partition these observations into a series of subgroups such that the total within groups sum of squares of Y is minimised. The procedure works as follows:

(1) The sample is divided into two groups by dichotomising a selected predictor variable so that this partition produces two groups with the property that the total within groups sum of squares of Y is minimised.

(2) The two groups so formed are subject to the same procedure and the process continued until certain "stopping rules" are satisfied.

The algorithm for locating the best partition for any set of predictor variables involves finding the maximum value of the statistic:

$$BSSikp = TSSi - (TSS1 + TSS2) \quad \ldots\ldots \text{(Eq. 1.3.6)};$$

where TSSi is the total within groups sum of squares of the group being partitioned and (TSS1 + TSS2) is the total within groups sum of squares for a partition of the sample at cutting point p on the kth predictor.

The statistic BSSikp is a measure of the absolute reduction in the total sum of squares that is achieved by a given partition on a predictor variable, hence finding the condition which maximises this statistic minimises the within subgroups variability of the criterion. When this procedure is applied successively the sample is partitioned into a dendrogram, or tree, of binary partitions with the terminal groups of this tree - representing sets of conditions which minimise the variability of the criterion.

In theory, if partitioning can be carried on indefinitely with a sufficient number of effective predictor variables, each of the terminal groups of the AID tree would be associated with a single value of the criterion variable Y. However, in practice it is neither possible nor desirable to carry out partitioning to this extent and the AID tree is terminated by a series of "stopping rules" which specify the conditions under which any partition is permissible. These stopping rules have no particular statistical justification. Their intent is to prevent the partition of groups having a negligible amount of variability;

to ensure that groups having a small number of observations are not partitioned; and to ensure that each partition reduces the variability of the criterion by an appreciable amount. Sonquist, Baker and Morgan (1971) recommend that the partitioning process should stop when at least one of the following conditions is met:

(1) The reduction in the total sum of squares if a split occurred would be less than 0.6% of the original total sum of squares around the mean.

(2) If a split were made on a group, one or both of the subgroups formed would contain fewer than 25 cases.

(3) Some maximum number of splits (25) has already been made.

It is important to recognise that the predictive power of any particular AID tree tends to be an over-estimate owing to the fact that at each stage of the analysis the method finds the partition which, for the particular sample, minimises the within groups variability of the criterion variable. Because of this the AID tree tends to have greater predictive power for the sample of observations on which it was constructed than for other samples.

Section 1.4 The Need for Validation

A point which has been mentioned in passing in the previous sections is that in the construction of a prediction instrument there is a tendency for the prediction method to capitalise on chance variation in the data and hence provide overly optimistic estimates of predictive power. The prediction rule is over-fitted to the sample of observations and will shrink, or lose predictive power, when applied to a fresh sample of observations. Two sources of shrinkage are possible for any set of data:

(1) The first arises from the estimation of parameters from the sample data. To the extent to which such estimates are subject to sample error and variation they tend to maximise predictive power.

(2) The second source of shrinkage is more elusive. In the process of constructing a prediction instrument, selection is usually made amongst a number of potentially predictive items. This process results in certain variables being identified as predictors by chance. Further, scoring procedures for variables may be selected to maximise prediction (Simon 1971). Thus, in constructing a prediction instrument, the investigator is often carrying out a series of procedures which maximise the likelihood that he will select spurious predictors.

The presence of shrinkage on prediction instruments is something which must always be taken into account and unvalidated instruments run the risk of misleading rather than helping their user. The conventional procedure to overcome the problem of shrinkage, and thence obtain unbiased estimates of predictive power, is to randomly partition a sample of observations into two groups of equal size. The first set of observations is used to construct the prediction rule and is called the construction sample. The second group of observations is used to test the prediction rule and is called the validation sample. The statistics for the validation sample give unbiased estimates of the predictive power of the instrument.

Section 1.5 The Base Rate Problem

A recurrent problem in prediction studies is that the incidence of detected juvenile offending is low. The low base rate of offending poses two problems for prediction research:

(1) If the data are collected using a simple random sample, the sample size has to be very large to ensure that a sufficient number of delinquent subjects are obtained.

(2) With a low base rate of offending, the variability in the criterion variables is small as most subjects have committed no offences. The limited variability of criterion variables makes it extremely difficult to find effective predictor variables. On this matter Simon (1971), quoting Gottfredson (1967), comments that the limited variance of the criterion reduces predictability as it is this variance that must be analysed in the search for effective predictors. This effect is also known in other areas of the behavioural sciences. For example, Magnusson (1967) shows formally how the concurrent validity of a test can be reduced by limiting the variability of the criterion variable. The effect may also be seen in considering the way in which the point-biserial correlation coefficient varies with the base rate. This coefficient is frequently used to assess predictive power when the criterion variable is dichotomous. A formula for the point-biserial is:

$$r_{pbis} = \frac{(Mp - Mq)\sqrt{pq}}{Sy} \quad \ldots\ldots\ldots \text{(Eq. 1.5.1);}$$

where Mp is the mean score of the group of successes on some test Y, Mq is the mean score of the group of failures, p is the proportion of successes, q is the proportion of failures, and Sy is the standard deviation of the test. It can be seen that the point-biserial reaches its maximum value when $p = q = .5$; as the base rate approaches either 0 or 1 the point-biserial tends to 0. This indicates that the most favourable situation for prediction occurs when half the sample are successes and half are failures, and

shows that as the base rate of offending becomes small the chances of finding effective predictors also become increasingly small.

Simon (1971) has suggested that these problems can be overcome by the use of a stratified sampling scheme in which half the sample are delinquent and the other half are non-delinquent. This scheme offers the advantages of giving maximum sensitivity to the predictors and of reducing the size of sample required for analysis. Despite the attractive features of the design, it has one major drawback: all inferences and statistics based on the design apply to an artificial population which contains 50% delinquents and 50% non-delinquents; the results do not apply to a population in which the base rate of offending is (say) 10%. An extremely lucid article by Meehl and Rosen (1955) outlines the liabilities of such a design if carelessly employed and shows how, by the use of Bayes theorem, estimates can be adjusted for a different base rate. However, while it is fairly easy to adjust prediction tables for base rate effects, the problems of translating an entire study based on a 50% base rate to another base rate are more complex. If the research is to be of maximum value, one would like to obtain estimates of all statistics reported as they apply to the general population. This poses quite knotty problems in transforming correlation coefficients and, more particularly, significance levels. Further, it is not altogether clear whether a prediction system which is optimal for a population in which the base rate is 50% is optimal for a population in which the base rate is (say) 10% as to some extent selection of predictors may be influenced by the distributional properties of the variables.

As the previous discussion implies, failure to take account of the base rate problem has been one of the most persistent errors in criminological research and is one which is still being perpetrated (see, for example, LaBrie 1972). The basis of this error is the making of unjustified inferences from stratified samples. In general, it would seem that the use of a simple random sampling scheme overcomes the problem in the most direct

fashion and avoids the possibility of erroneous inferences being made due to inadequate consideration of the complexities of the base rate effect. However, while a simple random sampling scheme considerably simplifies the problems of inference, it tends to be expensive in the data collection phase of research.

Section 1.6 The Measurement of Predictive Power

Once a prediction instrument has been constructed and validated the next problem is to determine the extent to which it is effective as a predictor. Broadly speaking, the measurement of predictive power involves the determination of the degree to which the instrument accurately predicts the scores of the subjects on the criterion variable for the validation sample.

Prediction instruments may be expressed in one of two forms:

(1) As a prediction equation which gives, for each subject, an estimated score on the criterion variable.

(2) As a table which partitions the sample into a series of classes; to each class there is attached some estimate of the likely score, on the criterion variable, of any member selected at random.

These two methods of presentation are not mutually exclusive and often it is possible to present a prediction equation as a prediction table and vice versa. Most commonly the results of prediction studies are presented as tables. Appropriate measures of predictive power for both situations are discussed below.

Variance Measures

A measure of predictive power that is frequently used is the amount of variance in the criterion variable that can be accounted for by the prediction rule. In the general case, variance prediction measures take the form:

$$\frac{\text{Amount of variation in criterion accounted for by rule}}{\text{Total variation in the criterion variable}}$$

This general form leads to a variety of statistics for measuring predictive power. When the prediction rule is in the form of a score which is assumed to be linearly related to the criterion variable the appropriate measure of predictive power is the square of the product moment correlation coefficient between the criterion values and the score values.

More commonly, the prediction rule is laid out empirically by partitioning the sample into a series of groups G1, G2 ..... Gk defined on predictor variable scores. To each class Gi there is attached some estimate of the likely score of the subject on the criterion. For a set of validation data such a table may be considered to be a stratified sample with each stratum, i.e. group, having a within groups score distribution. An alternative way of looking at the table is as a one-way analysis of variance table with k groups measured on a dependent variable Y. The problem here is to find some means by which to assess the extent to which the partitioning procedure reduces the within groups variability of the criterion variable. The appropriate measure is the correlation ratio - eta - which is defined as follows:[1]

$$eta^2 = \frac{TSSt - \sum_{i=1}^{k} TSSi}{TSSt} \quad \ldots\ldots\ldots\ldots \text{(Eq. 1.6.1)};$$

where TSSt is the total sum of squares around the mean of the unpartitioned sample and $\sum TSSi$ is the total within groups sum of squares for the partitioned sample. The reasoning behind this index is fairly obvious. The difference between the total sum of squares TSSt and the total within groups sum of squares $\sum TSSi$ represents that portion of the variation of the criterion variable that has been accounted for by the partitioning process and hence the ratio of this difference to the statistic TSSt is a measure of the proportionate reduction in variance achieved by the partitioning.

The statistic eta^2 has cropped up in a variety of guises as a measure of predictive power. For example, Sonquist and Morgan (1964) have defined the statistic R as a measure of prediction where:

$$R^2 = \frac{BSSt}{TSSt} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(Eq. 1.6.2)};$$

and BSSt is the total between groups sum of squares for the k terminal groups of an AID tree. Clearly, equations 1.6.1 and

---

1. There are a number of measures logically similar to eta which give measures of strength of effect. Hays (1963) describes these measures.

1.6.2 are identical. Further, Simon (1971) has proposed the use of the statistic ϕ^2 to evaluate predictive power in the case where the criterion variable is in dichotomous form. The formula for ϕ^2 is:

$$\phi^2 = \chi^2/N \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(Eq. 1.6.3)};$$

where χ^2 is the Pearson chi square value for a 2 x k risk table. It can be shown that when Y is in dichotomous form, eta^2 becomes ϕ^2.

However, while the correlation ratio eta is a fairly general measure of prediction or association for any set of sample data, it can be shown that as an estimator of the population correlation ratio it is biased (Peters and Van Voorhis 1940). The bias comes from the fact that the estimate of the within groups sum of squares ΣTSSi is based on k groups, whereas the estimate of the total sum of squares TSSt is based on a single group. There is thus a need to adjust the estimate to take account of the varying number of degrees of freedom used to estimate the total and within groups sums of squares. The unbiased estimator of the population correlation ratio is the statistic $epsilon^2$ discussed by Peters and Van Voorhis (1940). This statistic is defined as follows:

$$epsilon^2 = 1 - \frac{N - 1 \ (\Sigma \text{TSSi})}{N - k \ \ (\text{TSSt})} \quad \ldots\ldots \text{(Eq. 1.6.4)}.$$

However, in practical terms, the difference between eta^2 and $epsilon^2$ is negligible since N, the number of observations, is normally large and k, the number of groups, is normally small.

The correlation ratio has the disadvantage that it takes no account of the way in which the groups in the prediction table are laid out. When the table is formed from a series of discrete classes based on no underlying metric this is not a problem. If, however, the classes in the table are based on at least an ordinal measure, the correlation ratio may give a misleading impression of the relationship of this measure to the criterion variable as it takes account of all between groups variation in criterion scores.

Variance measures of predictive efficacy are useful summary statistics for describing the overall properties of any prediction rule. However, if they are used as the sole measure of predictive efficacy they may be quite misleading. This is because the purpose of a prediction instrument is prediction, not the reduction of variance. The two terms are not quite synonymous as can be seen from the table below which shows hypothetical data for predicting the risk of juvenile delinquency for six groups of children.

	G1	G2	G3	G4	G5	G6	Total
Risk of Offending	.30	.35	.45	.50	.55	.98	.48
N	100	100	100	100	100	50	550

The value of eta^2 for the above table is 0.134: a figure which might lead one to believe that the prediction table is of little value. However, further examination of the table reveals that its efficiency differs according to circumstances: for group 6 the table is an extremely efficient predictor as any child belonging to this category is almost certain to become a delinquent; for the other groups the efficiency of the table is poor. These distinctions are entirely glossed over by the variance measure statistics which are concerned with the overall performance of the prediction table not its utility in given circumstances.

A related problem with variance measures is that of translating them into intuitively meaningful terms. While the statement "the prediction method was able to account for 40% of the variation in the criterion" indicates that the method displayed some predictive power, it says little about the liabilities and advantages of the prediction method. In short, variance measures of predictive power are global measures which must be supplemented by more detailed information if a thorough evaluation of predictive efficacy is to be obtained.

Measures for a dichotomous criterion variable

Frequently, criminological research uses a simple criterion of success or failure and as a result emphasis has been placed on the development of indices of predictive power for dichotomous criterion variables. A brief review of these measures is given below:

(1) Predictive efficiency:- The most obvious and simple means of measuring prediction is to tally up the number of correct predictions. However, such information says very little about predictive power unless one compares it with the number of correct predictions that would be obtained by using the base rate information alone. Clearly, if the chance of success is 90%, then one can make a 90% correct prediction by simply predicting that everyone will be a success. On this basis an instrument which gives an 85% correct prediction is ineffective. These ideas underlie the development of the index of predictive efficiency (PE) devised by Ohlin and Duncan (1949).

A formula for PE is:

$$PE = \frac{\text{Number of misclassifications using instrument alone}}{\text{Number of misclassifications using base rate alone}}$$

While PE is intuitively appealing as a measure of prediction it has a number of liabilities. Simon (1971) suggests that it is susceptible to influence by the base rate: when the base rate is high the possibility of obtaining a high PE is limited. Further, the index does not adequately summarise the degree of separation between groups.

(2) Range and Selectivity:- A natural extension of the idea that a measure of predictive power may be based on the rate of misclassification is to consider the ways in which misclassifications are distributed. These considerations led Stott (1960c) to suggest that predictive efficiency could be better evaluated in terms of two measures which he describes as range and selectivity. Range is defined as the proportion of delinquents who are accurately classified as delinquent;

selectivity is the proportion of those classified as delinquent who are in fact delinquent. These concepts are sound and draw on the idea that decision errors vary both in their direction and their importance. This idea is elaborated formally in the Theory of Signal Detectability (TSD) and later in this chapter the way in which Stott's concepts may be subsumed under this theory will be shown.

(3) Chi square and phi:- A frequent but not entirely justifiable practice is to express the degree of prediction obtained by computing the Pearson chi square statistic for 2 x k risk table. This statistic is not an appropriate measure since it is concerned with testing the degree to which the within groups risk distribution is different from the overall base rate. The value of this statistic is to a considerable extent dependent on the sample size rather than upon the degree of prediction. This point is, of course, a particular instance of the more general distinction between size of effect and statistical significance (see for example Hays 1963, p.300).

Simon (1971) has suggested that when the criterion variable is dichotomous a useful index of predictive power is the point-biserial correlation between the within groups risk values and the dichotomous criterion values. She shows this coefficient to be equal to ϕ. As we have suggested earlier ϕ is also the correlation ratio computed for a dichotomous criterion variable.

(4) MCR:- Duncan et al (1953) have proposed a measure of statistical association which they describe as the mean cost rating (MCR).

A formula for the MCR is:

$$MCR = \sum_{i=1}^{k} C_i \, U_{i-1} - \sum_{i=1}^{k} C_{i-1} \, U_i \qquad \text{(Glaser 1955);}$$

where:

k = number of score classes or risk groups in the table arranged in order of decreasing risk

i = score class above which all cases are classified as failures

C_i = the proportion of successes who are incorrectly classified by cutting the table at score class i

U_i = the total proportion of failures who are correctly classified by cutting the table above score class i.

Simon (1971) argues that the MCR is an extremely useful measure of predictability in that it is not influenced by the base rate; it is sensitive to the order in which the risk table is laid out, and it involves no assumptions of normality, continuity or equality of score units. While these advantages must be admitted, a problem the authors have found is that of placing an interpretation on the MCR; a value of the MCR leads to no intuitively obvious account of the degree of prediction. Later we will show that the MCR in fact bears a systematic relationship to TSD statistics and can be most easily interpreted in this context.

Other indices

The measurement of predictive power is a special case of the more general problems of assessing goodness of fit and/or the degree of association between variables. At present a bewildering variety of indices designed to produce such measures are available.

Most of these measures are designed for use with ordinal or nominal data. To present a review of all these methods would be an almost impossible and confusing task. One might also observe that although many of these measures serve their purpose in particular applications, for the practical assessment of predictive power the presentation of a prediction table and limited use of statistics would appear to be just as efficient. Further we would argue that the development of indices of predictive power in isolation is a sterile practice and that what is required is the development of a systematic body of theory which will relate measures of prediction to decisions and the consequences of decisions. In the next section we will show how TSD fulfills these conditions.

The Theory of Signal Detectability

Most of the measures of predictive power discussed previously have relied on the use of a single summary statistic. The complexities of prediction systems are unlikely to be well represented in this way. A more systematic and comprehensive attack on the problem of measuring predictive power is offered by the Theory of Signal Detectability (TSD) (Green and Swets 1966). TSD is part of general information theory and was developed mainly in the context of electrical engineering to handle the problem of decision making from noisy or uncertain information sources. This is exactly the problem faced by the user of a statistical prediction device: perfect classification of individuals is not possible and one is in the position of making a decision which maximises the use of the available information. To do this the decision maker requires a strategy or decision rule which optimises the use of the available information. The way in which TSD handles this problem is discussed below.[1]

Consider a 2 x k risk table comprising k categories G1, G2 ..... Gk. Associated with each group there is a conditional probability P(Gi/S) that a subject who is a success is a member of Gi and a probability P(Gi/F) that a subject who is a failure is a member of Gi. The terms success and failure are used neutrally to denote the two (mutually exclusive and exhaustive) states of a dichotomous criterion variable; in the present context they may be interpreted as non-delinquent and delinquent. The overall probabilities of success and failure for the table are denoted P(S) and P(F). Conventionally, the probabilities P(Gi/S), P(Gi/F), P(S), P(F) are described as prior probabilities and the ratio P(F)/P(S) is known as the prior odds. The problem is to specify a decision rule based on the prior probabilities such that the outcome of this decision rule is optimum in some sense. A useful statistic on which to base decision rules is the likelihood ratio:

$$L(Gi) = \frac{P(Gi/S)}{P(Gi/F)}$$

---

1. There is a strong similarity between the application of TSD discussed here and the utility theory treatment of the assessment of predictive power presented by Duncan et al (1953).

This statistic is a measure of the likelihood that a member of Gi is a success. The likelihood ratio has two important properties:

(1) It is monotone with the posterior probabilities of success, P(S/Gi), and failure, P(F/Gi), associated with each group.

(2) It can be shown that if L(Gi) is greater than the prior odds then the members of Gi are more likely to be successes than failures (Coombs, Dawes and Tversky 1970).

These properties make the likelihood ratio a useful statistic for decision making and one which allows movement between the prior and posterior probabilities. The most obvious decision rule to formulate using the likelihood ratio is to classify all groups with values of L(Gi) greater than the prior odds as successes, and all other groups as failures. This procedure maximises the number of correct classifications made. However, such a decision rule does not take into account the fact that the costs of various decisions may vary. (For example, the consequence of a doctor classifying a patient as dead when he is alive is not the same as that of the patient being classified as alive when in fact he is dead). To meet this eventuality TSD introduces the idea of the pay off matrix:

PAY OFF MATRIX

		Predicted	
		Success(S')	Failure (F')
Actual	Success(S)	V11	- V12
	Failure(F)	- V21	V22

In the pay off matrix the predicted outcome is compared with the actual state of nature. There are two ways in which correct decisions can be made: the subject can be predicted to be a success and turn out to be a success or he may be predicted to be a failure and turn out to be a failure. TSD describes the first event as a hit and the second as a correct rejection. Similarly, there are two ways in which incorrect classifications

may occur: the subject can be predicted to be a success and turn out to be a failure or he may be predicted to be a failure and turn out to be a success. The first event is called a false alarm and the second a miss. Each outcome in the pay off matrix is associated with a cost Vij (i=1,2; j=1,2). As hits and correct rejections are correct decisions they receive positive values; misses and false alarms receive negative values.

The set of L(Gi)s and the elements of the pay off matrix provide the essential ingredients for formulating optimum decision rules. Most often the optimum decision rule is to maximise the expected pay off of the decision process.

The expected value of calling all subjects in group Gi successes is:

$$E(S'/Gi) = V11\ P(S/Gi) - V21\ P(F/Gi);$$

and the expected value of calling subjects in Gi failures is:

$$E(F'/Gi) = V22\ P(F/Gi) - V12\ P(S/Gi).$$

To maximise the expected pay off from the decision process we therefore call all subjects in Gi successes if and only if:

$$E(S'/Gi) > E(F'/Gi).$$

This decision rule can be shown to be equivalent to calling all subjects in Gi successes if and only if:

$$L(Gi) > \frac{P(F)(V22 + V21)}{P(S)(V11 + V12)}$$

This decision rule has the useful property that it is invariant over transformations of both the probability scale and the value scale i.e. the likelihood ratio criterion will remain invariant irrespective of the actual units in which costs are measured or the form in which the probabilities are specified (cf. Green and Swets 1966, p.23).

A recurrent problem in criminological prediction research is that while the set of L(Gi)s can be estimated from existing actuarial data, the elements of the pay off matrix remain ill-defined, if not non-existent. In this situation it is difficult

to decide on the procedure for optimising the pay off from prediction. There would seem to be two possible approaches to the problem. The first is to adopt arbitrary values for the elements of the pay off matrix and then to locate the decision rule which maximises the pay off. This strategy is perhaps the one most frequently employed in prediction research where investigators seek to locate a classification rule which maximises the number of correct predictions. This rule can be shown to be equivalent to assuming that the cost of a false alarm is equal to the cost of a miss. However, this is only one solution that can be applied to the incomplete information available for formulating decision rules. An alternative way of attacking the problem is to work backwards, as it were, from the cutting rule to the pay off matrix. One may observe that although it is not possible to specify a matrix of pay off values for a given prediction instrument, frequently users of such systems are in a position to specify which prediction rules are acceptable or not acceptable. Further, using the likelihood ratio criterion, there are only k + 1 decision rules which can be formulated of which one must be chosen (or the idea of prediction forgotten entirely). Thus from a presentation of the probability structure and operating characteristics of the prediction instrument, the user should be in a position to specify which rule, if any, is acceptable to him. In this instance, the prediction instrument can be viewed as an actuarial device to which the user applies his own subjective pay off matrix to reach the optimum decision rule. While this situation is not completely desirable in that TSD assumes that the elements of the pay off matrix are computed independently of the values of the likelihood ratio, it probably represents a better solution to the problem than the use of the arbitrary assumption that all kinds of error are of equal value.[1]

If this view is accepted the essential information to be presented about a prediction instrument is a summary statement of the properties of the instrument over all decision rules that can

---

1. One might also observe that such a posteriori selection of a cutting rule runs some risk of over-fitting the prediction in that the decision rule is chosen on the basis of a fallible set of probability estimates and thus is prone to capitalise on chance variation. Under ideal circumstances the cutting rule should be selected on one sample and validated on another.

be formulated. TSD provides an extremely succinct and useful method of presenting such a summary. This method is called the Receiver Operating Characteristic (ROC) curve and is presented as a table or graph which shows the consequences of decision rules. The basis of the ROC curve can be described as follows:

Consider the 2 x k risk table described earlier, laid out so that the k groups are arranged in ascending order of the value of L(Gi). Using this table there are k + 1 decision rules that can be formulated. These rules correspond to the sequence: call all subjects successes, call all subjects successes save those in the group with the lowest value of L(Gi), ... call all subjects failures.

The properties of these decision rules can be summarised by two statistics: the hit rate, P(S'/S), the probability that a subject who is a success will be predicted as a success; and the false alarm rate, P(S'/F), the probability that a subject who is a failure will be classified as a success. (It is easy to see that the miss and correct rejection rates, P(F'/S) and P(F'/F), are merely complements of these statistics).

Thus if the hit and false alarm rates are plotted against each other for each decision rule that can be formulated, the resulting curve describes the consequences of all decision rules. From this curve and information on the prior probabilities, P(S) and P(F), it is possible to generate a complete set of summary statistics for each decision rule. Some of the statistics that may be derived[1] are as follows:

(1) The proportion of delinquents correctly identified (i.e. the hit rate).

(2) The proportion of non-delinquents correctly identified (i.e. 1 - false alarm rate).

(3) The proportion of delinquents amongst those classified as delinquent. We will describe this statistic as the detection rate.

---

1. These statistics can all be derived using fairly simple applications of Bayes Rule to the hit and false alarm rates and the prior probabilities of success and failure. The derivations of the statistics are not shown here as they involve a rather tedious repetition of simple formulae.

(4) The proportion of non-delinquents among those classified as non-delinquents. We will describe this statistic as the rejection rate.

(5) The proportion of correct classifications resulting from the decision rule.

These statistics form a sufficient basis for describing the consequences of any decision rule: they indicate what proportion of delinquents will be correctly classified; what proportion of non-delinquents will be correctly classified; what proportion of those classified as delinquent are in fact delinquent; what proportion of those classified as non-delinquent are in fact non-delinquent; and the overall proportion of correct classifications.

In addition to describing the consequences of decision rules, the ROC curve can also be used to generate a number of indices of predictive power. Most of these indices involve rather restrictive assumptions concerning the distribution of the criterion. Perhaps the most useful measure for many applications is the non-parametric statistic P(A): the area under the ROC curve. This area is shown in Figure 1.6.1 below which shows a hypothetical ROC curve. In addition, the figure also shows the chance line: i.e. the plot of hit and false alarm rates that would emerge if children were classified as delinquents and non-delinquents at random using various sampling fractions. It will be observed that the ROC curve is contained in a square of unit area and because of this P(A) has the interpretation of being the proportion of this unit square which falls below the ROC curve. Under normal circumstances therefore P(A) varies from .5, for the case in which prediction is no better than chance, to 1 for the case in which prediction is perfect.

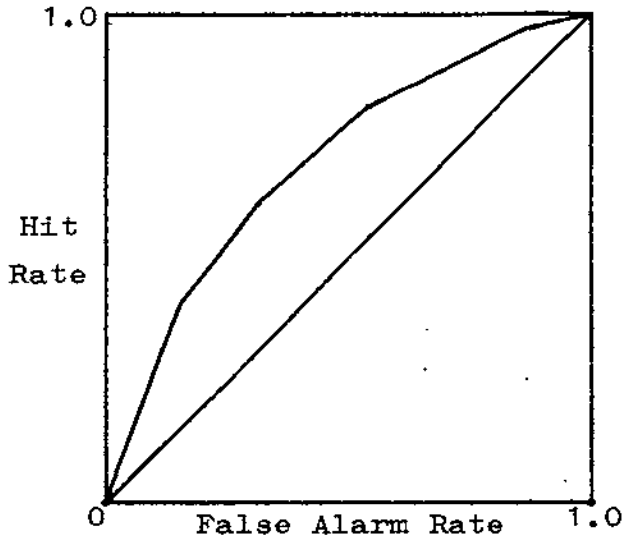


Figure 1.6.1 ROC CURVE AND CHANCE LINE

P(A) has a number of useful properties. First, like the MCR to which it is closely related, P(A) involves no assumptions concerning the equality, continuity or normality of predictor scales. Second, it is invariant under transformations of the base rate. Some comment on this feature is in order. Measures of predictive power may be classified into two groups: those like ϕ, eta or PE which are dependent on the base rate and those like P(A) or MCR which are independent of the base rate. The information conveyed by these measures would appear to be quite different. Base rate dependent statistics describe the predictive properties of an instrument when applied to a given situation with a given base rate. Base rate independent statistics describe the predictive properties of an instrument in a more abstract way and are not related to any particular base rate situation. Both measures serve different purposes. For the practical assessment of predictive power, base rate dependent statistics seem to be the most suitable as they describe the predictive capacity of the instrument as it applies to a particular situation. For theoretical purposes, base rate independent statistics seem to be more appropriate as they describe the predictive potential of the instrument irrespective of the limitations on this potential that are imposed by various base rate situations.

A theorem devised by Green and Swets (1966) makes it possible to place a relatively simple intuitive interpretation on any value of P(A). These authors have demonstrated that P(A) is in fact identical to the expected number of correct classifications that would arise from a two-alternative forced-choice experiment. This result may be explained as follows.

Imagine that every delinquent was paired at random with a non-delinquent and that for each such pair an observer was required to say which child was the delinquent and which child was the non-delinquent. If the observer had no information about the children he would respond more or less at random and achieve an expected rate of correct classification of 50%. Suppose, however, he had access to a test score about the child and he knew that delinquents were prone to receive higher scores than non-delinquents. He could therefore improve his prediction by classifying the child with the

higher score in each pair as a delinquent and the other child as a non-delinquent. Under this strategy, the expected proportion of correct classifications would be equal to the P(A) associated with the test instrument. For example, if P(A) were .70 then the expected rate of correct classification from the two-alternative forced-choice procedure would be 70%, which would represent a 20% increase on the rate of classification that would be achieved by chance.

Not only does TSD offer a comprehensive account of the properties of prediction systems, it can also be shown that some of the indices of predictive power that have been discussed previously are in fact special cases of TSD statistics. The relationship between TSD and these statistics is shown below:

(1) Range and selectivity:- the concepts of range and selectivity proposed by Stott (1960c) show a simple relationship to TSD statistics. Range is defined as the proportion of delinquents who are classified as delinquent; this statistic is simply the hit rate. Selectivity is defined as the proportion of those subjects who are classified as delinquent who turn out to be delinquent; this statistic is the posterior probability corresponding to the hit rate: P(S/S'). In the discussion above we have described this probability as the detection rate.

Stott claims that the concepts of range and selectivity provide an adequate basis for assessing predictive power. In fact this is not entirely true. These two statistics, in conjunction with the base rate data, do not provide the same amount of information about prediction that is conveyed by the ROC curve. The weakness of the concepts of range and selectivity is that they are concerned with the correct classification of delinquents not the correct classification of both delinquents and non-delinquents.

(2) The Mean Cost Rating:- MCR is in fact a simple linear transformation of P(A). In Appendix 2 it is proved

that MCR = 2P(A) - 1. This result has an easy geometric interpretation.[1] Figure 1.6.2 shows an ROC curve and the complement of this curve created by plotting the correct rejection rate against the miss rate. It will be seen that the area between the ROC curve and the chance line (Z3) is equal to the area between the complement of this curve and the chance line (Z2). It can be shown that the MCR is in fact equal to the ratio of the area between the chance line and the complement of the ROC curve to the entire area under the chance line (i.e. 0.5) (Duncan et al 1953). Thus MCR = Z2/(Z1 + Z2) = Z2/.5 = Z3/.5. Further, from the definition of P(A), it follows that P(A) = Z1 + Z2 + Z3. From this it follows readily that MCR = 2P(A) - 1.

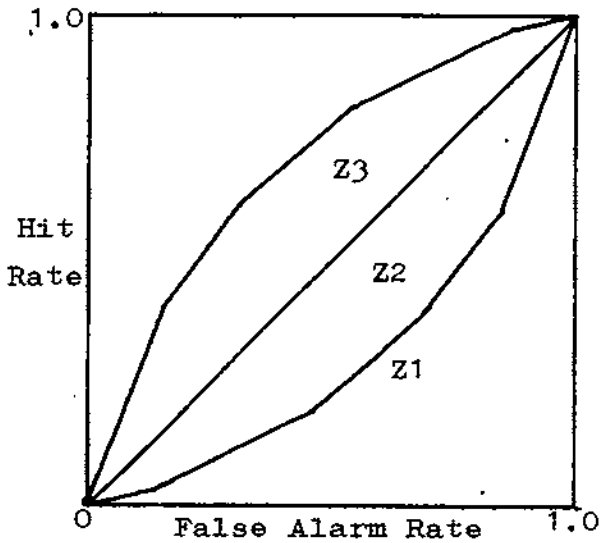


Figure 1.6.2

The relationship between the MCR and P(A) can be expressed most easily as follows: P(A) is the ratio of the area under the ROC curve to the unit square containing this curve; the MCR is the ratio of the area between the ROC curve and the chance line to half the unit square. It is clear from this result that both measures convey exactly the same information expressed in slightly different ways.

---

1. This geometric illustration holds as long as the MCR is positive; when the MCR is negative one must make certain conventions about the algebraic sign of the various values representing areas for the illustration to hold.

From the foregoing discussion it is clear that TSD offers many advantages as an approach to measuring and assessing predictive power. These advantages may be summarised as follows:

(1) TSD makes explicit the relationship between prediction and decision making by showing that decision making requires both the application of risk estimates and of pay off values.

(2) TSD demonstrates that the optimum statistic for forming decisions is the likelihood ratio.

(3) The theory provides a highly efficient and parsimonious method of displaying the probability structure of a prediction instrument via the ROC curve.

(4) Finally, TSD subsumes in one general theoretical framework a variety of indices of predictive power that have been developed in isolation in other areas of criminology.

The combination of a theory making explicit the theoretical underpinning of decision making and the general logic of the probability structure of a prediction instrument would suggest that TSD offers the most systematic means of assessing prediction instruments and, also, one which subsumes much of the previous work carried out in this area.

CHAPTER 2

DATA COLLECTION AND PREPARATION

Section 2.1 Sampling and Data Preparation

The population under study was all boys born in 1957 who were attending New Zealand State schools at 24th April 1967. Data on this population were obtained by having class teachers complete a standard questionnaire (see below). This procedure yielded a sample of 25,348 subjects. The sample was a good approximation to the total population of boys born in 1957 attending New Zealand State schools in 1967 and also covered the great majority (86%) of all boys born in 1957 (cf. Fergusson, Donnell and Slater 1975b).

The data collection for this sample was carried out in two phases:

(1) In the first phase of the study, class teachers completed a standard questionnaire for each boy in the sample. This questionnaire was described as a Child Data Booklet (CDB). Each CDB carried an anonymous code number which was used to identify the boy for the duration of the study.

(2) In the second phase of the study each boy was followed up until the end of 1973 to determine his frequency of appearance before the Children's Court.

The way in which the data were prepared is described below.

CDB Information

Each CDB[1] contained information on the following matters:

(1) The boy's promotional level and number of classmates at this promotional level.

---

1. A copy of the CDB is shown in the appendix to the paper The Structure of the Bristol Social Adjustment Guide, Fergusson, Donnell and Slater, 1975. In Press.

(2) The boy's race and the occupation of his parent/ guardian.

(3) The boy's school attendance, the number of schools attended and the date his schooling began.

(4) The boy's school attainment and personality ratings.

(5) Any intelligence or personality tests that the boy may have been given.

(6) Whether or not the boy was a twin.

(7) A copy of the 1956 version of the Bristol Social Adjustment Guide (BSAG).

(8) A number of supplementary questions on the boy's vision, health and hearing.

The preparation of the CDB data has been described in detail in a previous paper (Fergusson, Donnell and Slater 1975b) and the description given here is a brief summary of the comments provided in that paper. The main contents of the CDB were categorical data which were transcribed to coding sheets using a standard system of coding instructions. The contents of the BSAG were coded in the following way. The BSAG comprises a series of statements descriptive of the child's behaviour in school. This instrument is completed by the child's class teacher who endorses those statements applicable to the individual child. Each statement in the BSAG was treated as a binary variable which could assume the value 0 or 1. The item was scored 1 if endorsement of it implied something adverse about the child or if non-endorsement implied something adverse about the child. (For example, the item "absolutely never greets" was scored 1 if endorsed, whereas the item "walks alertly" was scored 1 if it was not endorsed). Otherwise the item was scored 0. Thus the BSAG data were represented by a string of 0s and 1s which reflected the pattern of endorsements on the instrument.

Follow-Up Information

Each boy in the sample was followed up until the end of 1973 to determine the frequency and nature of his appearances (if any) before the Children's Court. This information was provided by

the statistics section of the Department of Social Welfare which collects such data as part of its routine statistics. The data were provided in coded form identified by an anonymous code number (cf. Fergusson, Donnell and Slater (1975b) for a description of the code number system) and included information on the following matters:

(1) The boy's age, his race and the occupation of his father-figure at the time of the offence.

(2) The boy's school or work situation at the time of the offence.

(3) The reasons for the Court appearance.

(4) The social worker's recommendations about the disposal of the case.

(5) The disposal details of the case.

In any large scale longitudinal study, the follow-up of subjects presents a problem. The present study suffered from fewer of these problems than do most such studies as all follow-up material was obtained from nationally collected official statistics and there was no need to locate each subject in person. However, even such a simplified form of follow-up design has its difficulties. The main difficulty encountered in the present study was that of the matching of code numbers with the Court report data so that this information could be integrated with the CDB data. This resulted in a situation in which, for a number of cases, there was a Court report for a boy born in 1957 which could not be matched up with a corresponding CDB. There are several possible reasons for this situation.

(1) The boy's birth date could have been recorded incorrectly either at the initial data collection or on the Court report so that the two sets of information did not agree with each other. For example, a boy shown as being born in 1957 on a Court report may not have had a CDB completed owing to the fact that his birth year was 1956.

(2) The boy could either have been out of New Zealand or attending a private school at the time of the survey in which case he would not have been a <u>bona fide</u>

sample member and would not have received a code number.

(3) The boy could have been a member of the sample but owing to changes in family status (i.e. name change through mother's re-marrying) it might not be possible to find a code number for the boy.

The data provided by the Department of Social Welfare showed that there were 7,231 appearances made by boys allegedly born in 1957, between 1967 and 1973. Using a series of intensive search procedures including checks on Department of Social Welfare records, checks on birth dates at the Registrar of Births and checks on Catholic school enrolments, it was possible to attach code numbers to a total of 5,972 Court reports leaving a total of 1,259 reports to be accounted for. Of these 1,259 Court reports it was possible to account for the lack of a code number in 489 cases. Table 2.1.1. shows the reasons for the lack of code numbers.

Table 2.1.1 REASONS FOR LACK OF CODE NUMBERS

Reason	Number
Listed birth date incorrect	192
In private school at survey	250
Overseas at time of survey	47
Total	489

Of the 489 Court appearances for which there was a reason for the lack of a code number, 192 (39%) were excluded because of incorrect birth dates, i.e. the boys were not born in 1957, 250 (51%) were excluded because the boys were attending private schools at the time of the survey and 47 (10%) were excluded because the subjects were overseas at the time the study was carried out. In all cases the individuals involved were not bona fide sample members.

For the remaining 770 appearances it was possible, in some cases, to find tentative reasons for the lack of code number. Table 2.1.2 shows these reasons.

Table 2.1.2 POSSIBLE REASONS FOR LACK OF CODE NUMBERS

Reason	Number
No known reason	673
Possibly in private school at survey	39
Possible incorrect listing of birth date	9
Possibly overseas at time of survey	26
Possibly accidently omitted from sample	23
Total	770

Of the 770 appearances there was no apparent reason for the lack of a code number in 673 (87%) cases. In 39 cases there was some suggestion that the boy was in a private school at the time of the survey; in 26 cases there was some indication that the boy might have been overseas; and in 23 cases there was evidence that the boy should have been a sample member but that his Child Data Booklet had been omitted in the initial data collection process.

The implications of the foregoing may be summarised as follows:

(1) Making the conservative assumption that the subjects responsible for the 770 appearances described in Table 2.1.2 were all bona fide members of the sample, the data collection procedure captured 5,972/(7,231 - 489) or 88.6% of valid sample members who had Court reports.

(2) Making the liberal assumption that the reasons given in Table 2.1.2 for lack of code number are correct, the procedure captured 5,972/(7,231 - (489 + 97)) or 89.9% of bona fide sample members who had Court reports.

It can be seen that although the data capture procedure for the follow-up data was not perfect, it managed to account for approximately 90 per cent of the cases in which a Court report was present.

Section 2.2 Analysis Sample and Variable Definitions

The processing of a sample of over 25,000 records is extremely costly and time consuming and to ease this burden it was decided to process the data in batches of approximately 5,000 records. The present analysis is based on a sample of 5,472 records which were selected using a systematic sampling procedure. This procedure is described in detail in Fergusson, Donnell and Slater (1975b) and appeared to produce an acceptable approximation to a simple random sample of records extracted from the data.

The variables used in the analysis presented in this report are as follows:

(1) Criterion variables:- Two variables were used as criteria of juvenile offending. The first was whether or not the boy had appeared before the Children's Court for a charge or complaint of misbehaviour before the end of 1973. This measure was coded as a dichotomous variable which assumed the value 1 if the child had made an appearance and the value 0 if he had not. The second variable used in the analysis was the number of distinct appearances, for charges or complaints of misbehaviour, that a child had made before the end of 1973.

(2) Predictor variables:- The predictor variables for the study were extracted from the CDB information collected at age ten years. These variables may be loosely grouped into three categories:

  (i) Demographic data on the child's race and socio-economic status.
  (ii) Information on the child's school history and performance.
  (iii) Information on the child's social adjustment as measured by the BSAG.

In subsequent chapters of this report we examine the extent to which it is possible to predict the criterion variables from the set of predictor variables.

CHAPTER 3

PREDICTION RESULTS

Section 3.1 Introduction

This chapter examines the results of the application of a number of methods of predicting delinquency using the information in the Child Data Booklet. For analysis purposes, two criteria of offending are used:

(1) Appearance before the Children's Court by the end of 1973 for any charge or complaint involving misbehaviour.[1] By that time all boys in the sample would have been over 16 but under 17 years old. This variable is treated as a simple dichotomous measure which assumes the value 1 if the boy made such an appearance and 0 if he did not.

(2) The second measure is the number of distinct appearances before the Children's Court for charges or complaints of misbehaviour by the end of 1973. By and large, this measure may be treated as a proxy for a measure of seriousness of offending: in general, it is reasonable to assume that a boy who has made several appearances before the Children's Court is a more serious offender than a boy who has appeared only once. However, the measure is only approximate as it takes no account of the seriousness of each individual offence nor the number of separate offences dealt with at each appearance.

The analysis examines the following issues:

(1) The predictive efficiency of four additive prediction models based on the BSAG data: the Delinquency Pre-

---

1. During the period of the study a child could be charged with any offence with which an adult could be charged. For young persons under 17 years of age all such charges except those of murder, manslaughter or minor traffic offences were heard in the Children's Court. In addition, Section 13 of the Child Welfare Act 1925 provided that on the complaint that any child was delinquent or not under proper control the child and his parent might be summoned before the Court for the child to be dealt with under the Act.

diction Instrument designed by Stott (1960a); an unweighted version of this instrument; a points score system applied to a selection of items from the BSAG; and a multiple regression equation applied to the same data.

(2) The application of a non-linear prediction model (AID) to the BSAG data to determine the extent to which such a model improves predictive efficiency as compared to the simple additive models.

(3) A consideration of the extent to which information additional to the BSAG data improves the efficiency of prediction.

(4) A comparison of the predictive efficacy of the approaches described in (1) - (3) above as measured by signal detection statistics.

Section 3.2 Delinquency Prediction Scores

Stott (1960a) has designed a Delinquency Prediction Instrument (DPI) based on a selection of 54 items from the BSAG. A full description of the rationale and content of this instrument is given by Stott (1960a, 1963). Briefly, the DPI assigns to each subject a score based on a weighted sum of the number of items endorsed. Weights are assigned to each item on the basis of the relative frequency with which the item is endorsed in delinquent and non-delinquent populations.

Table 1 in Appendix 3 shows the distribution of DPI scores for the sample of 5,472 boys. This score identifies boys with as low as an 8% risk of offending and as high as a 33% risk of offending. The mean number of appearances associated with each score group shows a similar trend.

The general tendencies in these score distributions are shown in Table 3.2.1 below which presents the relationship between the two criterion variables and the DPI scores grouped into five class intervals. This method of presentation reduces the discriminability of the DPI by very little while expressing the trends in the data in a more readily interpretable fashion.

Table 3.2.1 RISK OF OFFENDING AND MEAN NUMBER OF COURT APPEARANCES BY DELINQUENCY PREDICTION SCORE.

Score	Number	Risk of Offending	Mean Appearances
0-3	3,663	7.7%	0.133
4-8	635	12.6%	0.220
9-20	648	14.7%	0.293
21-29	212	22.6%	0.547
30+	314	29.3%	0.863
Overall	5,472	10.9%	0.220

$\chi^2 = 188.434$
$\phi = 0.186$ ($p < 0.001$)
MCR = 0.254.

eta = 0.234
($p < 0.001$)

While these results indicate that the DPI does discriminate to some extent between delinquent and non-delinquent boys, its predictive accuracy is not great. The correlation between the DPI score and the first criterion variable is +.18 and for the second criterion variable it is +.23. Neither coefficient is large although both are highly statistically significant ($p < 0.001$). It should be noted that both of these estimates are unbiased in that no fitting procedures were used to devise the DPI scores for the present sample.

An issue of some interest is the extent to which the weights used in constructing the DPI score improve the efficiency of prediction. The effects of the weights on the predictive power of the DPI are examined in Table 3.2.2 which shows the matrix of intercorrelations between the two criterion variables, the DPI score, and a new score derived by taking an unweighted sum of the DPI items.

Table 3.2.2 CORRELATION MATRIX

	Appearance	Number of Appearances	Weighted Score	Unweighted Score
Appearance	X	0.76	0.18	0.18
Number of appearances		X	0.23	0.22
Weighted Score			X	0.97
Unweighted Score				X

It can be seen from Table 3.2.2 that the increase in predictive efficiency achieved by the weighting system is negligible: the correlations with the criterion variables are almost identical for the weighted and unweighted scores. Further, the two scores are extremely highly correlated. These results are, however, at variance with the findings presented by Stott (1963) who states "(the) efficiency (of the items) has been significantly increased by weighting them" (p.61).

The differences between the two conclusions can almost certainly be attributed to the fact that Stott's estimate of predictive power for the weighted score was based on the sample

data from which estimates of the item weights were obtained. This estimate is biased owing to the large number of parameters, in the form of weights, that were estimated from the sample data. By contrast, the present measures of predictive power are unbiased. It would seem likely, therefore, that the apparent superiority of the weighting system, as reported by Stott (1963), is due to a statistical artifact caused by over-fitting the sample of observations, and not to any intrinsic superiority of the weighted DPI score.

Section 3.3 A Revised Delinquency Prediction Instrument

In the preceding section it was shown that the predictive power of the DPI was not great. In this section, we examine the extent to which it is possible to increase the predictive power of the instrument by revising the item content and weighting system used. To ensure that the estimates of predictive power that were obtained were not inflated, the procedure described in this section of the report used a construction/validation procedure: the sample was randomly partitioned into two groups of 2,637 boys and 2,835 boys; the first group served as the construction sample and the second group as the validation sample.

The first stage of the revision procedure involved selecting a pool of items from the BSAG to serve as candidate items for the revised instrument. This was done by correlating all BSAG items with the two criterion variables and selecting those variables which were correlated greater than $|.10|$ with <u>either</u> criterion variable. The value of $|.10|$ was somewhat arbitrary but appeared to produce a reasonable number of items which had good face validity as predictors of delinquent behaviour. Table 3.3.1 shows the selected items and their correlations with the criterion variables for the construction sample.

Table 3.3.1 CORRELATIONS OF SELECTED ITEMS WITH APPEARANCE AND NUMBER OF APPEARANCES BEFORE THE CHILDREN'S COURT (CONSTRUCTION SAMPLE)

Item	Appearance	Number of Appearances
Sometimes eager, sometimes definitely avoids (greeting)	.076	.111
Offers except when in a bad mood (helping teacher)	.079	.106
Always keen to answer (answering questions)	-.104	-.081
Suspicious (on the defensive) (liking for attention)	.087	.113
Well behaved	-.112	-.095
Very naughty, difficult to discipline	.075	.117
Plausible, sly, will abuse trust	.117	.161
Always or nearly always truthful	-.124	-.158

Item	Appearance	Number of Appearances
Sometimes a fluent liar	.098	.167
Habitual slick liar; has no compunction about lying	.111	.161
Normal for age (attitude to correction)	-.089	-.111
Resentful muttering or expression at times (attitude to correction)	.104	.122
Cannot attend or concentrate for long	.155	.141
Works steadily (persistence (class-work))	-.127	-.112
Reading level (English)	.131	.120
Arithmetic skill (Maths)	.128	.125
Sticks to job (persistence (manual tasks))	-.109	-.098
Bad sportsman (plays for himself only, cheats, fouls) (team games)	.129	.193
Starts off others in scrapping and rough play	.098	.113
Can always amuse himself; works patiently at models, etc. (free activity)	-.120	-.104
Does not know what to do with himself, can never stick at anything long	.093	.117
Squabbles, makes insulting remarks (ways with other children)	.078	.102
Hurts by pushing about, hitting	.080	.149
Misbehaves when teacher is out of room	.108	.153
Disliked, shunned (attitude of other children)	.093	.114
Associates mostly with unsettled types	.103	.113
Has truanted once or twice	.117	.103
Has truanted often	.075	.158
Has cut lessons	.068	.170
Looks after books, etc.	-.127	-.137
Careless, untidy, often loses or forgets books, pen	.108	.128
Sensible (ability at class jobs)	-.125	-.130
Untrustworthy (ability at class jobs)	.103	.093
Scruffy, very dirty	.119	.150
Damage to public property, etc. (of school, fences, unoccupied houses)	.117	.122
Follower in mischief	.084	.150
Bad language, vulgar stories, rhymes, drawings	.123	.173

It can be seen from the above table that 37 of the BSAG items showed correlations of over |.10| with one or both of the criterion variables. In general, the items selected appear to be of three types:

(1) Items relating to dishonest or sly behaviour.

(2) Items relating to lack of concentration, carelessness or restlessness.

(3) Items relating to moody or variable behaviour.

The selected items were combined to produce two additive prediction scores:

(1) An unweighted sum of predictor items. This score was based on the convention that the item was scored 1 if its endorsement implied something adverse about the child and 0 otherwise. For ease of future identification this score will be described as the Unweighted Points Score (UPS).

(2) A weighted sum of the same predictor items. This score was derived from a sum of the items weighted by the (raw score) regression coefficients for the regression equation between the first criterion variable and the 37 predictor items. Separate regression equations were not used for each criterion variable as prior investigation had revealed that the scores derived from separate regressions were highly correlated, and thus the development of separate scoring systems was redundant.
For ease of identification this score will be described as the Regression Score (RS).

Table 3.3.2 shows the correlations between the two scoring methods and the two criterion variables for the construction sample.

Table 3.3.2 CORRELATIONS BETWEEN APPEARANCE AND NUMBER OF APPEARANCES BEFORE THE CHILDREN'S COURT AND TWO PREDICTION SCORES (CONSTRUCTION SAMPLE)

	Appearance	Number of Appearances
UPS	.220	.244
RS	.291	.321

It can be seen from the above that both scoring methods produce a moderate degree of prediction and, as would be expected, the multiple regression method gives the superior results. However, it must be realised that the estimates provided are likely to be over-optimistic; unbiased estimates of predictive power were obtained by applying the prediction equations to the validation sample. Table 3.3.3 shows the correlations of the two methods of scoring with the criterion variables for the validation sample.

Table 3.3.3 CORRELATIONS BETWEEN APPEARANCE AND NUMBER OF APPEARANCES BEFORE THE CHILDREN'S COURT AND TWO PREDICTION SCORES (VALIDATION SAMPLE)

	Appearance	Number of Appearances
UPS	.243	.259
RS	.232	.257

It can be seen that, on validation, there would appear to be little reduction in the predictive power of the equations, in fact the correlations have increased slightly for the unweighted points score system while those for the regression score have decreased slightly. The results indicate that the unweighted points system is as effective as the multiple regression equation. This conclusion is consonant with the findings of Simon (1971) who reports similar results for the prediction of probation success.

At this point it is possible to compare the predictive efficiency of the four additive prediction models considered:

the weighted DPI score, the unweighted DPI score, the unweighted points score and the multiple regression score.

This comparison is given in Table 3.3.4 which shows the matrix of intercorrelations between the four scoring methods and the two criterion variables for the validation data.

Table 3.3.4 MATRIX OF INTERCORRELATIONS BETWEEN FOUR PREDICTION SCORES AND TWO CRITERION VARIABLES (VALIDATION SAMPLE)

	Appearance	Number of Appearances	DPI Wgted	DPI Unwgted	UPS	RS
Appearance	X	.748	.197	.193	.243	.232
No. Appearances		X	.242	.232	.259	.257
DPI (Weighted)			X	.963	.693	.667
DPI (Unweighted)				X	.723	.651
UPS					X	.747
RS						X

Inspection of the above correlation matrix indicates the following:

(1) All scoring methods produce about the same degree of prediction as measured by the correlation coefficient.

(2) All scoring methods are highly correlated and it would appear that they are measuring the same general set of conditions.

The implications of the above are that the additive models examined are all equally, or nearly equally, efficient as predictors of delinquency and that in the practical situation any one will do as well as any other. This would suggest that for practical purposes the most efficient method of scoring is the one which is most simple to apply. This is the UPS which has no complicated weighting system and involves fewer items than the DPI. To illustrate the level of prediction displayed by this score, Table 2 in Appendix 3 shows a cross-tabulation of the UPS by the risk of offending and the mean number of Court appearances.

Although the level of prediction achieved by the UPS is not great, the score does distinguish between delinquent and non-delinquent boys: at the lowest level of the score only 1.9% of children offend and this group has an average of .05 appearances before the age of 17 years; at the highest score level 31.3% of children offend and this group has an average of .84 appearances before the age of 17 years.

The general trend in this distribution is shown in Table 3.3.5 below which presents the relationship between the two criterion variables and the UPS grouped into five class intervals.

Table 3.3.5 RISK OF OFFENDING AND MEAN NUMBER OF COURT APPEARANCES BY UNWEIGHTED POINTS SCORE

Score	Number	Risk of Offending	Mean Appearances
0-3	819	3.9%	0.057
4-7	759	7.8%	0.129
8-11	526	9.9%	0.163
12-16	460	18.5%	0.407
17+	271	27.7%	0.694
Overall	2,835	10.7%	0.214

χ^2 = 157.730 for 4df eta = 0.256
ϕ = 0.236 ($p < 0.001$) ($p < 0.001$)
MCR = 0.393

The relationship between the UPS and the two criterion variables can be seen more clearly from the graphs presented in Figures 3.3.1 and 3.3.2.

In further sections of the report we examine the way in which the extent of prediction displayed by the simple additive models discussed here can be improved: (a) by applying a non-linear prediction model and (b) by introducing additional information about a child.

Figure 3.3.1 RISK OF OFFENDING BY UNWEIGHTED POINTS SCORE (VALIDATION SAMPLE)

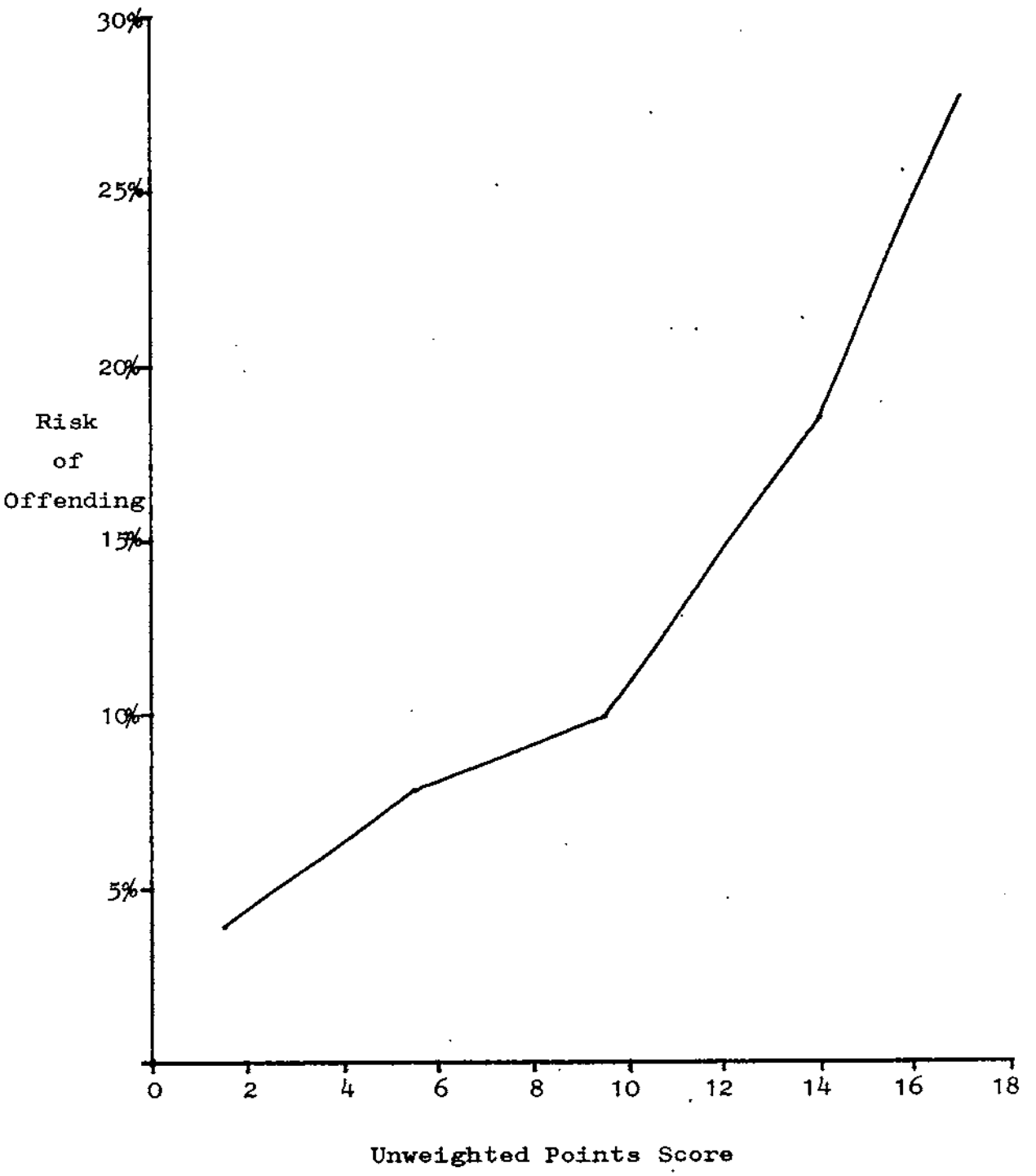
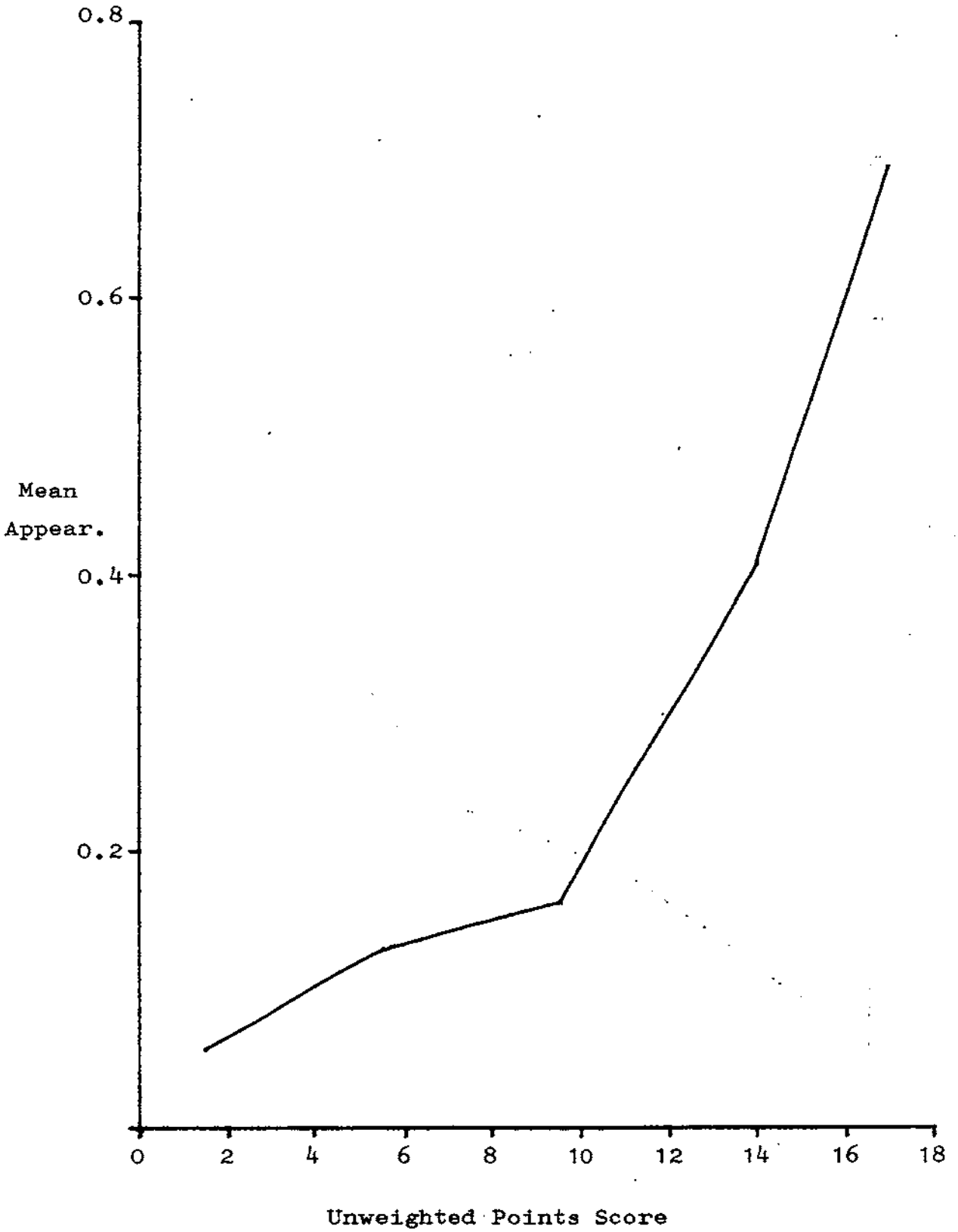


Figure 3.3.2 MEAN NUMBER OF APPEARANCES BY UNWEIGHTED POINTS SCORE (VALIDATION SAMPLE)



Section 3.4 AID Analysis

The 37 selected items were run in two separate AID analyses for the two criterion variables. The stopping rule values for these analyses were set as follows:

(1) The minimum size for any terminal group was 40 observations.

(2) The maximum number of terminal groups was 25.

(3) Any partition was accepted as valid if it accounted for .1% of the total variation in the criterion variable.

These rules differ somewhat from those recommended by Sonquist et al (1971) who suggest a minimum group size of 25 and accept a partition if it accounts for more than .6% of the total variation. The reasons for the differing parameter values were as follows. First, it was felt that with a sample of the present size it was possible to increase the size of any terminal group without doing much violence to the overall structure of the data. Second, in so far as interest was in finding groups of BSAG items which defined subjects with high and low risks of offending, it was felt that a liberal partitioning criterion would allow for a more extensive analysis of these item combinations. In short, we increased the minimum size of the terminal groups to ensure that any combinations of predictor variables yielded fairly stable risk estimates, and at the same time we allowed a very liberal partitioning criterion.

Figure 3.4.1 shows the AID tree for the risk of offending criterion for the construction sample. It can be seen from the figure that the tree partitions the sample into 11 terminal groups. To summarise the results of the analysis, Table 3.4.1 shows the pattern of BSAG item endorsements associated with each of these groups, the risk of offending for the group and the ratio of the group risk to the base rate for the construction sample (11.19%).

Figure 3.4.1 AID TREE FOR RISK OF OFFENDING CRITERION (CONSTRUCTION SAMPLE)

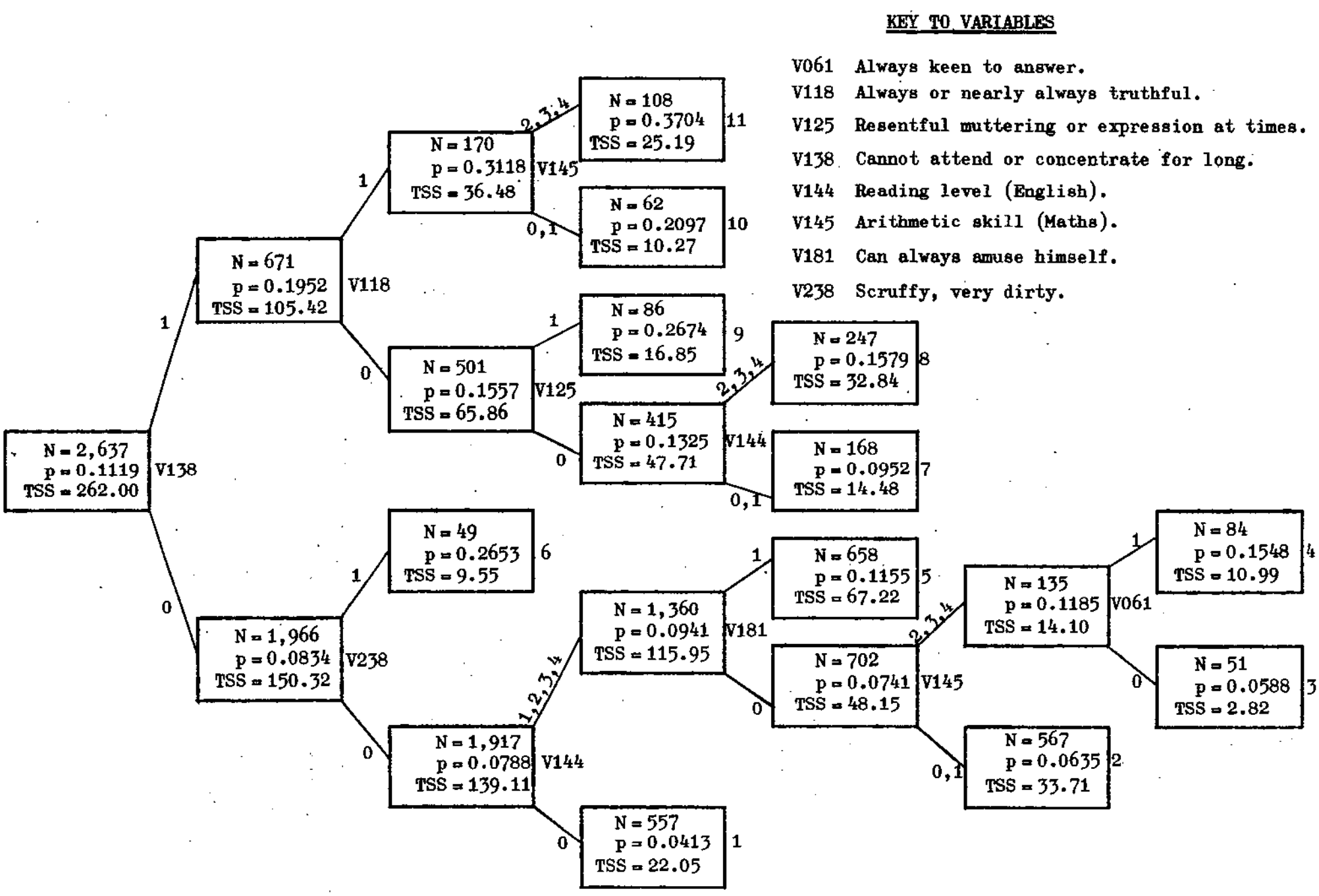


Table 3.4.1 DESCRIPTION OF THE TERMINAL GROUPS OF THE AID TREE: RISK OF OFFENDING CRITERION (CONSTRUCTION SAMPLE)

	Group Description	Risk of Offending	Ratio of Group Risk to Base Rate
1.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (not endorsed); good at reading (endorsed).	4.13%	0.37
2.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (not endorsed); good at reading (not endorsed); can always amuse himself (endorsed); good or average at arithmetic (endorsed).	6.35%	0.57
3.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (not endorsed); good at reading (not endorsed); can always amuse himself (endorsed); good or average at arithmetic (not endorsed); always keen to answer (endorsed).	5.88%	0.53
4.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (not endorsed); good at reading (not endorsed); can always amuse himself (endorsed); good or average at arithmetic (not endorsed); always keen to answer (not endorsed).	15.48%	1.38
5.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (not endorsed); good at reading (not endorsed); can always amuse himself (not endorsed).	11.55%	1.03

	Group Description	Risk of Offending	Ratio of Group Risk to Base Rate
6.	Cannot attend or concentrate for long (not endorsed); scruffy, very dirty (endorsed).	26.53%	2.37
7.	Cannot attend or concentrate for long (endorsed); always or nearly always truthful (endorsed); resentful muttering or expression at times (not endorsed); good or average at reading (endorsed).	9.52%	0.85
8.	Cannot attend or concentrate for long (endorsed); always or nearly always truthful (endorsed); resentful muttering or expression at times (not endorsed); good or average at reading (not endorsed).	15.79%	1.41
9.	Cannot attend or concentrate for long (endorsed); always or nearly always truthful (endorsed); resentful muttering or expression at times (endorsed).	26.74%	2.39
10.	Cannot attend or concentrate for long (endorsed); always or nearly always truthful (not endorsed); good or average at arithmetic (endorsed).	20.97%	1.87
11.	Cannot attend or concentrate for long (endorsed); always or nearly always truthful (not endorsed); good or average at arithmetic (not endorsed).	37.04%	3.31

The terminal groups fall into three categories: those having risks of offending about half that for the total sample (Groups 1, 2 and 3); those having an average or slightly higher risk (Groups 4, 5, 7, 8 and 10) and those having over twice the risk of offending of the sample as a whole (Groups 6, 9 and 11). The level of prediction achieved by the 2 x 11 risk table defined on these terminal groups is modest (ϕ = 0.247; $p < 0.001$).

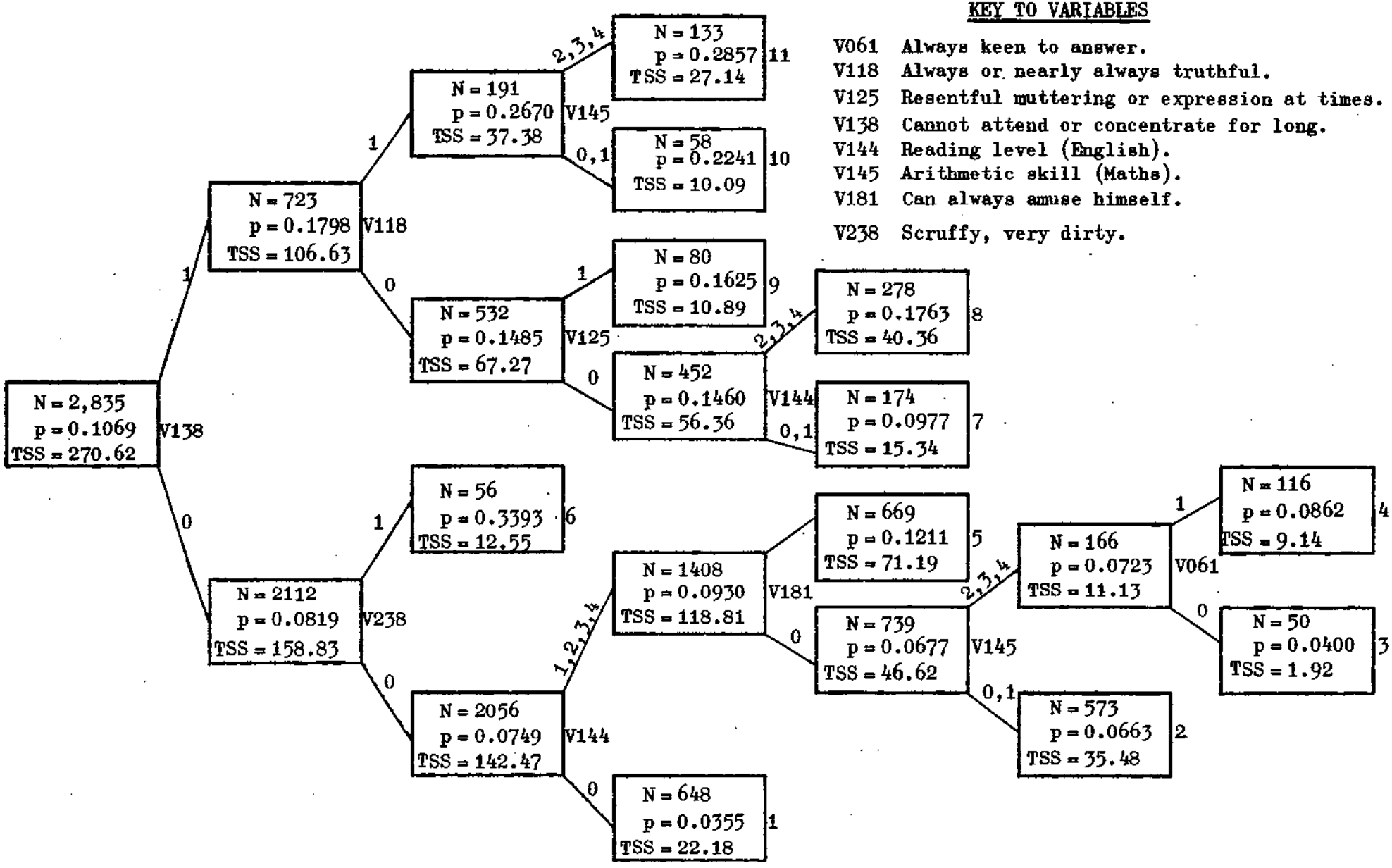
The corresponding AID tree for the validation sample is shown in Figure 3.4.2. It can be seen that the risk values in the tree are similar to the risk values for the construction sample which suggests that the amount of shrinkage on validation was comparatively small. The final risk table for the validation sample is shown in Table 3.4.2 which shows the risk groups arranged in ascending order of risk: the lowest risk in the table is 4% and the highest risk is 34%. The overall level of prediction for the table is similar to that for the construction sample (ϕ = 0.230; $p < .001$) and is slightly less than that achieved with the UPS.

Table 3.4.2 TERMINAL GROUPS FOR AID TREE IN ASCENDING ORDER OF RISK OF OFFENDING : VALIDATION SAMPLE

Group	Delinquent		Non-delinquent		Total	
	Number	Percentage	Number	Percentage	Number	Percentage
1	23	3.55	625	96.45	648	100.00
3	2	4.00	48	96.00	50	100.00
2	38	6.63	535	93.37	573	100.00
4	10	8.62	106	91.38	116	100.00
7	17	9.77	157	90.23	174	100.00
5	81	12.11	588	87.89	669	100.00
9	13	16.25	67	83.75	80	100.00
8	49	17.63	229	82.37	278	100.00
10	13	22.41	45	77.59	58	100.00
11	38	28.57	95	71.43	133	100.00
6	19	33.93	37	66.07	56	100.00
Total	303	10.69	2,532	89.31	2,835	100.00

(χ^2 = 150.112 for 10 df; ϕ = 0.230 ($p < 0.001$); MCR = 0.387).

Figure 3.4.2 AID TREE FOR RISK OF OFFENDING CRITERION (VALIDATION SAMPLE)

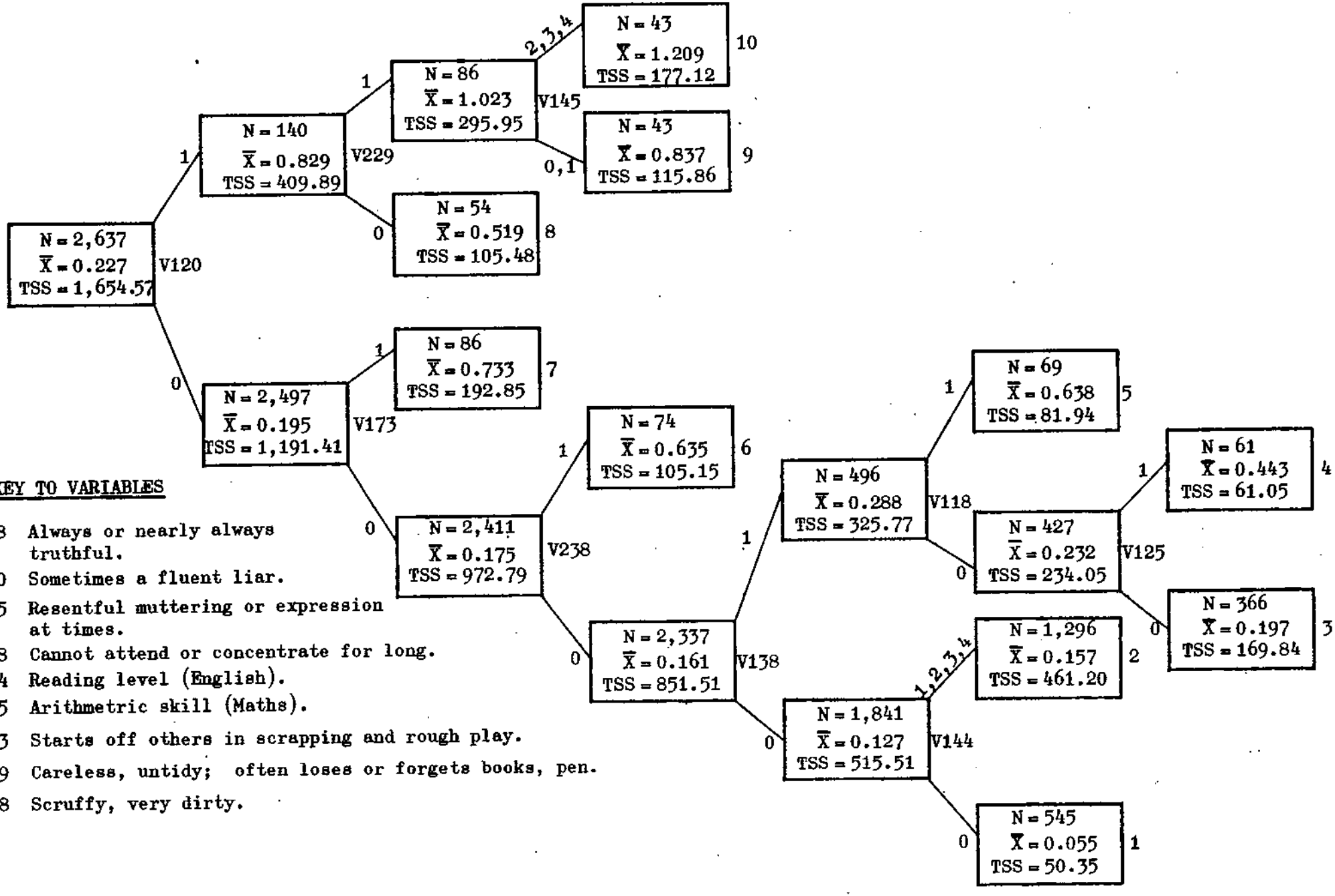


At this point a comment should be made on the validation method used. Simon (1971) adopts a stringent criterion for validating tree structures in that she terminates the validation tree at the point at which no groups can be partitioned according to the partitioning criterion used for the construction sample. This approach is extremely demanding in that one or two spurious splits early in the partitioning process can lead to the rejection of what is otherwise a valid tree structure. Here, we have adopted the less stringent approach of treating the terminal groups of the construction sample as defining a k-way partition which is validated on the validation sample. In using this method, concern is not with validating each partition in the tree structure but with the predictive utility of the final risk table. Both approaches to validation have their liabilities: Simon's approach is liable to reject a tree having predictive power because of the presence of spurious splits; the present approach is prone to accept a tree providing it is predictive irrespective of the presence of redundant splits. The choice of the method of validation depends largely on the purpose to which the validated tree is to be put: if concern is with the interpretation of the tree structure, Simon's approach would seem to be preferable; if concern is with devising a predictive classification then the method described above would seem to be more useful.

The same procedure was applied to the data for the mean number of appearances criterion. Figure 3.4.3 shows the construction tree for this analysis. It can be seen that this tree divides the sample up into ten groups which range in the mean number of appearances from .06 per boy to 1.2 per boy.

Table 3.4.3 presents a description of each group, the mean number of appearances for the group and the ratio of the group mean to the overall mean for the construction sample (0.227 appearances per boy).

Figure 3.4.3 AID TREE FOR MEAN NUMBER OF APPEARANCES CRITERION (CONSTRUCTION SAMPLE)



KEY TO VARIABLES

V118 Always or nearly always truthful.

V120 Sometimes a fluent liar.

V125 Resentful muttering or expression at times.

V138 Cannot attend or concentrate for long.

V144 Reading level (English).

V145 Arithmetic skill (Maths).

V173 Starts off others in scrapping and rough play.

V229 Careless, untidy; often loses or forgets books, pen.

V238 Scruffy, very dirty.

Table 3.4.3 DESCRIPTION OF THE TERMINAL GROUPS OF THE AID TREE : MEAN NUMBER OF APPEARANCES CRITERION (CONSTRUCTION SAMPLE)

	Group Description	Mean Number of Appearances	Ratio of Group Mean to Overall Mean
1.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (not endorsed); cannot attend or concentrate for long (not endorsed); good at reading (endorsed).	0.055	0.24
2.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (not endorsed); cannot attend or concentrate for long (not endorsed); good at reading (not endorsed).	0.157	0.69
3.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (not endorsed); cannot attend or concentrate for long (endorsed); always or nearly always truthful (endorsed); resentful muttering or expression at times (not endorsed).	0.197	0.87
4.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (not endorsed); cannot attend or concentrate for long (endorsed); always or nearly always truthful (endorsed); resentful muttering or expression at times (endorsed).	0.443	1.95

	Group Description	Mean Number of Appearances	Ratio of Group Mean to Overall Mean
5.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (not endorsed); cannot attend or concentrate for long (endorsed); always or nearly always truthful (not endorsed).	0.638	2.81
6.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (not endorsed); scruffy, very dirty (endorsed).	0.635	2.80
7.	Sometimes a fluent liar (not endorsed); starts off others in scrapping and rough play (endorsed).	0.733	3.23
8.	Sometimes a fluent liar (endorsed); careless, untidy, often loses or forgets books, pen (not endorsed).	0.519	2.29
9.	Sometimes a fluent liar (endorsed); careless, untidy, often loses or forgets books, pen (endorsed); arithmetic good or average (endorsed).	0.837	3.69
10.	Sometimes a fluent liar (endorsed); careless, untidy, often loses or forgets books, pen (endorsed); arithmetic good or average (not endorsed).	1.209	5.33

The terminal groups for this criterion tend to fall into only two categories - those having a mean number of appearances falling below the mean for the sample (Groups 1, 2 and 3) and those groups having a mean number of appearances ranging from

twice to five times that of the average for the sample. The power of the 2 x 10 prediction table defined on these terminal groups is again modest (eta = 0.279; $p < 0.001$).

Figure 3.4.4 shows the AID tree applied to the validation sample. The overall distribution of mean appearances for this tree is similar to that for the construction sample.

Table 3.4.4 presents the final prediction table for the mean number of appearances criterion. The table divides the sample from the lowest group having 0.062 appearances per boy to the highest group having 0.879 appearances. The degree of prediction possible for this table shows little shrinkage but is slightly lower than that obtained with the UPS (eta = 0.244; $p < 0.001$).

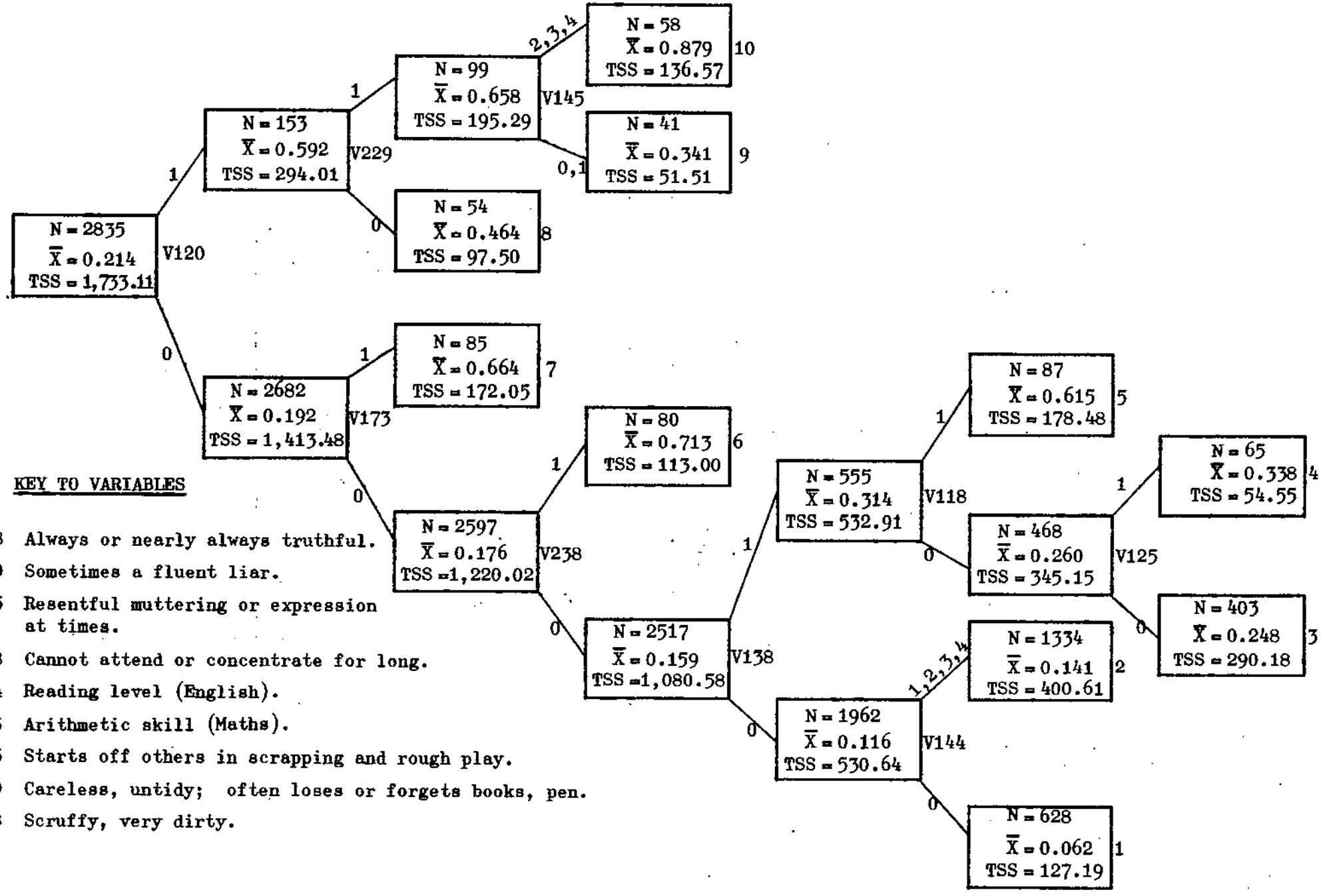
Table 3.4.4 TERMINAL GROUPS FOR AID TREE FOR VALIDATION SAMPLE (IN ASCENDING ORDER OF MEAN NUMBER OF APPEARANCES)

Group	Number	Mean
1	628	.062
2	1,334	.141
3	403	.248
4	65	.338
9	41	.341
8	54	.464
5	87	.615
7	85	.664
6	80	.713
10	58	.879
Total	2,835	.214

eta = 0.244; $p < .001$

The above analysis shows that the predictive power achieved by the AID analysis is no greater than that for the additive models discussed in the previous sections. In addition, the AID results are considerably more cumbersome to use and interpret.

Figure 3.4.4 AID TREE FOR MEAN NUMBER OF APPEARANCES CRITERION (VALIDATION SAMPLE)



KEY TO VARIABLES

V118 Always or nearly always truthful.

V120 Sometimes a fluent liar.

V125 Resentful muttering or expression at times.

V138 Cannot attend or concentrate for long.

V144 Reading level (English).

V145 Arithmetic skill (Maths).

V173 Starts off others in scrapping and rough play.

V229 Careless, untidy; often loses or forgets books, pen.

V238 Scruffy, very dirty.

Section 3.5 The Effects of Other Variables

So far we have considered the extent to which BSAG information collected at the age of ten years may be an effective predictor of future juvenile delinquency. In this section of the report, we extend the argument by considering the extent to which this prediction can be augmented by the introduction of additional information about the child at age ten years.

All variables in the CDB, excluding the BSAG data, were correlated with both criterion variables for the construction sample data and any variable which showed a correlation of greater than $|.10|$ with <u>either</u> criterion variable was selected as a candidate variable to be combined with the UPS. Table 3.5.2 shows the candidate variables selected and their correlations with both criterion variables. The variables described in this table are defined in the following way.

(1) Race: initially this variable was coded into 11 categories using a standard classification which was completed by the boy's class teacher. Table 3.5.1 shows a summary of the racial classification and the distribution of the sample over this classification.

Table 3.5.1 RACIAL DISTRIBUTION OF SAMPLE

Race	Number	Percentage
European	4,511	82.4
Maori (half or more)	630	11.5
Part Maori (less than half)	193	3.5
Pacific Islander	54	1.0
Other	63	1.2
Not specified	21	0.4
Total	5,472	100.0

For analysis purposes, race was redefined as a dichotomous variable: European/Non-European. This was done because a finer racial classification did not provide sufficient numbers of observations to carry out analysis. It should be noted that of the 17.6% of boys who were classified as Non-Europeans, 66% were Maori and a further 20% were part-Maori. Thus this classification could also be loosely interpreted as Maori/Non-Maori.

(2) Socio-economic Status (SES): this variable was based on information collected on the occupation of the boy's parent or guardian and was coded into six categories based on a classification devised by Elley and Irving (1972). These categories can be loosely described as follows:

Category 1 : Professional Workers.
Category 2 : Executive, managerial workers and farmers.
Category 3 : White collar and service workers.
Category 4 : Skilled workers.
Category 5 : Semi-skilled workers.
Category 6 : Unskilled workers.

(3) Oral language
(4) Written language
(5) Reading
(6) Spelling
(7) Arithmetic

These variables consisted of teacher ratings of achievement on a 5 point scale from 1 "outstanding" to 5 "extremely limited".

Table 3.5.2 CORRELATIONS OF CANDIDATE ITEMS WITH CRITERION VARIABLES

Variable	Risk of Offending	Mean Number of Appearances
Race	-.203	-.202
SES	.180	.183
Oral language	.161	.150
Written language	.139	.137
Reading	.129	.116
Spelling	.111	.104
Arithmetic	.157	.132

It can be seen that the selected variables fall into two general classes: demographic variables (race and SES) and teacher ratings of scholastic achievement.[1] To combine these variables with the UPS a stepwise regression procedure was applied to the construction sample data; the variables were entered into a stepwise regression equation in the order of their correlations with the criterion variables. In this analysis, subjects with data missing on either the race or SES variables were deleted. Tables 3.5.3 and 3.5.4 show summary statistics for the stepwise regressions on both criterion variables.

Table 3.5.3 STEPWISE REGRESSION ON OFFENDING/NON-OFFENDING CRITERION

Variable	Multiple R	R^2	R^2 Change
UPS	.211	.044.	.044
Race	.276	.076	.032
SES	.293	.086	.010
Written language	.294	.086	.000
Reading	.296	.087	.001
Arithmetic	.296	.087	.000
Spelling	.296	.088	.001
Oral language	.299	.090	.002

1. At the inception of the study it was suggested that simple teacher ratings of the child would be as effective a predictor as the BSAG. As Table 3.5.2 implies this was not the case: teacher ratings of academic performance correlated only about .13 with both criterion variables; and ratings of such traits as stability, co-operation, perseverance and independence all correlated below .10. The implication of these results is that such global teacher ratings are not as efficient a predictor as the BSAG.

Table 3.5.4 STEPWISE REGRESSION ON NUMBER OF APPEARANCES CRITERION

Variable	Multiple R	R^2	R^2 Change
UPS	.237	.056	.056
Race	.295	.087	.031
SES	.311	.096	.009
Written language	.311	.097	.001
Reading	.311	.097	.000
Arithmetic	.312	.097	.000
Spelling	.314	.099	.002
Oral language	.314	.099	.000

Both tables reveal the same trend: the addition of the variables race and SES increases the predictive power of the equation (as measured by the change in R^2); the introduction of the teacher rating variables adds little or nothing in the way of predictive power. The results indicate that the most effective and parsimonious means of predicting the criterion variables is to combine information on race, SES and the UPS. The problem is that of deciding the appropriate method of combination.

The use of a prediction equation involving race, SES and the UPS is undesirable as it combines demographic and behavioural measures into a global score the interpretation of which would be extremely difficult. Further, it is almost certain that the reason for race and SES improving prediction is that these variables define groups of the population having markedly differing rates of offending (cf. Fergusson, Donnell and Slater 1975a) and that these different rates contaminate the simple regressions of the criterion variables against the UPS. To put the matter another way, the regression equations of the criterion variables against the UPS are subject to multicollinearity effects introduced by the presence of several subpopulations having markedly differing offending rates. The most sensible way to overcome this problem is to partition the sample of observations into a series of sub-groups defined by race and SES, and within each subgroup to derive

an appropriate prediction rule relating the UPS to the criterion variables.[1] This procedure makes explicit the fact that the sample is not homogeneous with respect to the risk of offending and can be partitioned into a number of identifiable subsamples.

In line with this reasoning the sample of observations was partitioned into three groups:

(1) European children of white collar parents:- those European children who were described by categories 1, 2, 3 of the Elley and Irving Scale.

(2) European children of non-white collar status or whose SES was unknown.

(3) Non-European children.

At first sight this partitioning appears to be incomplete in that Non-European children are not differentiated with respect to SES. The reason for this was that there were so few Non-European children of white collar status that the partitioning was not justified. Table 3.5.5 shows the distribution of the two criterion variables over the UPS for the three subgroups using the construction sample. UPS categories were combined to enable stable risk estimates to be made.

The table shows that the three subgroups differ quite markedly with respect to the values of the criterion variables: these differences can be most clearly seen from the plots given in Figures 3.5.1 and 3.5.2 which show the distributions of the two criterion variables by the UPS for the three subgroups.

There are two ways of assessing the predictive power of this table. The first is to examine the level of prediction for

---

1. This approach has in fact been developed more formally by the most recent version of the AID program (Sonquist et al 1971) which incorporates a routine for partitioning samples of observations to maximise the precision of within group regressions.

Table 3.5.5 RELATIONSHIP BETWEEN TWO CRITERION VARIABLES AND THE UPS FOR THREE SUBPOPULATIONS (CONSTRUCTION SAMPLE)

UPS	EUROPEAN WHITE COLLAR			EUROPEAN NON-WHITE COLLAR OR N.S.			NON-EUROPEAN		
	Number	Risk of Offending	Mean Appearances	Number	Risk of Offending	Mean Appearances	Number	Risk of Offending	Mean Appearances
0 - 5	514	2.7%	0.043	482	6.6%	0.093	152	15.1%	0.342
6 - 10	258	4.3%	0.043	325	9.5%	0.151	122	24.6%	0.492
11 - 15	140	7.9%	0.100	229	13.5%	0.279	100	26.0%	0.580
16 +	76	11.8%	0.276	158	27.2%	0.677	81	42.0%	1.185
OVERALL	988	4.6%	0.069	1,194	11.5%	0.222	455	24.8%	0.585
		χ^2=16.811 for 3 df ($p<0.001$) ϕ=0.130 MCR=0.294	eta=0.166 ($p<0.001$)		χ^2=51.794 for 3 df ($p<0.001$) ϕ=0.208 MCR=0.309	eta=0.240 ($p<0.001$)		χ^2=20.491 for 3 df ($p<0.001$) ϕ = 0.212 MCR=0.256	eta=0.224 ($p<0.001$)

Figure 3.5.1 RISK OF OFFENDING BY UNWEIGHTED POINTS SCORE FOR THREE SUBGROUPS (CONSTRUCTION SAMPLE)

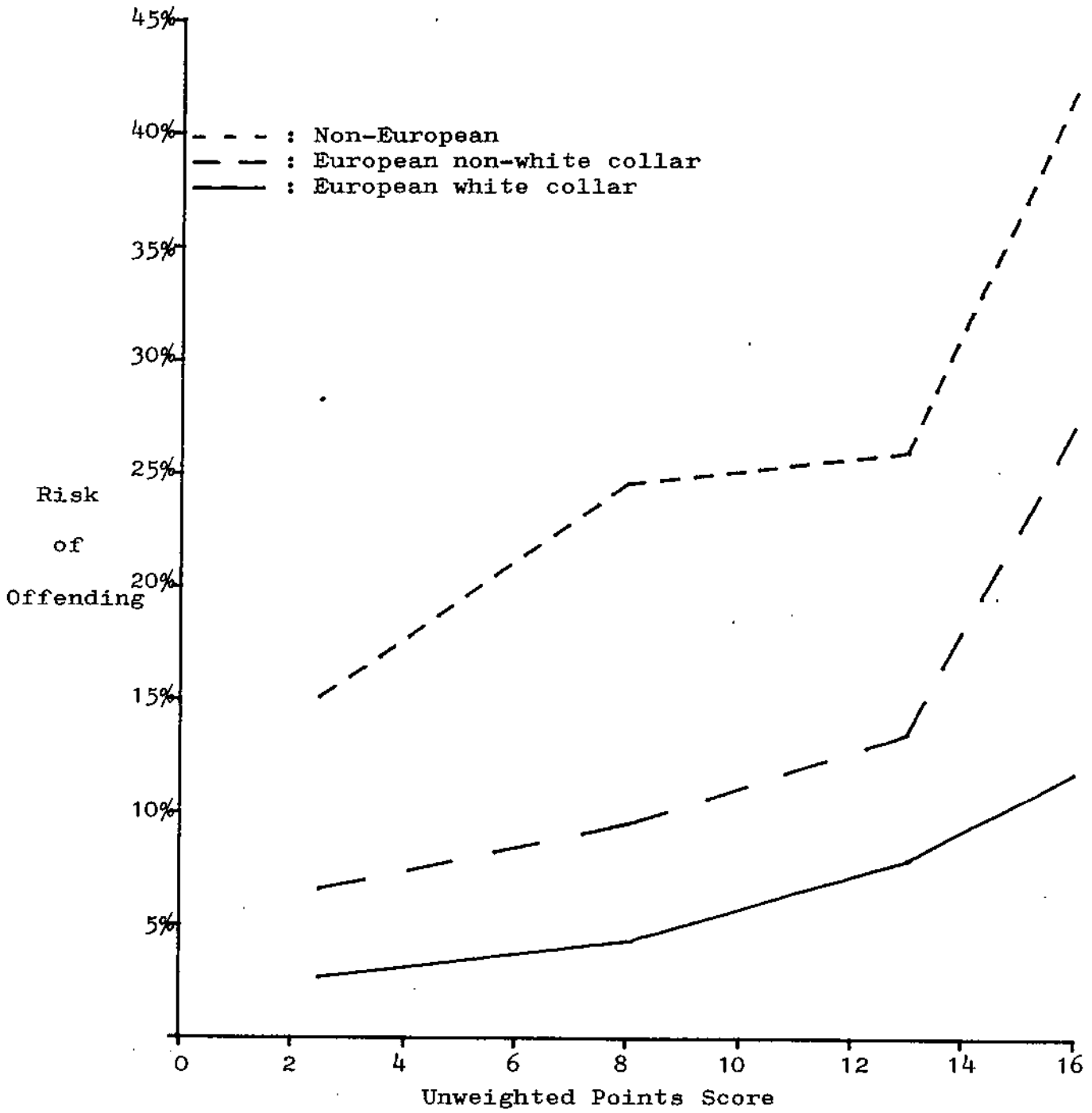
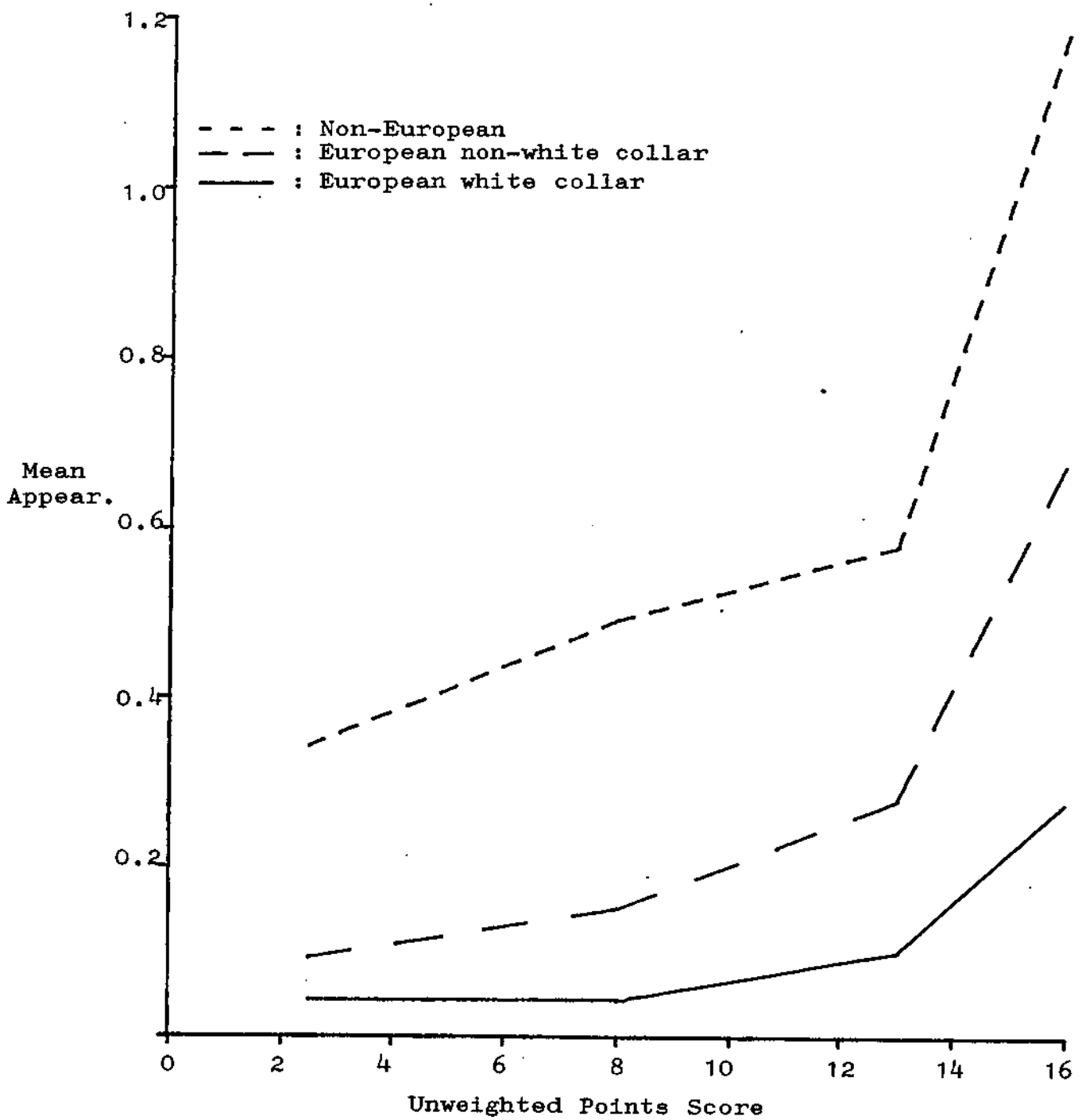


Figure 3.5.2 MEAN NUMBER OF APPEARANCES BY UNWEIGHTED POINTS SCORE FOR THREE SUB-GROUPS (CONSTRUCTION SAMPLE)



the subpopulations in the table; the second is to assess the predictive capacity of the entire table. Measures of prediction for each subgroup are given at the foot of the table and show that, within groups, the level of prediction achieved is not particularly high: the European white collar group tends to have the least amount of prediction and the European non-white collar group tends to have the greatest amount of prediction as measured by ϕ and eta. This difference in level of prediction most probably reflects the influence of the base rate on the predictive capacity of the UPS. This view is supported by the results shown for the base rate independent measure of MCR; using this measure the level of predictability achieved by the European white collar group is higher than that for the Non-European group.

To assess the predictive capacity of the entire table, the correlation ratio was computed for the sample partitioned into 12 groups defined on the three subpopulations and four score intervals. The results of this computation show that the level of prediction obtained by the partitioning process is comparable with that for the multiple regression equation (ϕ = 0.29; eta = 0.31).

Table 3.5.6 shows the corresponding results for the validation sample. By and large, the structure of the table is similar to that for the construction sample and little shrinkage is in evidence. The overall level of prediction achieved is similar to that for the construction sample (ϕ = 0.31; eta = 0.33). The trends in the table are shown in Figures 3.5.3 and 3.5.4.

An explanation of the structure of the validation table is given below:

(1) The group of European children of white collar status have the lowest overall rate of offending: at the lowest range of the points score (0 to 5) less than 3% of these children are offenders, whereas at the highest range (16+) 12% are offenders. In general, while there is some tendency for the risk of offending to increase with the points score the degree of discrimination obtained is not great.

Table 3.5.6 RELATIONSHIP BETWEEN TWO CRITERION VARIABLES AND THE UPS FOR THREE SUBPOPULATIONS (VALIDATION SAMPLE)

UPS	EUROPEAN WHITE COLLAR			EUROPEAN NON-WHITE COLLAR OR N.S.			NON-EUROPEAN		
	Number	Risk of Offending	Mean Appearances	Number	Risk of Offending	Mean Appearances	Number	Risk of Offending	Mean Appearances
0 - 5	531	2.6%	0.038	536	4.1%	0.049	161	13.0%	0.267
6 - 10	271	3.0%	0.041	350	8.3%	0.157	142	23.2%	0.394
11 - 15	152	5.3%	0.105	243	20.2%	0.370	116	31.0%	0.681
16 +	75	12.0%	0.160	171	24.0%	0.596	87	37.9%	1.103
OVERALL	1,029	3.8%	0.057	1,300	10.9%	0.210	506	24.3%	0.542
		χ^2=17.228 for 3 d.f ($p<0.001$) ϕ=0.129 MCR=0.255	eta=0.113 ($p<0.001$)		χ^2=79.876 for 3 d.f ($p<0.001$) ϕ=0.248 MCR=0.416	eta=0.242 ($p<0.001$)		χ^2=22.819 for 3 d.f ($p<0.001$) ϕ=0.212 MCR=0.275	eta=0.266 ($p<0.001$)

Figure 3.5.3 RISK OF OFFENDING BY UNWEIGHTED POINTS SCORE FOR THREE SUBGROUPS (VALIDATION SAMPLE)

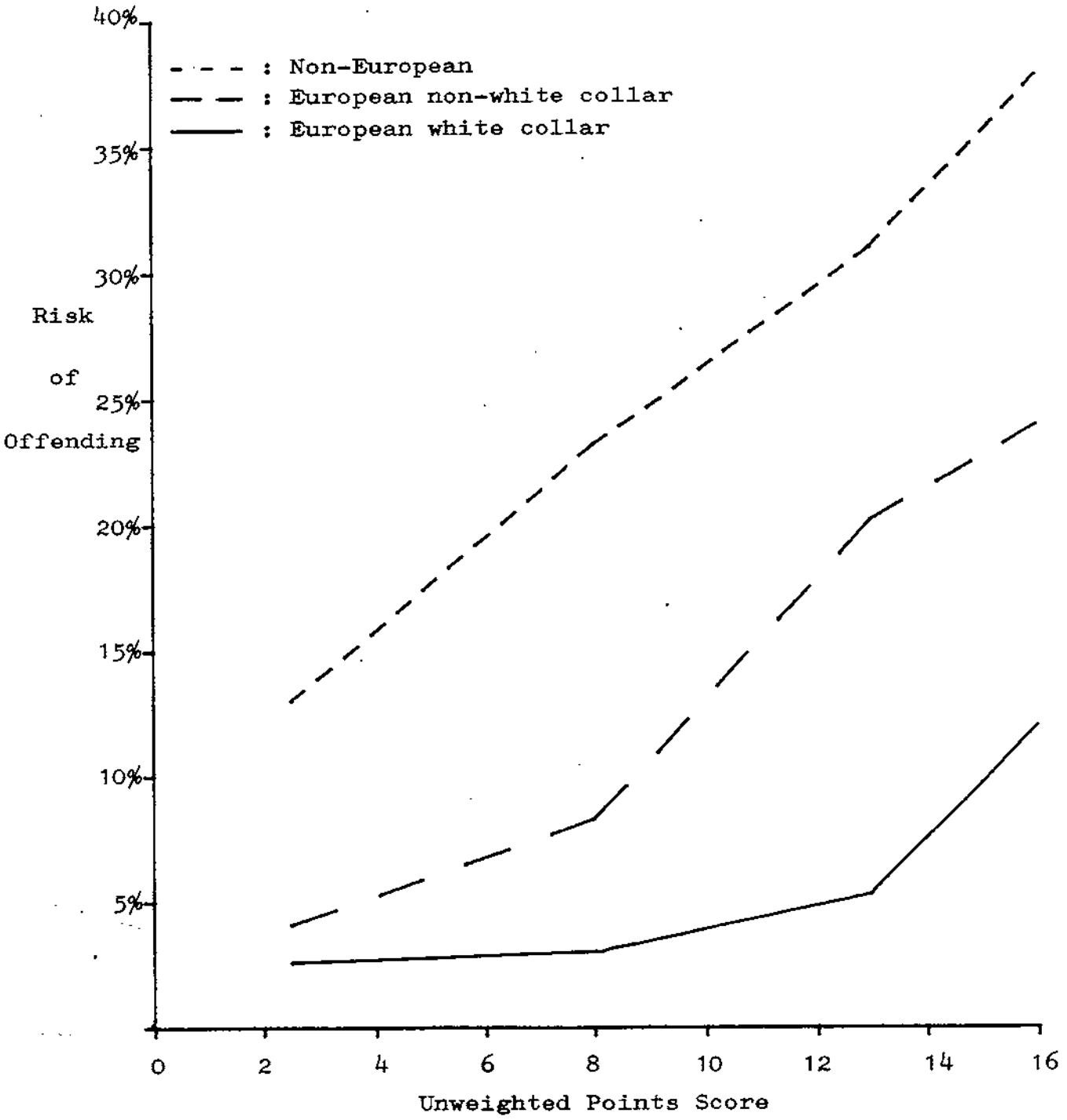
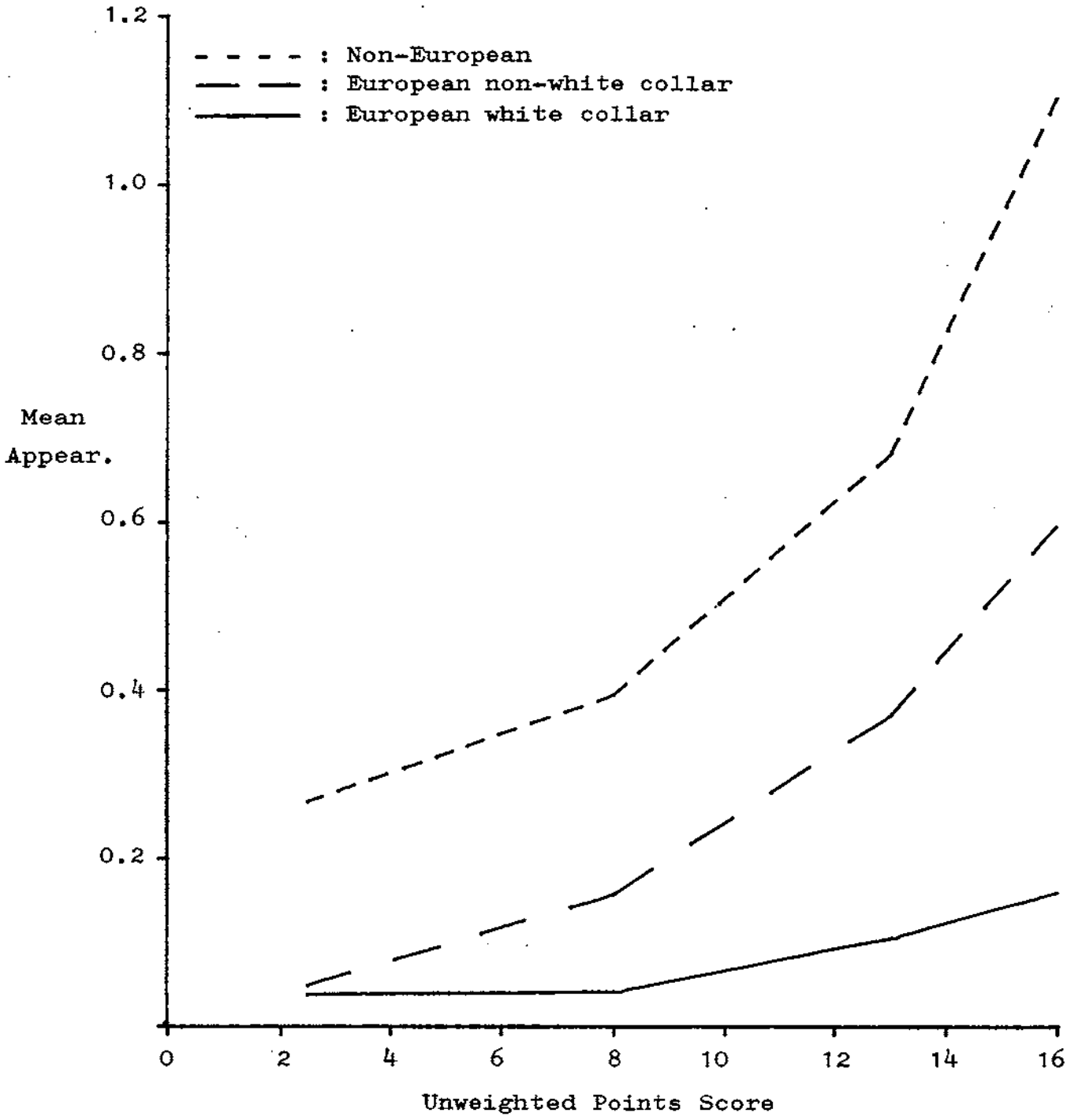


Figure 3.5.4 MEAN NUMBER OF APPEARANCES BY UNWEIGHTED POINTS SCORE FOR THREE SUBGROUPS (VALIDATION SAMPLE)



(2) The group of European children of non-white collar status has an intermediate rate of offending: at the lowest range of the points score 4% of these children are offenders; at the highest range 24% are offenders. The level of discrimination within this group is somewhat greater than that for the European white collar group.

(3) The group of Non-European children has the highest rate of offending: at the lowest range of the points score 13% of these children are offenders; at the highest range 38% are offenders.

In this way, Table 3.5.6 shows a partition of the population of boys into a series of subgroups defined on the UPS, race and SES, which vary in the risk of offending from less than 3% to about 40%. In the next section of the report we examine the ways in which this table can be used to make predictions of those children who are likely to become delinquent and the consequences of such predictions.

Section 3.6 Evaluating the Results

The preceding analysis indicates that two approaches offer the most effective means of identifying potential delinquents:

(1) The unweighted points score system.

(2) The same system applied to the population partitioned into subgroups defined on the basis of race and SES.

In this section of the report we examine the statistical consequences of predictions made on the basis of these methods. The method of evaluation used rests heavily on the discussion of TSD presented in Chapter 1.

The most reasonable set of decision rules for any 2 x k prediction table based on an underlying score distribution X can be formed by successively partitioning the sample at some score Xi and classifying all subjects scoring below this level as non-delinquent and all subjects scoring above this level as delinquent.[1] Thus, for such a prediction table k + 1 decision rules can be formulated: call all subjects delinquent, call all subjects delinquent save those with the lowest score ..... call all subjects non-delinquent. Associated with each decision rule there are a series of statistics which describe the consequences of the decision. As we have stated earlier, these statistics are all implied by the hit rate, the false alarm rate and the base rate. However, in the present case it is useful to present the relevant statistics for each decision in the form of a prediction summary table. These statistics are:

(1) The hit rate: the proportion of delinquents correctly identified.

(2) The false alarm rate: the proportion of non-delinquents incorrectly identified.

---

1. It will be noted that if the score distribution is not perfectly monotone with the values of L(Gi) these decision rules are not optimal by a likelihood ratio criterion. However, with an underlying score distribution it would seem more reasonable to make the cutting values monotone with this distribution than with the set of L(Gi)s.

(3) The detection rate: the proportion of delinquents amongst those classified as delinquent.

(4) The rejection rate: the proportion of non-delinquents amongst those classified as non-delinquent.

(5) The overall proportion of correct classifications.

These statistics, when tabulated for all decision rules that can be formulated, provide an efficient summary of the properties of the prediction table. They tell the user what proportion of delinquents will be detected; what proportion of non-delinquents will be detected; how many of those classified as delinquent will turn out to be delinquent; how many of those classified as non-delinquent will turn out to be non-delinquent; and how many times the classifications made will be correct. The results below show this method applied to the unpartitioned sample and to the sample partitioned into the three groups defined by race and SES.

Table 3.6.1 shows the prediction summary table for the UPS for the entire (unpartitioned) sample. The cutting points on the UPS are defined in equal steps of two score units. The table explores the consequences of 12 decision rules which vary from classifying all subjects as delinquent to classifying all subjects as non-delinquent.

Table 3.6.1 PREDICTION SUMMARY TABLE FOR THE UNWEIGHTED POINTS SCORE (VALIDATION SAMPLE)

Score range of groups classified as non-delinquent	Proportion classified as non-delinquent	Proportion classified as delinquent	Hit rate	False alarm rate	Detection rate	Rejection rate	Proportion correctly classified
NONE	0.000	1.000	1.000	1.000	0.107	*	0.107
0 - 2	0.203	0.797	0.947	0.779	0.127	0.972	0.298
0 - 4	0.364	0.636	0.865	0.609	0.145	0.960	0.442
0 - 6	0.499	0.501	0.749	0.472	0.150	0.946	0.552
0 - 8	0.606	0.394	0.654	0.363	0.177	0.939	0.639
0 - 10	0.702	0.298	0.581	0.264	0.209	0.936	0.720
0 - 12	0.790	0.210	0.452	0.181	0.230	0.926	0.780
0 - 14	0.851	0.149	0.337	0.126	0.242	0.917	0.817
0 - 16	0.904	0.096	0.248	0.077	0.277	0.911	0.850
0 - 18	0.951	0.049	0.145	0.037	0.319	0.904	0.875
0 - 20	0.970	0.030	0.106	0.021	0.372	0.901	0.885
ALL	1.000	0.000	0.000	0.000	*	0.893	0.893

The contents of the table lead to the following conclusions:

(1) The best decision rule, in terms of the number of subjects correctly classified, is to call all subjects non-delinquent. Using this rule 89.3% of subjects are correctly classified. The worst rule is to call all subjects delinquent; this results in 10.7% correct classifications. Therefore, the intuitively appealing strategy of maximising the number of correct classifications would result in the UPS having no utility whatsoever: the best prediction would be achieved using base rate information alone.

However, the errors associated with this decision are all of one type: potential delinquents are classified as non-delinquents. The practical utility of this rule seems to be low.

(2) In a similar way the consequences of decision rules based on non-trivial cutting points on the UPS can be explored using the table. The most useful decision rules would appear to lie within the range of score values from 10 - 14. Using the rules in this region about 45% of potential delinquents and 80% of potential non-delinquents are correctly classified. The probability of a child classified as delinquent turning out to be delinquent is about one in four, and about 93% of non-delinquent classifications are correct. However, these decision rules entail a large number of false alarms: approximately one in five of those who are non-delinquent are wrongly classified as delinquent.

In practical terms, the predictive utility of the table is low. Either one makes the trivial decision to classify all children as non-delinquent or alternatively a decision is made which involves misclassifying a large number of non-delinquents as potential delinquents. Although objective pay off values for the decision process are not available, intuitively it seems unlikely that the high false alarm rate associated with the UPS would allow the instrument to be acceptable for prediction purposes.

Figure 3.6.1 shows the ROC curve derived from Table 3.6.1. The value of P(A) for this curve is .707.[1] This result suggests that when the power of the UPS is measured independently of the base rate, the predictive capacity of the instrument is quite good. Recalling the two-alternative forced-choice interpretation of P(A), it can be seen that in such a situation the use of the UPS would result in a 71% rate of correct classification in contrast to the 50% rate that would be achieved by chance. However, while the instrument appears to be quite effective when measured by the base rate independent measure, P(A), the preceding analysis shows that when it is applied to a prediction situation in which the base rate of offending is 10% the power of the instrument is not sufficient to produce useful predictions.

Table 3.6.2 shows the prediction summary tables for the UPS for the sample partitioned into the three subgroups defined on race and SES. The table presents, for each subgroup, the consequences of decision rules defined in steps of two score units on the UPS. The following comments apply to the results.

(1) For all subgroups the best decision rule, in terms of correct classifications, is to classify all subjects as non-delinquent. This strategy results in 96% correct classifications for the European white collar group; 89% correct classifications for the European non-white collar group; and 76% correct classifications for the Non-European group.

(2) Inspection of the table for the European white collar group indicates that the scope for prediction using the UPS is extremely limited. Even the children with the worst prognosis have only a 17% chance of becoming delinquent and all decision rules entail either a high rate of false alarms or a low hit rate.

---

1. The values of P(A) reported in this section are slightly larger than implied by the values of the MCR given in Tables 3.3.5 and 3.5.6. This is because the values of P(A) computed here are based on a larger number of cutting points than are presented in these tables. This will tend to increase the value of P(A) slightly.

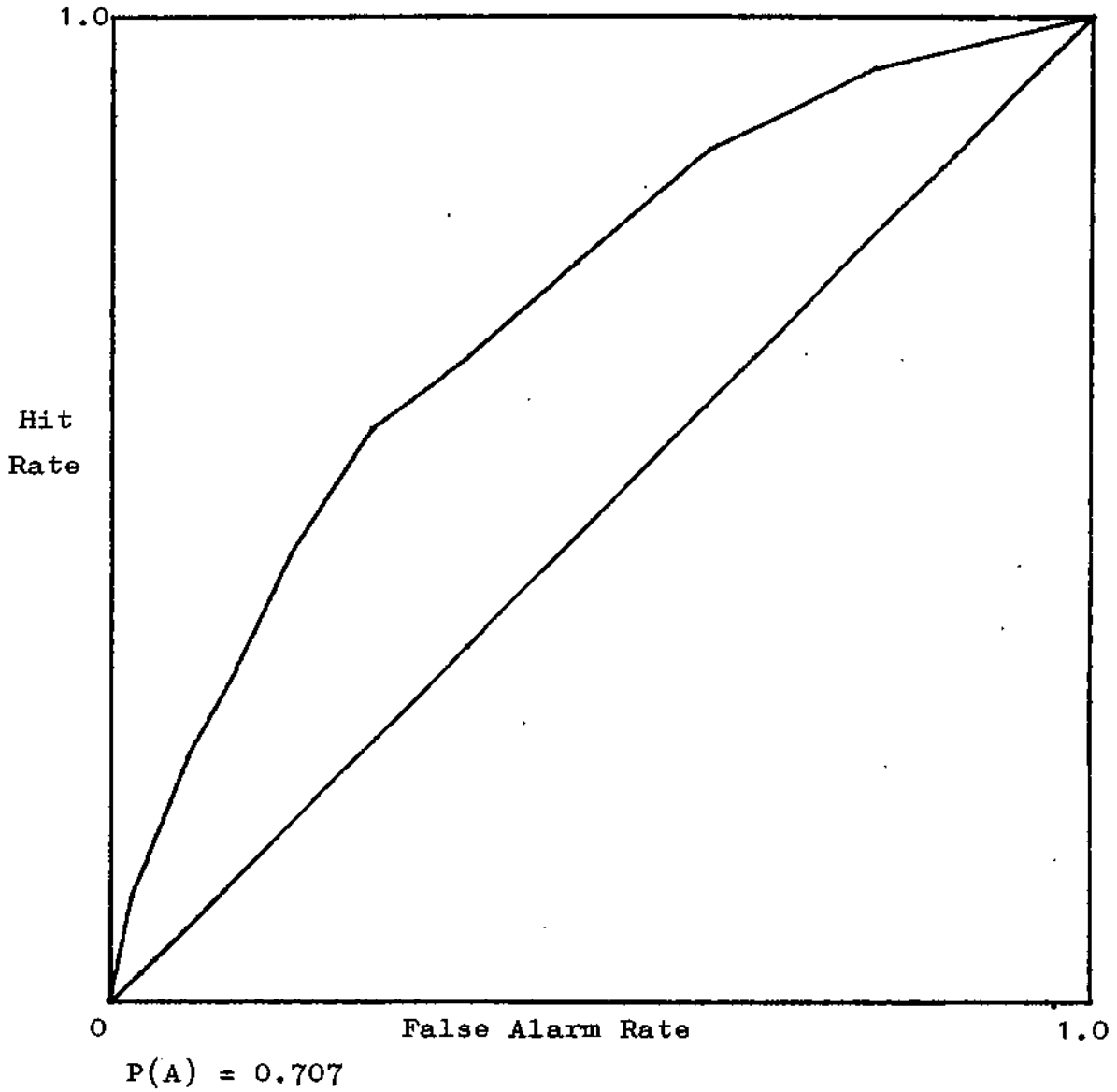


Figure 3.6.1 ROC CURVE FOR UNWEIGHTED POINTS SCORE (VALIDATION SAMPLE)

Table 3.6.2 PREDICTION SUMMARY TABLE FOR THE PARTITIONS OF THE SAMPLE DEFINED BY RACE AND SES (VALIDATION SAMPLE)

EUROPEAN WHITE COLLAR

Score range of groups classified as non-delinquent	Proportion classified as non-delinquent	Proportion classified as delinquent	Hit rate	False alarm rate	Detection rate	Rejection rate	Proportion correctly classified
NONE	0.000	1.000	1.000	1.000	0.038	*	0.038
0 - 2	0.261	0.739	0.974	0.729	0.050	0.996	0.297
0 - 4	0.438	0.562	0.769	0.554	0.052	0.980	0.459
0 - 6	0.588	0.412	0.564	0.406	0.052	0.972	0.593
0 - 8	0.697	0.303	0.513	0.295	0.064	0.974	0.698
0 - 10	0.779	0.221	0.436	0.212	0.075	0.973	0.775
0 - 12	0.863	0.137	0.308	0.130	0.085	0.970	0.848
0 - 14	0.908	0.092	0.256	0.086	0.105	0.969	0.889
0 - 16	0.948	0.052	0.205	0.046	0.148	0.968	0.925
0 - 18	0.977	0.023	0.103	0.020	0.167	0.965	0.947
ALL	1.000	0.000	0.000	0.000	*	0.962	0.962

EUROPEAN NON-WHITE COLLAR AND NOT SPECIFIED

NONE	0.000	1.000	1.000	1.000	0.108	*	0.108
0 - 2	0.188	0.812	0.936	0.796	0.125	0.963	0.283
0 - 4	0.346	0.654	0.887	0.626	0.147	0.964	0.430
0 - 6	0.477	0.523	0.773	0.493	0.160	0.948	0.536
0 - 8	0.577	0.423	0.695	0.390	0.178	0.943	0.619
0 - 10	0.682	0.318	0.638	0.280	0.217	0.942	0.712
0 - 12	0.769	0.231	0.482	0.200	0.227	0.927	0.765
0 - 14	0.838	0.162	0.348	0.140	0.232	0.916	0.805
0 - 16	0.886	0.114	0.262	0.096	0.250	0.910	0.835
0 - 18	0.942	0.058	0.170	0.044	0.320	0.904	0.871
ALL	1.000	0.000	0.000	0.000	*	0.892	0.892

Table 3.6.2 PREDICTION SUMMARY TABLE FOR THE PARTITIONS OF THE SAMPLE DEFINED BY RACE AND SES (VALIDATION SAMPLE)

NON - EUROPEAN

Score range of groups classified as non-delinquent	Proportion classified as non-delinquent	Proportion classified as delinquent	Hit rate	False alarm rate	Detection rate	Rejection rate	Proportion correctly classified
NONE	0.000	1.000	1.000	1.000	0.243	*	0.243
0 - 2	0.121	0.879	0.951	0.856	0.263	0.902	0.340
0 - 4	0.257	0.743	0.870	0.702	0.285	0.877	0.437
0 - 6	0.374	0.626	0.780	0.577	0.303	0.857	0.510
0 - 8	0.498	0.502	0.650	0.454	0.315	0.829	0.571
0 - 10	0.599	0.401	0.561	0.350	0.340	0.822	0.628
0 - 12	0.696	0.304	0.463	0.253	0.370	0.813	0.678
0 - 14	0.773	0.227	0.350	0.188	0.374	0.795	0.700
0 - 16	0.864	0.136	0.244	0.102	0.435	0.787	0.739
0 - 18	0.923	0.077	0.130	0.060	0.410	0.771	0.743
ALL	1.000	0.000	0.000	0.000	*	0.757	0.757

(3) The prospects for prediction with the European non-white collar group are slightly better. Decision rules in the region of UPS scores of 14 - 16 capture about 30% of delinquents; the chance of any child classified as delinquent becoming delinquent is about one in four and over 90% of non-delinquents are correctly classified. However, even these decision rules involve a relatively high frequency of false alarms; about one in ten non-delinquents are wrongly classified as delinquents.

(4) The Non-European group appears to offer the most favourable situation for prediction. The children with the worst prognosis have just over a 40% chance of offending. The most useful decision rules appear to lie in the region of cutting scores of 14 - 16. Using these scores about 24% - 35% of delinquents and between 81% - 90% of non-delinquents are correctly identified. The chance of a child classified as a delinquent becoming a delinquent is between 37% and 44%. However, these decision rules entail a comparatively high rate of false alarms; approximately 15% of non-delinquents are classified as delinquent.

The ROC curves for the three subgroups are shown in Figure 3.6.2. The values of P(A) for the partitioned sample show that within groups the predictive power of the UPS tends to reduce: the value of P(A) is .675 for the European white collar group; .710 for the European non-white collar group and .649 for the Non-European group. This tendency for P(A) to reduce within groups reflects the fact that the partitioning procedure tends to reduce the variance of the UPS and hence reduces prediction.

To assess the predictive capacity of the procedure based on the partitioning of the sample, it is necessary to re-order Table 3.6.2 in terms of the likelihood ratio and to derive a further ROC curve. Figure 3.6.3 shows the ROC curve for the entire partitioned risk table ordered on the basis of the likelihood ratio and, for comparison, the ROC curve for the

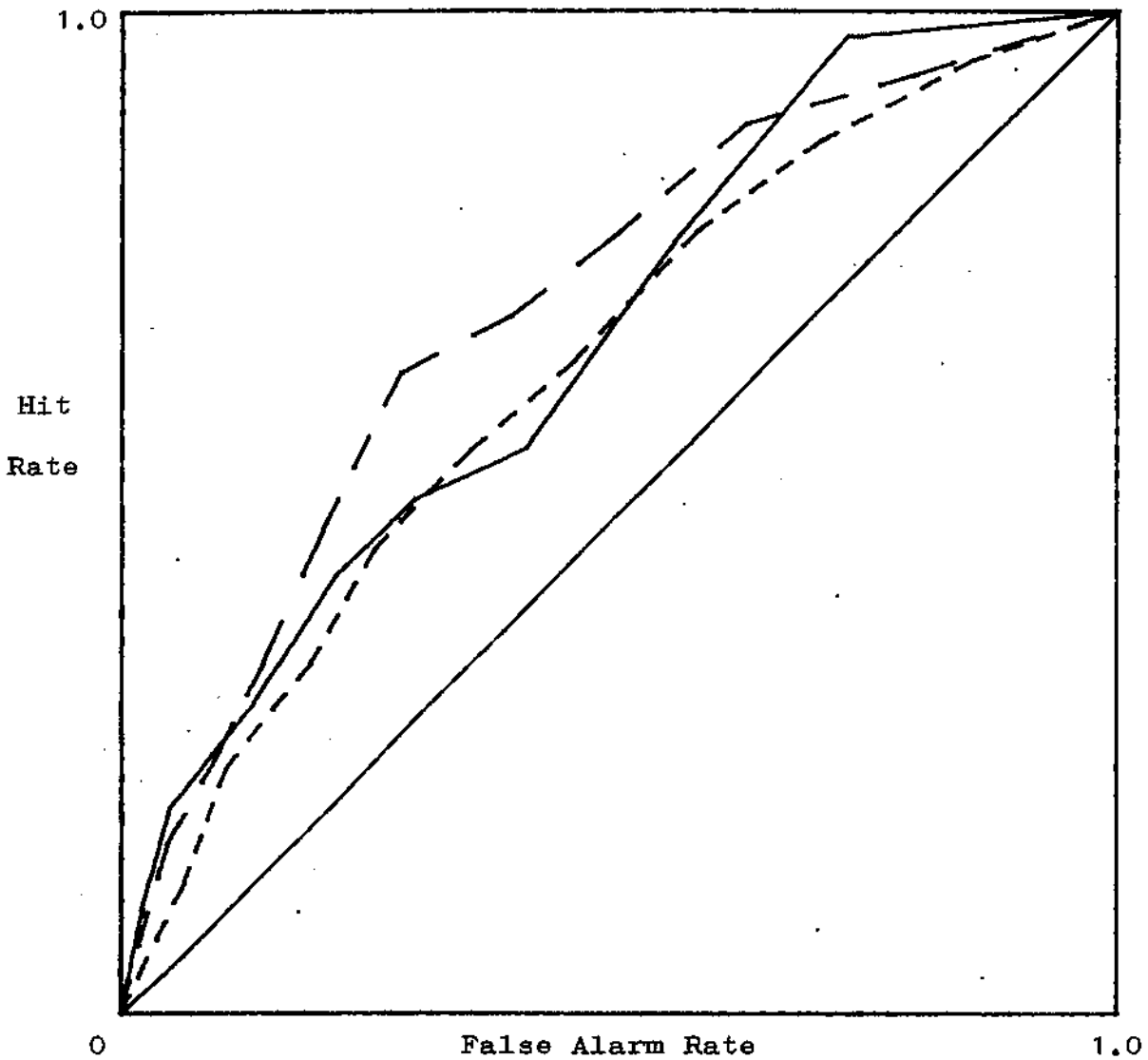
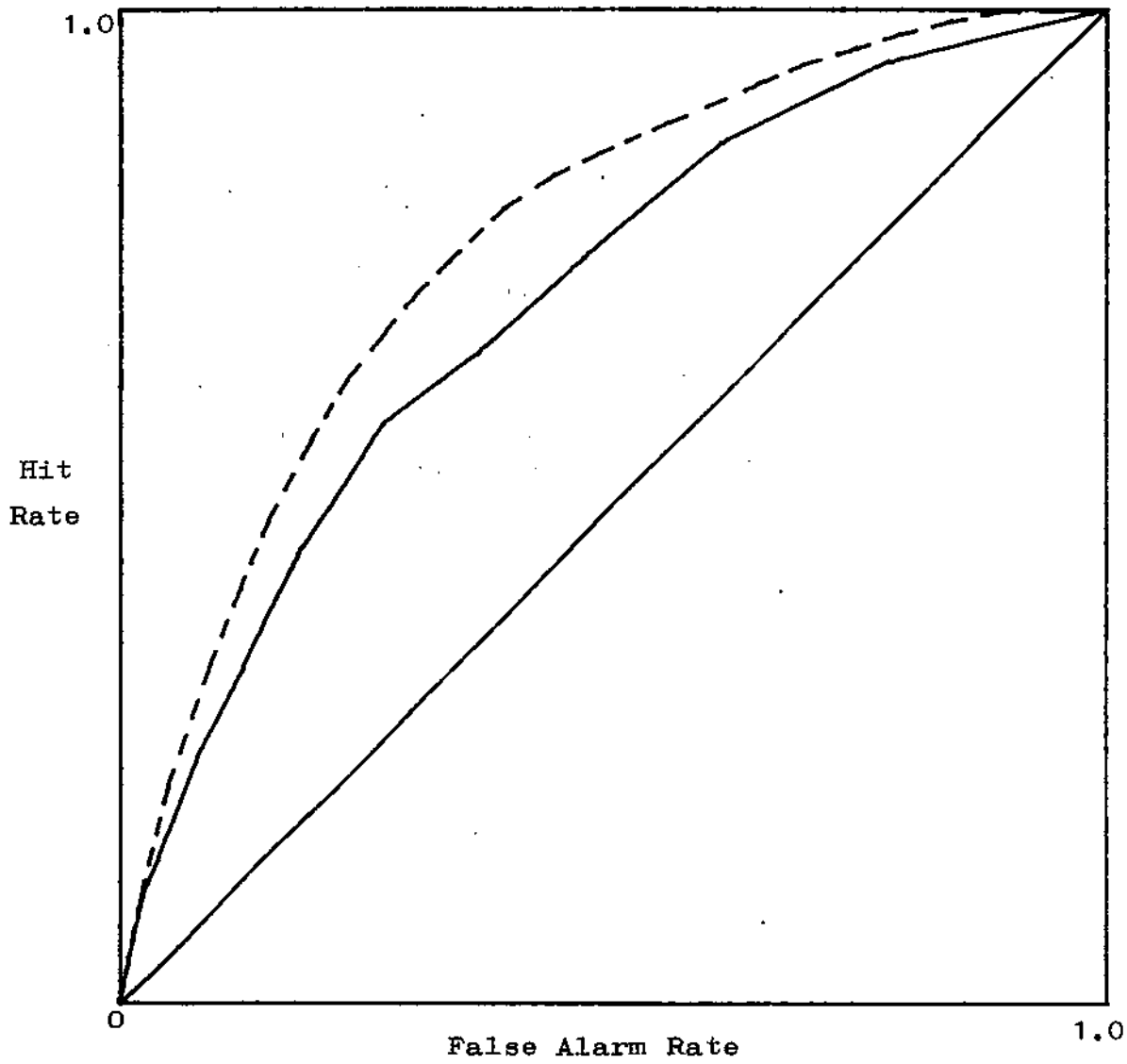


Figure 3.6.2 ROC CURVES FOR UNWEIGHTED POINTS SCORE FOR THREE SUBGROUPS (VALIDATION SAMPLE)



- - - : Partitioned Sample, P(A) = 0.768
——— : Unpartitioned Sample, P(A) = 0.707

Figure 3.6.3 ROC CURVES FOR PARTITIONED AND UNPARTITIONED SAMPLE (VALIDATION SAMPLE)

unpartitioned sample. This comparison reveals, quite clearly, the superiority of the partitioning approach; the value of P(A) for the partitioned sample is .768, whereas that for the unpartitioned sample is .707. To put the matter another way, a two-alternative forced-choice experiment based on the information used to construct the UPS would result in a correct classification rate of 71%; whereas if the experiment used the rules for the partitioned sample a 77% correct classification rate would be obtained.

However, even though the partitioning procedure does increase the predictive capacity of the UPS, this increase is still not sufficient to make the instrument an efficient predictor in a situation where the base rate of offending is 10%, although the analysis does indicate that in a more favourable base rate situation, the instrument may be quite effective.

CHAPTER 4

CONCLUDING COMMENTS

Section 4.1 The Predictive Power of the BSAG

The preceding analysis suggests that the BSAG is capable only of low to moderate prediction of juvenile offending in the general child population: even the most efficient method of data combination accounts for only about 10% of the total variation in the risk of offending. Further, the instrument does not meet the expected level of predictive power suggested by Stott (1960a, 1961, 1963). Stott indicates that the BSAG is able to identify groups of children with a risk of offending as low as 4% and as high as 100%; in the present study, the group with the best prognosis had a 3% risk of offending and the group with the worst prognosis about a 40% risk of offending. As the analysis shows, this separation of risk groups is not sufficient to provide accurate prediction of delinquent behaviour.

The reasons for the lower power of the BSAG in this study are not entirely clear. However, the following points should be noted:

(1) In Chapter 1 we suggested that Stott's risk estimates could have been biased as his method of adjusting the base rate did not entirely overcome the influence of a 50% base rate on the level of predictive power. However, as we also noted, the method by which Stott did adjust his base rate is not entirely clear. The point at issue is whether the adjustment that Stott made - multiplying the frequency distribution of DPI scores for the non-delinquent controls by a factor of 20 - is sufficient to give unbiased estimates of the predictive power of the BSAG when applied to a normal population. The answer depends entirely on the way in which the multiplication was done. If Stott multiplied his non-delinquent distribution by a factor of 20 and then simply applied this distribution to his original data to gain risk estimates, then the results he presents are biased. If, on the other hand, the base rate

frequency of offending was similarly adjusted then the estimates are unbiased. At present we do not know in detail which method of adjustment was used. It is possible, therefore, that the apparently reduction in predictive power reported in this study is due to the fact that the risk estimates presented by Stott have been miscomputed.

(2) A second point that must be borne in mind is that Stott's results were based on a cross-sectional comparison of known delinquents and non-delinquents. Moreover, the boys in Stott's sample varied from eight to 15 years of age. These two factors could have had quite a large effect on the predictive power of the instrument. As we have pointed out earlier, teachers may have rated known delinquents more adversely than non-delinquents; this would have artificially inflated the predictive power of the instrument. Further, the age structure of Stott's sample could have had two consequences for the apparent predictive power of the BSAG. The time lapse between collection of BSAG data and offending was likely to be shorter than for the present study. For example, a 15 year old boy in Stott's sample who offended would have probably done so within one year of the data being collected. In the present study, the time lapse between collection of the original data and offending would have been up to seven years. These differences in the interval between offending and the collection of BSAG data could account for some of the differences in the level of predictive power that have been reported. Second, it may be that the predictive capacity of the BSAG varies with the age at which measurement is made. It seems possible, indeed likely, that the predictive power of the instrument will be higher for older children. If this is the case Stott's results could be better as a consequence of the fact that his prediction was based on a more favourable age distribution.

(3) A final point that must be considered is the extent to which the differences in the samples used may have effected the results. Stott's original findings were

based on a sample of urban children living in a heavily industrialised area; the present results apply to a nationwide sample obtained in a country where a sizeable proportion of the population is rural and where there are few large industrial areas. These differences could, to some extent, account for the differences in the results.

Not only has the study suggested that the predictive power of the BSAG is low but also the findings suggest that the items identified in Stott's Delinquency Prediction Instrument are not the optimum set of predictors. The results indicate that an unweighted sum of 37 selected items does slightly better than Stott's original set of 54 weighted items. However, one must bear in mind that the two scores are highly correlated and in fact the improvement in prediction is marginal. The results suggest also that the system of weighting proposed by Stott improves the predictive power of the DPI only very slightly and that it is likely that the superiority of the weighting system as reported by Stott is due to over-fitting the original sample of observations.

From the above analysis it is reasonable to conclude that the BSAG is unlikely to provide a complete and effective method of predicting juvenile offending and we are in agreement with Marsh (1969) that the power of the instrument makes it suspect as a means of identifying potential delinquents in the general child population.

In the opinion of the authors, the level of predictive validity of the BSAG is such that the instrument should never be used in any of the following circumstances[1]:

(1) As a basis for sentencing or deciding on the disposal of any young offender.

(2) As the basis of any probation officer's or social worker's report on a child.

(3) For the large scale screening of potential delinquents.

---

1. This view is fully endorsed by all the member departments of the JCYO.

In these situations, the predictive accuracy of the instrument is such that it could cause the misclassification of many children. However, while the results suggest that the formal prediction of young offending from BSAG data is a suspect procedure, this does not mean that the BSAG is entirely without use in the identification and treatment of young offenders. In the next section we examine ways in which BSAG data can be used for this purpose.

Section 4.2 The Use of the BSAG

While the BSAG is not efficient as a predictor of juvenile offending, it is not true to say that the instrument conveys no information about such behaviour. In fact, as we have shown, it is possible to divide the population up into groups having relatively low risks of offending and quite high risks of offending. The results thus provide quite substantial amounts of information about the distribution of the risk of offending. This information may be used in a variety of ways by those who deal with young offenders. In particular, prior evidence of the risk of offending for any child allows the professional worker to reach some judgement about the allocation of his own and other resources in dealing with the child. Clearly, one's reaction and method of approach to a child who comes from a group with (say) only a 3% chance of offending will differ from one's reaction to a child who comes from a group with (say) a 40% risk of offending. In the first instance, consideration of the factors likely to lead the child to future juvenile offending will be minimal, whereas in the second case some effort would be made to locate further factors in the child's background which might influence his future offending behaviour. In short, the results presented here can provide useful information which may help the clinical professional to allocate his resources in dealing with the allegedly troublesome or disturbed child. This is a far cry from predicting whether or not a child will be a delinquent solely on the basis of his BSAG score. Further, one may observe that such practice is congruent with practice in other areas (for example, job selection, vocational guidance, marriage guidance counselling) in which test scores are used as an adjunct and guide to the counsellor not as a substitute for him.

However, even such a mild and reasoned application of the results is open to some criticism. It could be objected that prior information about the risk of juvenile delinquency associated with any boy is open to abuse in that persons dealing with children may tend to classify those children with high scores as delinquent, even though the results show that all children are more likely to be non-delinquent than delinquent. The authors are in agreement that such use of the instrument would amount to an abuse of the

results. To make this matter completely explicit, we are of the view that in the hands of a competent professional worker BSAG data can provide valuable information about a child but this information is fallible and must be treated with appropriate caution: the BSAG is not a substitute for the proper clinical evaluation of a case.

Further, it may be argued that although BSAG information may be subject to abuse through thoughtless or careless use, the reverse side of the coin is the misdiagnosis and errors that would occur if this information was not available. It would seem to us a principle of high importance that the treatment of any child referred to any professional worker should be based on the maximum amount of information that can be obtained. The BSAG is one source of such information.

Finally, we would stress that the present paper is a technical report on the predictive capacity of the BSAG; it is not intended as a manual for its use in the prediction of delinquency. Thus it is important that any person who wishes to use the results as presented is prepared to take the time and trouble to become fully conversant with these results before attempting to apply them in practice. At a later date, when further work has been carried out, a manual for the use of the BSAG for prediction purposes may be prepared.

Section 4.3 Reasons for the Low Predictive Validity of the BSAG

So far we have examined the predictive power of the BSAG and have considered the ways in which the results may be applied in the treatment of children referred to professional services because of behaviour problems. At this point it becomes necessary to consider some of the reasons for the low predictive validity of the instrument. These reasons are discussed below:

(1) The validity of the criteria: the measures of offending used in the study are both based on officially recorded delinquent behaviour. There is a growing body of opinion which asserts that officially recorded offending gives a biased measure of delinquent behaviour (Kitsuse and Cicourel 1963; Sellin and Wolfgang 1964; Gold 1966; Gottfredson 1967; Gould 1969; Schur 1971; Simon 1971). Further, the measures have been concerned solely with the frequency of delinquent acts and not with the type and seriousness of the acts. The possible lack of validity of the criteria and their crudeness may have limited the level of prediction achieved.

While this view is worthy of consideration, it should be noted that in many applied situations the prediction of officially detected delinquency is probably more appropriate than the prediction of all forms of behaviour that might be classified as delinquent, irrespective of whether these behaviours come to official attention. Thus, there is a need to weigh up the practical utility of the criterion against its theoretical validity. On balance, the use of a criterion of official offending is not without its merits for many applied situations.

On the issue of the crudeness of the criteria, it must be pointed out that the present report is the first stage of a series of analyses designed to examine the predictive capacity of the BSAG. For the purposes of exploring the data structure a broad definition of offending seems to be the most appropriate. In later papers, we will attempt to examine the relationship between BSAG scores and more refined measures of offending.

(2) The reliability of the BSAG: a further feature which may have reduced the predictive validity of the results is the reliability of the BSAG ratings. One source of predictive error may have been that there are quite marked variations between teachers in completing the instrument. These variations introduce a source of error in prediction. In this respect it is worth noting that the reliability of the BSAG is not particularly high (Fergusson, Donnell and Slater 1975b) and it is possible therefore that the level of prediction has been substantially reduced by error from this source.

However, while teacher descriptions of children have their defects as predictors of future delinquency, one must bear in mind that such data would appear to be amongst the best predictors of juvenile offending (West and Farrington 1973).

(3) The base rate problem: as we have suggested earlier, with a low base rate of offending it is extremely difficult to find effective predictor variables. It is probably the low base rate of offending, more than any other factor, which limits the predictive validity of the BSAG in the present study. In order to improve substantially on the level of prediction provided by the base rate we would have had to identify a sizeable group of children with a risk of offending in excess of 90%. This level of prediction does not seem possible with the BSAG or for that matter with any other existing delinquency prediction system: the detection of pre-delinquents in the general population requires a very powerful predictor.

At the same time it can be observed that the results reported here show that for certain populations the BSAG may be a very useful instrument. For example, if the results were applied to a population in which the base rate of offending was 50%, the level of prediction possible with the instrument would be quite impressive: it would correctly classify about 70% of cases in comparison to the 50% classification rate achieved by

the base rate. In short, the BSAG does discriminate between delinquents and non-delinquents, but the level of predictive power displayed by the instrument is not sufficient to handle the extremely difficult task of predicting delinquency in a population in which the risk of offending is of the order of 10%.

(4) Unknown factors: an essential feature of prediction research is that it attempts to predict unknown future outcomes from limited prior information. To the extent to which the prior information fails to take account of all the factors likely to influence the outcome, prediction must necessarily be imperfect. In the present research this is a matter of obvious importance since one is often attempting to predict events which will occur six or seven years subsequent to the collection of the predictive data. When one considers the variety of influences and factors which impinge on the child and the adolescent it is amazing that any prediction is possible at all.

In short, the low level of prediction achieved by the BSAG is to be expected given the complexity of the behaviour predicted, the lapse of time from the collection of the predictive data and the limited prior information on which the predictions are based.

Section 4.4 In Defence of Prediction Research

Attempts to predict criminality and delinquency have been a source of controversy in criminology. In this section we examine some of the major objections that have been raised and consider ways in which these objections may be answered.

(1) Prediction and Self-Fulfilling Prophecies

The sociologist Merton has suggested that the prediction of human behaviour has certain distinctive properties in that human behaviour is purposive. This implies that the agent about whom prediction is made may respond to the prediction and so influence the predicted outcome. Merton describes situations in which such responses favourably influence the predicted outcome as "self-fulfilling prophecies" (Merton 1957). In a somewhat different context, Lemert (1951) has discussed the development of what he describes as "secondary deviance". Lemert's argument is that the intervention of official agencies in dealing with offenders or potential offenders causes these individuals to be labelled as deviant and hence forces them, or tends to force them, into deviant roles.

An extension of this argument has been applied specifically to prediction studies by several researchers. Kahn (1965) states in reference to the Glueck's Social Prediction Table that "Labeling ..... may worsen a bad situation by influencing school attitudes toward the identified predelinquent and persuading him that he is irremediably "bad"" (p.217). Toby (1961) clarifies this objection by pointing out that "Early identification does not necessarily imply early stigmatization, but early discriminatory treatment seems to" (p.5).

These arguments deserve serious consideration as they suggest that early attempts to identify and treat delinquents may do more harm than good. However, this view is not entirely consistent with the available data on early treatment. In general, one would expect that if self-fulfilling prophecies cause large effects then those children subject to early treatment would

tend to show a greater frequency of deviance than would non-treated children.

The largest single study on the effects of early intervention is the Cambridge-Somerville study in which two groups of children, identified on prior criteria as being pre-delinquent, were subject to two treatment regimes: one group was given various forms of counselling and social work assistance and the other group was left untreated. The results suggest that in terms of delinquent behaviour and personal adjustment there were few differences between the two groups (McCord and McCord 1959).

A further well known early intervention programme is the New York City Youth Board's validation study of the Glueck's Social Prediction Table (Craig and Furst 1965). Boys identified as having a greater than 50% chance of becoming delinquent according to the table were selected for treatment and matched with boys in a control group. The treatment took the form of extensive child guidance therapy. A comparison of the two groups revealed that the same number of serious delinquents appeared in each group.

These results indicate that either the effects of self-fulfilling prophecies associated with treatment were small or that such effects were cancelled out by treatment effects operating in the opposite direction. Toby (1961) takes the latter stance and, using a peculiar mixture of fact, conjecture and common sense, implies that the failure of the Cambridge-Somerville study, in particular, may have been due to the stigmatising effect of early treatment. Perhaps so. However, Toby's argument appears to be far-fetched in comparison to the simpler hypothesis that there were neither large treatment effects nor large effects attributable to the self-fulfilling prophecy.

A point which is not always made completely clear is the way in which predictions become self-fulfilling. A moment's reflection on the matter suggests that it

is not the act of prediction that causes further delinquency but the reaction of official agencies and individuals to the information conveyed by the prediction: no child is made any more or less delinquent by a table which asserts that his risk of offending is (say) 60%.

It is important to recognise this point as it is apparent that the alleged criticism of prediction research is in fact a criticism of the unintended consequences of early intervention not a criticism of prediction per se. Further, one can argue that such unintended consequences may occur in any early treatment project irrespective of whether or not predictions are made. However, it must also be recognised that the act of identifying a child as a potential delinquent may increase tendencies for the treatment program to have stigmatising effects. Finally, it should be observed that these criticisms are not sufficient grounds for rejecting either prediction or early treatment; they simply alert one to the problems that must be faced in developing such programmes.

(2) Clinical and Statistical Prediction

A view that is sometimes advanced is that attempts at statistical prediction are unnecessary since the individual clinician with a greater body of information at his finger-tips is in a position to make more accurate predictions. This view is not supported by the available data. Two major reviews of the efficiency of clinical prediction versus statistical prediction (Meehl 1954; Sawyer 1966) show that statistical prediction is superior. Simon (1971), in considering these results, suggests that in part the superiority of statistical prediction lies with the fact that clinical judgements are unreliable and that this unreliability limits their predictive validity. Further, it should be noted that Sawyer (1966) comments that the best predictions occur when clinical and other data are combined statistically. This would suggest that

although the clinician may have useful information at his finger-tips he is not efficient in combining this information to make predictions. There are several reasons for this. First, one may observe that the number of cases dealt with by the individual clinician is necessarily limited and this limits the scope of the data on which prediction rules are to be formed. Secondly, feedback in the clinical situation may be poor and such feedback as is available is likely to be contaminated by treatment effects. Finally, it is known that human beings tend not to use available prior information in a statistically optimal way in forming predictions (Phillips and Edwards 1966; Peterson and Beach 1967; Wendt 1969).

However, while statistical predictions are more efficient than clinical predictions this should not be construed as suggesting that the statistician is a substitute for the clinician; rather statistical prediction devices should be looked on as useful clinical aids. On this point Meehl (1954) writes:

> "For practical purposes, the concept of efficiency must include some reference to the amount and level of work required to arrive at a given degree of predictive success. Once some sort of statistical backlog has been collected (and this takes no more time than is needed for the clinician to get experience), the actuarial method almost invariably takes less time, less effort, and - no minor point - can be entrusted to lower paid persons possessing much less skill" (p.127).

We would argue, in line with Meehl's comments, that statistical prediction devices offer the advantage of providing the clinician with information which allows him to allocate his resources more effectively and alerts him to features of the case which might otherwise go undetected.

(3) Prediction, Theory, Cause and Treatment

A group of criticisms that have been levelled at prediction devices (notably by Toby 1961) is that they are not well grounded in theory, they fail to consider the causes of crime and they do not indicate likely treatment methods. These criticisms in the main reflect a misunderstanding of the relationship between prediction, theory and cause and effect in science.

Toby (1961) stresses the need for the integration of prediction efforts with a consistent theoretical framework, arguing that predictions made without a theoretical basis are uninterpretable as the researcher has no insight into why they are correct.

However, it is not always the case that an efficient predictor reveals causes or that good predictors are of theoretical importance. The primary criterion of an effective predictive device is that it predicts the dependent variable with optimal accuracy, not that it leads to or supports theoretical conclusions or specifies causes or treatment. Prediction devices are, more generally, empirical generalisations which describe systematic relationships between sets of variables. The relationship between such generalisations and theory is not a simple one. In some instances, empirical generalisations may follow from existing theory or serve to test the theory, and in other cases such generalisations may suggest theory or force the redefinition of existing theory: it is not the case that theory is necessarily antecedent to the development of such generalisation. In criminology this latter argument has particular force as existing theory is ill-defined and appears unlikely to be specified with sufficient precision to lead to the development of effective predictive devices.

In relation to the question of causal factors and prediction devices, Toby suggests that in both the Cambridge-Somerville study and in the New York Youth Board study attention to socio-cultural factors (which he suggests are important causal factors in juvenile crime) can improve the accuracy of prediction. It is reasonable

to expect that causal factors will lead to good prediction, however, it is also the case that good prediction does not necessarily rest on the identification of causes. For example, many diagnostic procedures in medicine rest on the identification of symptoms which are non-causal in nature; the non-causal nature of the symptoms does not impair the efficiency of diagnosis.

It has been asserted that since predictions made without a theoretical basis or incorporation of causal factors are mechanical, they provide no guidance for effective forms of treatment. This argument is even more muddy. The development of a predictive device and the determination of treatment procedures involve two logically and empirically distinct problems. In the first case, one is trying to identify from available data those factors which are symptomatic of a future outcome; in the second case, concern is with evaluation of methods of treating this outcome. It is fairly evident that the former procedure is only indirectly related to the latter. While it is possible that some factors identified as predictors may be useful in treatment regimes this is not necessarily the case. If such outcomes do not occur it does not mean that the prediction instrument has failed. Diagnosis and treatment are two distinct procedures which are not logically dependent on each other.

Lest the reader think we are advocating that prediction devices should be developed without reference to theory, cause or treatment, we would point out that the systematic relationships of a prediction device to all three of these factors is a highly desirable state of affairs and indeed one which should be sought. However, the failure of a device to show such relationships is not sufficient grounds for rejecting the device; the primary criterion for evaluating a predictor is the extent to which it leads to effective prediction.

REFERENCES

Blalock, H. M. (ed.) Causal Models in the Social Sciences. Washington: MacMillan, 1971.

Capwell, D. F. "Personality Patterns of Adolescent Girls. II. Delinquents and non-delinquents". Journal of Applied Psychology, 1945, 29, 289-297.

Challinger, D. "A Predictive Device for Parolees in Victoria". Australian and New Zealand Journal of Criminology, 1974, 7, 44-54.

Coombs, C. H., Dawes, R. M., and Tversky, A. Mathematical Psychology. New Jersey: Prentice-Hall, 1970.

Craig, M. M. and Furst, P. W. "What Happens after Treatment? A Study of Potentially Delinquent Boys". The Social Service Review, 1965, 39, 165-171.

Darlington, R. B. "Multiple Regression in Psychological Research and Practice". In Heermann, E. F. and Braskamp, L. A. Readings in Statistics for the Behavioural Sciences. New Jersey: Prentice Hall Inc., 1970.

Duncan, O. D. "Review of "Predicting Delinquency and Crime" by S. and E. T. Glueck". American Journal of Sociology, 1960, 65, 537-539.

Duncan, O. D., Ohlin, L. E., Reiss, A. J. Jr., and Stanton, H. R. "Formal Devices for Making Selection Decisions". The Americal Journal of Sociology, 1953, 58, 573-584.

Elley, W. B. and Irving, J. C. "A Socio-Economic Index for New Zealand Based on Levels of Education and Income from the 1966 Census". New Zealand Journal of Educational Studies, 1972, 7, 153-167.

Ezekiel, M. and Fox, K. A. Methods of Correlation and Regression Analysis. (3rd edit.) New York: Wiley, 1959.

Fergusson, D. M., Donnell, A. A. and Slater, S. W.
The Effects of Race and Socio-Economic Status on Juvenile Offending Statistics. 1975a. In Press.

Fergusson, D. M., Donnell, A. A. and Slater, S. W.
The Structure of the Bristol Social Adjustment Guide. 1975b. In Press.

Glaser, D. "The Efficacy of Alternative Approaches to Parole Prediction". American Sociological Review, 1955, 20, 283-287.

Glueck, E. T. "Efforts to Identify Delinquents". Federal Probation, 1960, 24, 49-56.

Glueck, S. and Glueck, E. Unraveling Juvenile Delinquency. New York: The Commonwealth Fund, 1950.

Glueck, S. and Glueck, E. Predicting Delinquency and Crime. Cambridge, Massachusetts: Harvard University Press, 1959.

Glueck, S. and Glueck, E. (eds.)
Identification of Predelinquents. New York: Intercontinental Medical Book Corporation, 1972.

Gold, M. "Undetected Delinquent Behaviour". Journal of Research in Crime and Delinquency, 1966, 3, 27-46.

Gottfredson, D. M. "Assessment and Prediction Methods in Crime and Delinquency". Task Force Report: Juvenile Delinquency and Youth Crime. Washington: U.S. Government Printing Office, 1967.

Gould, L. C. "Who Defines Delinquency: A Comparison of Self-Reported and Officially-Reported Indices of Delinquency for Three Racial Groups". Social Problems, 1969, 16, 325-336.

Green, D. M. and Swets, J. A.
Signal Detection Theory and Psychophysics. New York: Wiley, 1966.

Hathaway, S. R. and Monachesi, E. D.
Analyzing and Predicting Juvenile Delinquency with the M.M.P.I. Minneapolis: The University of Minnesota Press, 1953.

Hays, W. L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, 1963.

Herzog, E. Identifying Potential Delinquents: Juvenile Delinquency Facts, Facets No. 5. U.S. Department of Health, Education, and Welfare. Social Security Administration, Children's Bureau, 1960.

Hood, R. Borstal Re-assessed. London: Heinemann, 1965.

Kahn, A. J. "The Case of the Premature Claims. Public Policy and Delinquency Prediction". Crime and Delinquency, 1965, 11, 217-228.

Kitsuse, J. I. and Cicourel, A. V. "A Note on the Uses of Official Statistics". Social Problems, 1963, 11, 131-139.

LaBrie, R. A. "Verification of Glueck Prediction Table by Mathematical Statistics Following Computerized Procedure of Discriminant Function Analysis". In Glueck, S. and Glueck, E. (eds.) Identification of Predelinquents. New York: International Medical Book Corp., 1972.

Lemert, E. M. Social Pathology. New York: McGraw-Hill Book Co., 1951.

MacNaughton-Smith, P. "The Classification of Individuals by the Possession of Attributes Associated with a Criterion". Biometrics, 1963, 19, 364-366.

McCord, W., McCord, J. with Zola, I. K. Origins of Crime. New York: Columbia University Press, 1959.

McNicol, D. A Primer of Signal Detection Theory. London: George Allen and Unwin Ltd., 1972.

Magnusson, D. Test Theory. Reading, Massachusetts: Addison-Wesley, 1967.

Mannheim, H. and Wilkins, L. T. Prediction Methods in Relation to Borstal Training. London: H.M.S.O., 1955.

Marsh, R. W. "The Validity of the Bristol Social Adjustment Guides in Delinquency Prediction". The British Journal of Educational Psychology, 1969, 39, 278-283.

Marshall, T. F. Review of "Identification of Predelinquents: Validation Studies and Some Suggested Uses of Glueck Table". British Journal of Criminology, 1973, 13, 410-411.

Meehl, P. E. Clinical versus Statistical Prediction. Minneapolis: University of Minnesota Press, 1954.

Meehl, P. E. and Rosen, A. "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores". Psychological Bulletin, 1955, 52, 194-216.

Merton, R. K. Social Theory and Social Structure. Glencoe, Ill.: Free Press, 1957.

Monachesi, E. D. "Some Personality Characteristics of Delinquents and Non-Delinquents". Journal of Criminal Law and Criminology, 1948, 38, 487-500.

Monachesi, E. D. "Personality Characteristics of Institutionalized and Non-Institutionalized Male Delinquents". Journal of Criminal Law and Criminology, 1950, 41, 167-169.

Ohlin, L. E. and Duncan, O. D. "The Efficiency of Prediction in Criminology" The American Journal of Sociology, 1949, 54, 441-452.

Peters, C. C. and Van Voorhis, W. R. Statistical Procedures and their Mathematical Bases. New York: McGraw-Hill, 1940.

Peterson, C. R. and Beach, L.R. "Man as Intuitive Statistician". Psychological Bulletin, 1967, 68, 29-46.

Phillips, L. D. and Edwards, W. "Conservatism in a Simple Probability Task". Journal of Experimental Psychology, 1966, 72, 343-354.

Prigmore, C. S. "An Analysis of Rater Reliability on the Glueck Scale for the Prediction of Juvenile Delinquency". Journal of Criminal Law, Criminology and Police Science, 1963, 54, 30-41.

Rao, C. R. "The Problem of Classification and Distance Between Two Populations". Nature, 1947, 160, 835-836.

Reiss, A. J. "Unraveling Juvenile Delinquency. II. An Appraisal of the Research Methods". American Journal of Sociology, 1951, 57, 115-120.

Rempel, P. P. "The Use of Multivariate Statistical Analysis of Minneasota Multiphasic Personality Inventory Scores in The Classification of Delinquent and Nondelinquent High School Boys". Journal of Consulting Psychology, 1958, 22, 17-23.

Rubin, S. "Unraveling Juvenile Delinquency. I. Illusions in a Research Project Using Matched Pairs". The American Journal of Sociology, 1951, 57, 107-114.

Sawyer, J. "Measurement and Prediction, Clinical and Statistical". Psychological Bulletin, 1966, 66, 178-200.

Schumacher, M. "Predicting Subsequent Conviction for Individual Male Prison Inmates". Australian and New Zealand Journal of Criminology, 1974 7, 22-30.

Schur, E. M. Labeling Deviant Behaviour. New York: Harper and Row, 1971.

Sellin, T. and Wolfgang, M. E.
The Measurement of Delinquency. New York: John Wiley, 1964.

Shaplin, J. T. and Tiedman, D. V.
"Comment on the Juvenile Delinquency Prediction Tables in the Gluecks' Unraveling Juvenile Delinquency". American Sociological Review, 1951, 16, 544-548.

Simon, F. H. Prediction Methods in Criminology. London: Her Majesty's Stationery Office, 1971.

Snedecor, G. W. Statistical Methods (4th edit.). Ames, Iowa: Iowa State College Press, 1946.

Sonquist, J. A., Baker, E. L. and Morgan, J. N.
Searching for Structure. Ann Arbor, Michigan: Institute for Social Research, 1971.

Sonquist, J. A. and Morgan, J. N.
The Detection of Interaction Effects. Ann Arbor, Michigan: Institute for Social Research, 1964.

Stott, D. H. "Spotting the Delinquency - Prone Child". The Howard Journal of Penology, 1959, 10, 87-95.

Stott, D. H. "A New Delinquency Prediction Instrument Using Behavioural Indications". The International Journal of Social Psychiatry, 1960a, 6, 195-205.

Stott, D. H. "Delinquency, Maladjustment and Unfavourable Ecology". British Journal of Psychology, 1960b, 51, 157-170.

Stott, D. H. "The Prediction of Delinquency from Non-delinquent Behaviour". British Journal of Delinquency, 1960c, 10, 195-210.

Stott, D. H. Delinquency Prediction Instrument. Manual Based on the Bristol Social Adjustment Guides. London: University of London Press, 1961.

Stott, D. H. "Measuring Crime-proneness in Children". New Scientist, 1962, 13, 633-635.

Stott, D. H. The Social Adjustment of Children. Manual to the Bristol Social Adjustment Guides (2nd edit.) London: University of London Press, 1963.

Stott, D. H. and Sykes, E. G.
The Bristol Social Adjustment Guides. London: University of London Press, 1956.

Taft, D. R. "Implication of the Glueck Methodology for Criminological Research". Journal of Criminal Law, Criminology and Police Science, 1951, 42, 300-316.

Tatsuoka, M. M. Multivariate Analysis. New York: Wiley, 1971.

Thurston, J. R., Benning, J. J. and Feldhusen, J. F.
"Problems of Prediction of Delinquency and Related Conditions over a Seven-year Period". Criminology, 1971, 9, 154-165.

Toby, J. "Early Identification and Intensive Treatment of Predelinquents: A Negative View". Social Work USA, 1961, 6, 3-13.

Venezia, P. S. "Delinquency Prediction: A Critique and a Suggestion". Journal of Research on Crime and Delinquency, 1971, 8, 108-117.

Wendt, D. "Value of Information for Decisions". Journal of Mathematical Psychology, 1969, 6, 430-443.

West, D. J. Present Conduct and Future Delinquency. London: Heinemann, 1969.

West, D. J. and Farrington, D. P. Who Becomes Delinquent? London: Heinemann, 1973.

APPENDIX 1

The general logic of MacNaughton-Smith's (1963) predictive attribute analysis (PAA) and the Sonquist and Morgan (1964) automatic detection of interaction effects (AID) is identical: both procedures partition the sample into a series of subgroups defined on binary splits on the predictor variables. The methods differ in the following details:

(1) The predictor variables for AID may be on nominal, ordinal, interval or ratio scales, whereas those for PAA must be in dichotomous form.

(2) The criterion variable for AID must be on an interval scale with the minimum requirement that the criterion is in dichotomous form. For PAA the criterion variable must be dichotomous.

(3) The AID model selects a split at any point of the analysis by maximising the statistic BSSikp; whereas PAA selects a split by maximising the value of chi square between the predictor and the criterion variable.

It would be reasonable, therefore, to regard PAA as a special case of AID if it could be shown that for any set of data to which PAA could be applied a corresponding AID analysis would produce the same tree structure. This involves the condition that the statistic BSSikp is a monotone increasing function of chi square for any set of totally dichotomous data. The proof is given below.[1]

Consider the 2 x 2 table below which shows the relationship between a dichotomous criterion variable Y and a dichotomous predictor variable Xk.

---

1. The gist of this proof is anticipated in McNaughton-Smith's (1963) paper on PAA in which he comments on the relationship between chi square and reduction of variance.

	Xk = 0	Xk = 1	
Y = 0	a1	a2	a
Y = 1	b1	b2	b
TOTAL	n1	n2	N

The table shows the sample of observations partitioned into two groups: the group of observations for which Xk = 0 and the group of observations for which Xk = 1. Within the first group there are a1 observations for which the criterion variable assumes the value 0 and b1 observations for which the criterion assumes the value 1; similarly there are a2, b2, observations assuming the criterion values 0, 1 respectively in the second group. The total number of observations assuming the criterion value 0 is a, (a = a1 + a2) and the total number of observations assuming the criterion value 1 is b, (b = b1 + b2).

The value of BSS for this partition is:

$$BSSk = TSSi - (TSS1 + TSS2) \quad \ldots\ldots \text{(Eq. 1)}$$

Where TSSi is the total sum of squares of Y for the unpartitioned sample and TSS1, TSS2 are the within groups sums of squares for the subgroups formed by the partition. Equation 1 can be re-expressed as:

$$BSSk = \frac{Nab}{N^2} - \sum_{j=1}^{2} \left[\frac{nj\ aj\ bj}{nj^2} \right] \quad \ldots\ldots \text{(Eq. 2)}$$

It is convenient at this point to define the statistic:

$$Rik^2 = \frac{BSSk}{TSSi} = \frac{\frac{Nab}{N} - \sum_{j=1}^{2} \left[\frac{nj\ aj\ bj}{nj^2} \right]}{\frac{Nab}{N^2}} \quad \ldots\ldots \text{(Eq. 3)}$$

The statistic Rik^2 is, in fact, the proportionate reduction in the sum of squares of the criterion variable that is achieved

by the partition. Multiplying both sides of equation 3 by N, replacing a and aj in the numerator by N-b and nj - bj respectively and re-arranging terms gives:

$$N.Rik^2 = \frac{\sum_{j=1}^{2} \frac{bj^2}{nj} - \frac{b^2}{N}}{\frac{ab}{N^2}} \quad \ldots\ldots \text{(Eq. 4)}$$

The R.H.S. of equation 4 is Brandt and Snedecor's formula for chi square (Snedecor 1946, p.206) and hence the following identity is established:

$$R^2 = \frac{\chi^2}{N} = \phi^2$$

and

$$BSSk = \frac{\chi^2}{N} \quad (TSSi)$$

The above demonstrates that the value of BSS for any set of totally dichotomous data is, in fact, a linear function of the chi square value. Further, since TSSi/N is always non-negative it follows that BSS is a monotone increasing function of chi square, save for the trivial case where TSSi = 0. The implications of this are that the results of attempting to partition any set of totally dichotomous data using either BSS or chi square will produce the same structure of results: or, to put the matter another way, MacNaughton-Smith's predictive attribute analysis is merely a special case of AID.

APPENDIX 2

In the text of the report we presented a geometric illustration of the relationship between the TSD statistic P(A) and the Mean Cost Rating (MCR). We now prove algebraically that MCR = 2P(A) - 1.

Consider a 2 x k prediction table successively partitioned into two classes defined on some series of cutting points defined on the predictor variable. For each cutting rule, one class of observations is predicted as successes and the other class as failures. The consequences of a prediction made from the ith cutting point on the predictor can be described by the following statistics:

(1) The Hit Rate (HR_i): the proportion of successes who are predicted as successes.

(2) The Miss Rate (MR_i): the proportion of successes who are predicted as failures.

(3) The Correct Rejection Rate (CR_i): the proportion of failures who are predicted as failures.

(4) The False Alarm Rate (FA_i): the proportion of failures who are predicted as successes.

These statistics specify the ROC curve for the table. P(A) - the area under the ROC curve - is:

$$P(A) = \frac{1}{2}\left[\sum_{i=1}^{k} (FA_i - FA_{i-1})\ (HR_i + HR_{i-1})\right] \quad \ldots\ (\text{Eq. } 1)$$

(McNichol 1972, p.115);

where the summation is over the series of cutting rules defined on the predictor variable (usually these rules are based on an ordering of the table on the basis of the likelihood ratio).

The MCR for a prediction table laid out as above is defined as:

$$MCR = 1 - \sum_{i=1}^{k} (C_i + C_{i-1})(U_i - U_{i-1}) \quad \text{. . . . (Eq. 2)}$$

(Duncan, Ohlin et al. 1953, p.579);

where C_i denotes the "cost" associated with the ith cutting rule; defined as the proportion of successes predicted as failures. It is immediately apparent that $C_i = MR_i$. U_i denotes the "utility" associated with the ith cutting rule; defined as the proportion of failures predicted as failures. Thus, $U_i = CR_i$.

An equivalent formula for the MCR, and one which is more convenient here, is that derived by Glaser:

$$MCR = \sum_{i=1}^{k} (C_i U_{i-1}) - \sum_{i=1}^{k} (C_{i-1} U_i) \quad \text{. . . . (Eq. 3)}$$

(Glaser 1955, p.248).

By recalling that $MR_i = 1 - HR_i$ and that $CR_i = 1 - FA_i$, and by multiplying out, rearranging and cancelling terms, Equation 3 can be re-expressed as:

$$MCR = \sum_{i=1}^{k} (HR_i FA_{i-1}) - \sum_{i=1}^{k} (HR_{i-1} FA_i) \quad \text{. . . . (Eq. 4)}$$

Further, by a transformation analagous to that used by Glaser for the MCR, it can be shown that P(A) is equivalent to:

$$P(A) = \tfrac{1}{2} \left[\sum_{i=1}^{k} (HR_i FA_{i-1}) - \sum_{i=1}^{k} (HR_{i-1} FA_i) + 1 \right] \quad \text{. . . . (Eq. 5)}$$

From Equations 4 and 5 it follows easily that:

$$MCR = 2P(A) - 1$$

APPENDIX 3

Tables 3.2.1, 3.3.5 and 3.5.6 in the text of the report present tabulations of the risk of offending and the mean number of appearances by two prediction scores: the DPI and the UPS. In order to show the overall trends in the data these distributions were grouped into broad class intervals. This appendix presents the source data for these tables consistent with the convention that approximately 100 observations must be present in each class interval for stable estimates to be made.

Table 1. RISK OF OFFENDING AND MEAN NUMBER OF COURT APPEARANCES BY DELINQUENCY PREDICTION SCORE

Score	Number	Risk of Offending	Mean Appearances
0	2,616	7.57%	.124
1	416	8.65%	.183
2	313	7.03%	.144
3	318	8.49%	.132
4	168	12.50%	.167
5	121	9.92%	.124
6	136	19.12%	.419
7	105	13.33%	.229
8	105	6.67%	.152
9	84	13.10%	.274
10	104	14.42%	.298
11,12	143	8.39%	.147
13,14	105	11.43%	.171
15-17	118	20.34%	.441
18-20	94	22.34%	.479
21-24	111	21.62%	.514
25-29	101	23.76%	.584
30-37	104	26.92%	.923
38-51	105	27.62%	.629
52 +	105	33.33%	1.038
Overall	5,472	10.93%	.220

Table 2. RISK OF OFFENDING AND MEAN NUMBER OF APPEARANCES BY UNWEIGHTED POINTS SCORE (VALIDATION SAMPLE)

UPS	Number	Risk of Offending	Mean Appearances
0	155	1.94%	.045
1	188	2.66%	.032
2	232	3.45%	.056
3	244	6.56%	.086
4	212	4.25%	.075
5	197	8.12%	.132
6	186	10.22%	.167
7	164	9.15%	.152
8	141	9.93%	.149
9	154	8.44%	.175
10	118	7.63%	.153
11	113	14.16%	.177
12	136	16.91%	.324
13	98	16.33%	.357
14-15	164	23.17%	.524
16-17	157	17.83%	.395
18 +	176	31.25%	.841
Overall	2,835	10.69%	.214

Table 3. RELATIONSHIP BETWEEN TWO CRITERION VARIABLES AND THE UNWEIGHTED POINTS SCORE FOR THREE SUBPOPULATIONS (VALIDATION SAMPLE)

EUROPEAN WHITE COLLAR			
UPS	Number	Risk of Offending	Mean Appearances
0 - 2	269	0.37%	.015
3 - 5	262	4.96%	.061
6 - 7	133	2.26%	.045
8-10	138	3.62%	.036
11-14	132	5.30%	.114
15 +	95	10.53%	.137
Overall	1,029	3.79%	.057

EUROPEAN NON-WHITE COLLAR OR NOT SPECIFIED			
UPS	Number	Risk of Offending	Mean Appearances
0 - 1	135	2.96%	.030
2	110	4.55%	.055
3	113	3.54%	.035
4 - 5	178	5.06%	.067
6 - 7	154	10.39%	.175
8 - 9	137	7.30%	.175
10 - 11	109	12.84%	.138
12 - 13	111	17.12%	.297
14 - 16	105	21.90%	.505
17 +	148	25.00%	.642
Overall	1,300	10.85%	.210

NON-EUROPEAN			
UPS.	Number	Risk of Offending	Mean Appearances
0 - 3	100	12.00%	.220
4 - 7	124	19.36%	.355
8 - 11	97	21.65%	.402
12 - 15	98	33.67%	.745
16 +	97	37.93%	1.103
Overall	506	24.31%	.542

A. R. Shearer, Government Printer, Wellington, New Zealand—1975