U.S. Department of Justice
Office of Justice Programs
Office of Juvenile Justice and Delinquency Prevention

The Mathematics of Risk Classification:

Changing Data into Valid Instruments for Juvenile Courts

OJJDP Report

U.S. Department of Justice
Office of Justice Programs
810 Seventh Street NW.
Washington, DC 20531

Alberto R. Gonzales
Attorney General

Tracy A. Henke
Acting Assistant Attorney General

J. Robert Flores
Administrator
Office of Juvenile Justice and Delinquency Prevention

Office of Justice Programs
Partnerships for Safer Communities
www.ojp.usdoj.gov

Office of Juvenile Justice and Delinquency Prevention
www.ojp.usdoj.gov/ojjdp

This report was prepared under a contract with the National Center for Juvenile Justice, the research division of the National Council of Juvenile and Family Court Judges for their National Juvenile Justice Data Analysis Project funded by a cooperative agreement with the Office of Juvenile Justice and Delinquency Prevention of the U.S. Department of Justice. The opinions expressed are those of the authors and do not necessarily reflect the views or endorsement of the National Center for Juvenile Justice, the National Council of Juvenile and Family Court Judges, the Office of Juvenile Justice and Delinquency Prevention, the U.S. Department of Justice, the Maricopa County Juvenile Court or its Juvenile Justice Center, or any other agency or person.

The Office of Juvenile Justice and Delinquency Prevention is a component of the Office of Justice Programs, which also includes the Bureau of Justice Assistance, the Bureau of Justice Statistics, the National Institute of Justice, and the Office for Victims of Crime.

The Mathematics of Risk Classification:

Changing Data into Valid Instruments for Juvenile Courts

Don M. Gottfredson and Howard N. Snyder

National Center for Juvenile Justice

Report

July 2005

Office of Juvenile Justice and Delinquency Prevention

NCJ 209158

Prologue

The primary goal of a risk scale is to predict—not to explain, but to predict. Individual items are placed on the risk instrument because they are believed to be better at predicting the criterion (which in the work that follows is recidivism to juvenile court) than are other available measures. Some risk instruments are developed subjectively, by asking knowledgeable professionals to select items their experience tells them are predictive. Other instruments are developed empirically, using statistical analyses to select among a set of available data elements that capture various attributes of the youth and his/her juvenile court history. Statistically, when an item is a significant predictor of the criterion, it improves the power of the instrument to predict and should be used in the scale.

However, problems arise when empirical power is the sole criterion used to select risk scale predictors. Items that predict may also carry with them value judgements or ethical connotations that threaten the face validity of the instrument and may erroneously imply certain motives to its developers or users. These concerns cannot, and should not, be ignored or discounted. The obvious example of such a dilemma for risk scale developers in juvenile justice is the use of race (and/or ethnicity) as a predictor of recidivism (or program failure). The fact is that most juvenile justice research has found a correlation between race and negative outcomes. Knowing nothing else about a youth, one would be on statistically-sound footing to predict that minority youth are more likely to recidivate than white youth. However, few today would argue that a child's race is causally related to his likelihood of recidivism. Indeed, such a use of race would violate our sense of equal justice and equal protection under the law. Recidivism differences may be correlated with race but they are not caused by race. The recidivism differences are caused by attributes that are themselves correlated with race (e.g., poverty, school failure, a high proportion of unsupervised time in a day, the levels of community disruption, and the amount of police surveillance in the community). Statistically, if these items were available for each child, they could be incorporated into the pool of potential risk scale items and race would "drop out" as a predictor in the statistical analysis (i.e., found not to be a significant predictor of recidivism). However, when they are not available, their predictive power will be partially captured by a race variable that will then show a significant correlation with recidivism. In other words, when the range of available data is limited, race will "stand in for" the absent variables that are causally related to recidivism.

So what are risk scale developers to do? Statistically, including the race variable improves the predictions made by the risk scale. Ethically, excluding the race variable prevents the assignment of risk scores that may inappropriately lead to greater sanctions for minority youth. There is a solution that enables risk scale developers to accommodate both concerns. This solution is central to the work that follows. Briefly, the authors suggest that risk scale developers include a race variable in the early stages of risk scale development, but then remove the race variable from the published instrument. The authors argue that some statistical methods used in the development of risk scales enable the unique (e.g., independent) effect of race to be removed from the prediction process. In fact, the authors argue that unless race is mathematically included in the initial steps of the development of a risk scale when race is found to correlate with the criterion measure, racial bias cannot be removed from the resulting risk scale, but remains unobtrusively beneath the surface influencing each risk scale score. As discussed on page 31, mathematically removing the race variable from the published risk instrument treats all youth as having the same race. Ethically, statistically removing the influence of race on risk scores goes a long way to assure users that race is not a factor in the assignment of these scores. Practically speaking, in order to avoid having racial biases in the final product, the developer must use the race variable and then remove it.

As you read the material that follows, be assured that while race was found to be a predictive factor when using various statistical techniques to develop a risk scale, there is no claim that race causes recidivism. More importantly, it is the clear intent of this work to demonstrate how the influence of race (and other invidious variables) can and should be systematically removed from risk scales by those tasked to develop these important decisionmaking aides. The risk scales presented in the report include race as a predictive factor. These prototype scales are presented as models of those experimental scales that will be prepared during the development stages and used in the validation phases, to be seen and used only by the technical staff tasked with developing the scales. In the end, when the scales move from development into field-testing and then into day-to-day use, the race variable would not be a part of risk scales, either explicitly as an item on the scale or indirectly as a spurious correlation with other scale items. Part of the real value of this work is to demonstrate the method by which race (and other such variables) may be initially used to create the prediction scale, but then removed so that the vestiges of discrimination are removed from the final scale. As the discussion on page 31 notes, such a process is feasible when multiple regression techniques are used, and the use of such a multiple step process (first using race as a predictor, then removing it from the scale) can assist in creating prediction scales that do not exacerbate the issues of minority overrepresentation in contact with the justice system.

Acknowledgments

The permission of the Maricopa County Juvenile Court (Phoenix, AZ) and its Juvenile Justice Center was much appreciated. Terrence A. Finnegan prepared the data file from collection of the National Juvenile Court Data Archive maintained by the National Center for Juvenile Justice. David P. Farrington reviewed the manuscript and provided many helpful suggestions. Nancy Tierney prepared the manuscript for publication. We also valued the support and advice of Joseph Moone of the Office of Juvenile Justice and Delinquency Prevention, who is Program Monitor for the National Juvenile Justice Data Analysis Project (grant number 99–JN–FX–K002).

Table of Contents

List of Tables

List of Figures

Preface

This report is meant to help juvenile courts develop practical risk-screening instruments. Courts increasingly are using some method of risk classification to assist in assignment of youth to differential service/supervision programs. A comparison of commonly used or advocated risk classification methods may provide courts with guidance in selecting a method for developing a screening instrument. This report assesses the strengths and weaknesses of several prediction methods in the context of juvenile courts' risk classification needs, based on analyses of one court's data that were submitted to the National Juvenile Court Data Archive maintained by the National Center for Juvenile Justice.

Although the authors hope such advice will be useful, they recognize that giving it is problematic. What works best in one situation or with one set of data may work less well in another situation or with another set of data. The authors have sought to offer a basic description of the main statistical and practical problems associated with the methods compared and to provide a fair assessment of the advantages and disadvantages of each method. These assessments are based largely on the predictive validity of classification instruments similar to those that a court might devise by using the methods compared. Because the emphasis is on practice rather than on statistics, the authors often simplified concepts in the interests of helping courts devise clear classification procedures that are easily understood and used. In comparing classification methods, the authors placed major importance on the validity of the measures as a court might actually use them.

In order to focus on comparing methods of combining predictors into a single instrument for risk classification, the authors ignore a number of issues that are of great importance in prediction, including concerns related to sampling, reliability, and discrimination in predictor variables and in the criterion. Selection problems are influenced not only by these concerns and by issues of validity, but also by the proportions of a population to be selected and the shapes of distributions. In addition, the authors devote little attention to the subject of "what to predict" (i.e., the criterion) and do not address the importance of the base rate. Lastly, except for basing conclusions on results derived from a validation sample, the authors do not consider general problems associated with validation studies or with measurement of the accuracy of prediction. This report provides references to discussions of these topics.

Introduction

The juvenile courts are not alone in their need to classify persons on the basis of predicted future behavior. College administrators and admissions committees, personnel managers, parolling authorities, medical practitioners, and others must attempt to predict future behaviors when making decisions about people. The problem of prediction is central to all behavioral science, and it is at the core of decision-making problems in many areas of life.

Because of this fundamental importance, a great deal of effort has been devoted to establishing means of predicting specific behaviors (or events related to these behaviors). The fields of juvenile and criminal justice have been at the forefront of much of this effort and have produced a long history of research and a rich literature on predicting offense behaviors.[1]

Prediction methods in juvenile or criminal justice have various utilities, and research workers may have different purposes in developing risk classifications. Major benefits from prediction research include such contributions as increasing understanding, testing hypotheses, doing research on effectiveness of treatment, and planning programs for different categories of offenders. Each application raises different problems for comparative assessments of methods. This report focuses only on applications used to classify youth into a small number of groups according to risk as a part of the ordinary operating procedures of a juvenile court. It compares the various statistical methods by assessing the validity of the resulting simple classification instruments, although it also reports the validity of the prediction methods on which they are based.

Because this report emphasizes the comparative validity of the various classification devices, it ignores or mentions only briefly other topics of great importance in comparisons of prediction methods. These include problems associated with the following:

- The relative efficiency of informal clinical or other subjective ("in-the-head") predictions and more formal ("actuarial" or "statistical") methods of prediction.[2]
- Base rates.[3]
- Unreliability.[4]
- Selection of criteria (outcomes) to be predicted.[5]
- Measurement of the accuracy of prediction.[6]
- Validation issues.

---

[1] For histories and reviews of methodological issues, see H. Mannheim and L. Wilkins, *Prediction Methods in Relation to Borstal Training* (London, England: Her Majesty's Stationery Office, 1955); D.M. Gottfredson, "Assessment and Prediction Methods in Crime and Delinquency," in *Task Force Report: Juvenile Delinquency and Youth Crime* (Washington, DC: U.S. Government Printing Office, 1967); H.F. Simon, *Prediction Methods in Criminology* (London, England: Her Majesty's Stationery Office, 1971); D.M. Gottfredson and M. Tonry (eds.), *Prediction and Classification: Criminal Justice Decision Making*, vol. 9 of *Crime and Justice: A Review of Research* (Chicago, IL: The University of Chicago Press, 1987).

[2] For a recent review, see P.E. Meehl and W.M. Grove, "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy," *Psychology, Public Policy, and Law* 2, no. 2 (1996).

[3] See P.E. Meehl and A. Rosen, "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores," *Psychological Bulletin* 52 (1955).

[4] For discussions, see D.T. Campbell and J.C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago, IL: Rand McNally, 1963); L.J. Cronbach, *Essentials of Psychological Testing* (New York, NY: Harper Brothers, 1960); and E.F. Cureton, "Validity, Reliability, and Baloney," in *Problems of Human Assessment*, edited by D.M. Jackson and S. Messnick (New York, NY: McGraw-Hill, 1967).

[5] See D.M. Gottfredson and M. Tonry (eds.), *Prediction and Classification: Criminal Justice Decision Making,* vol. 9 of *Crime and Justice: A Review of Research* (Chicago, IL: The University of Chicago Press, 1987).

[6] See S.D. Gottfredson and D.M. Gottfredson, "The Accuracy of Prediction Models," in *Research in Criminal Careers and "Career Criminals,"* vol. 2, edited by A. Blumstein, J. Cohen, J.A. Roth, and C.A. Visher. (Washington, DC: National Academy Press, 1986).

These topics are interrelated, and full consideration of any one requires consideration of the others. Such consideration is not within the scope of this report.

Purpose of This Report

This report is directed to persons who are responsible for developing and using risk classification instruments in juvenile courts and is intended to help them classify youth into a small number of risk groups as an aid to program assignments. These individuals must decide which of several competing, commonly used statistical procedures will work best in selecting and combining variables for prediction. Given a data set that records characteristics of youth at referral, they must decide which characteristics will be most useful for risk classification. They must also decide how to combine the characteristics selected as predictors, usually into a single scale that will be the basis for the classification tool. When risk classification is actually put into practice, the validity of the classification tool is more important than the validity of the prediction instrument on which it is based. Although some predictive efficiency may be lost in the conversion from the prediction instrument to the simpler classification tool, it is the validity of this tool, as it might be used in practice, that is of utmost importance.

Prior Studies

Empirical comparisons of the predictive efficiency of different statistical methods for selecting and combining predictors have been carried out occasionally; most have been based on adult offender samples.[7] Generally, statistical methods that are theoretically less powerful and computationally and procedurally rather simple have been found to be equal or superior in predictive validity to the more complex and theoretically better methods. Debate continues, though, about the most efficient (most valid, least costly, most operationally useful) way to develop a risk classification instrument.

Procedures

This report compares the validities of several instruments derived by three markedly different types of statistical methods that are the most commonly used for selecting and combining predictor variables to yield risk scores, such as the risk of a new referral to the court. The first procedure is based on the assumption that a linear equation can be found that adds the scores of the predictor variables to provide a total score related to the outcome of interest (such as a new referral to the court). Two instruments based on this procedure are compared—in the first, the values of the predictor variables are simply added; in the second, weights are applied to each variable, and the weighted scores are added. The second procedure uses a more statistically sophisticated model, based not on the risk of failure but rather on the logarithm of the odds of "succeeding" or "failing" associated with each predictor variable. The third procedure is quite different from the others. It uses a "clustering" or group classification method to subdivide a sample into risk subgroups directly rather than on the basis of scores on a single scale. A fourth approach, which combines features from two of these procedures, is also considered. This approach first subdivides the sample into more homogeneous subgroups and then develops linear equations separately for each subgroup.

Throughout this report, the statistical solutions are used to devise classification tools, because that is how the solutions are commonly used in practice. As these tools need to be not only credible but also easy to use, the report introduces some simplifications. For example, in practice, classifications of youth usually define only a small number of groups (no more than five and often only three). Therefore, emphasis is placed herein on comparisons of validity based on

[7] D.M. Gottfredson et al., *The Utilization of Experience in Parole Decision-Making: Summary Report* (Washington, DC: U.S. Government Printing Office, 1974); H.F. Simon, *Prediction Methods in Criminology* (London, England: Her Majesty's Stationery Office, 1971); S.D. Gottfredson and D.M. Gottfredson, *Screening for Risk: A Comparison of Methods* (Washington, DC: U.S. Department of Justice, National Institute of Corrections, 1979); D.P. Farrington and R. Tarling, "Criminological Prediction: The Way Forward," in *Prediction in Criminology*, edited by D.P. Farrington and R. Tarling (Albany, NY: State University of New York Press, 1985); W. Wilbanks and M. Hindelang, "The Comparative Efficiency of Three Predictive Methods," app. B, in D.M. Gottfredson, L.T. Wilkins, and P.B. Hoffman, *Summarizing Experience for Parole Decision Making* (Davis, CA: National Council on Crime and Delinquency Research Center, 1972).

classifications into just five groups, although a substantially larger number of classifications could be devised with the statistical methods considered.[8]

At the end, the authors provide a few recommendations that focus on practical aspects of risk classification.

General Limitations of Risk-Screening Instruments

The instruments discussed in this report have been developed and validated with respect to specific criteria, using available data in a specific jurisdiction during a specific time period. As a result, any generalizations—to other outcomes of interest, with modifications of the item definitions used, or to other jurisdictions or time periods—should be regarded as uncertain.

---

[8] The emphasis on a relatively small number of categories limits the utility of the comparisons herein for other applications of prediction methods. When a prediction method results in a larger number of categories (commonly, a continuously distributed score), there are more choices of cutting points—a factor that often is important (particularly for selection problems), in part because the distributions usually are markedly skewed. See L.J. Cronbach and G.C. Gleser, *Psychological Tests and Personnel Decisions* (Urbana, IL: University of Illinois, 1957). To simplify the comparisons, this problem has been ignored.

General Method[9]

Steps Followed in Prediction Studies

Five steps should be followed in developing and using a risk assessment (prediction) measure: (1) defining what is to be predicted (the criterion categories), (2) selecting the predictors, (3) measuring relations between the predictors and the criterion categories, (4) verifying the relations found (validating the method), and (5) applying the method.

Defining Criterion Categories

The first step is to establish the criterion categories of "favorable" or "unfavorable" performance, or "new offense," or some other event. This involves defining the behavior or event to be predicted and developing procedures for classifying persons on the basis of their performance in regard to that behavior or event. This step is of utmost importance, because it defines the standard for selecting predictors and testing the validity of results. In addition, it sets limits to generalization.

Selecting Predictors

Second, the attributes or characteristics on which the predictions may be based are selected and defined. These "predictor candidates" are expected to relate significantly to the criterion categories. (Critics of prediction methods often argue that the procedures ignore differences among individuals. However, such differences, often assumed to be a source of error in other problems, are in fact the basis for any prediction effort. If the persons studied are alike with respect to the predictor candidates, no differential prediction can be made. If they are alike with respect to the criterion categories, there is no prediction problem.)

Measuring Relations

The third step is to determine the relations between the predictor candidates and the criterion categories (and, for some methods, among the predictor variables) in a sample representative of the population for which inferences are to be drawn. These relations ordinarily are measured by the Pearson product moment correlation coefficient (for continuous variables), the point biserial correlation (an estimate of the correlation coefficient for one continuous and one dichotomous variable), or the phi coefficient (an estimate of the correlation for two dichotomous attributes).[10] A representative sample usually is sought by random selection from the population to which generalizations are to be made. Any haphazard sample is apt to be biased, and procedures for sample selection should ensure that each individual in the population has an equal chance of inclusion in the sample.

Verifying Relations

Fourth, the relations found in the original sample must be verified by testing the prediction procedures in a new sample, or samples, of the population. Although this verification (referred to as cross-validation) often is omitted, it is a critical step. Without it, there can be little confidence in the utility of a prediction method for any practical application.

Applying the Method

Fifth, the prediction method may be applied in situations for which it was developed, provided that the stability of predictions has been supported in the cross-validation step and appropriate samples have been used.

---

[9] This section is adapted from D.M. Gottfredson (ed.), *Juvenile Justice with Eyes Open* (Pittsburgh, PA: National Center for Juvenile Justice, 2000).

[10] In the case of dichotomous variables, other indices, such as the odds ratio of "successes" and "failures," may be preferred; see, e.g., D.P. Farrington and R. Loeber, "Some Benefits of Dichotomization in Psychiatric and Criminological Research," *Criminal Behavior and Mental Health* 10 (2000):100–122.

Simplification Practices and Assumptions

To simplify procedures, prediction studies commonly engage in certain practices and make certain assumptions. Some practices that are not strictly justified on the basis of measurement considerations are followed nevertheless. For example, numerals may be used to represent differences in attributes (variates that are dichotomous)[11] because this serves a practical purpose. In addition, qualitative attributes may be assigned numbers, which are then used in statistical computations. These practices no doubt will continue, because the results are useful. The measurement and scaling problems involved in these practices, however, should not be ignored, because improving the measurement of individual differences by examining their hypothesized relations with later delinquency behavior is one means of improving current abilities to predict risk.

A second example of simplification is the assumption that the criterion magnitudes are linear functions of the predictor variables. This assumption is common in prediction studies, including some (but not all) of the studies considered in this report. Improvements of prediction methods, however, may also require study of nonlinear relations.

A third example of simplification is the common assumption that the relations among predictors and between predictors and the criterion hold for subgroups of a heterogeneous population. Prediction methods might be improved, however, by separate study of various subgroups of youth.

In addition, prediction studies commonly use certain statistical methods in a way that ignores generally accepted requirements for the appropriate use of the methods. One example of this practice is the use of multiple linear regression when the criterion is dichotomous rather than continuous. This practice probably also will continue because, although questionable, it produces results that often are useful nevertheless.

These considerations may point to avenues for improving risk classifications, but they do not preclude the use of prediction methods. That is because results of these methods are constantly being judged by their ability to discriminate criterion classifications in new samples. How well a method works is the ultimate test.

Requirements for Predictors

Two requirements for any predictor candidate (besides an expected relation to the criterion) are *discrimination* and *reliability*.

Discrimination

Discrimination is reflected in the number of categories to which significant numbers of persons are assigned. A classification that assigns all persons to one category does not discriminate. The best procedure is one that provides a "rectangular" distribution of categorizations of persons for the sample—i.e., one with about equal numbers of persons in each category. When only two categories are involved, the most discriminating item is one with half of the persons in the sample in each category.

One consequence of the discrimination requirement is that predictor items found useful in one jurisdiction may not prove useful in another. For example, a drug-involvement item that is a useful predictor in a jurisdiction where many youth have histories of drug abuse may be of no use in another jurisdiction where few youth use drugs. The discrimination requirement points to the need for cross-validation studies in various jurisdictions. It also points to one hazard of the untested acceptance of prediction measures developed outside a given target jurisdiction.

Considerations related to sampling and the concept of validity suggest similar conclusions. An item may effectively discriminate within a sample but not be a valid predictor for a given criterion, but an item with good discrimination within the sample may prove to be more useful than a valid predictor with little discrimination.

[11] This report generally uses the term "attribute" to refer to items or characteristics that are dichotomous (e.g., white and not white) and the term "variable" to items or characteristics that are continuous (e.g., age).

Reliability

Reliability refers to the consistency or stability of repeated observations, scores, or other classifications. If a procedure is reliable, then repetitions of the procedure lead to similar classifications. Valid prediction is not possible with completely unreliable measures, but all measurement is relatively unreliable. This means that some variation is always to be expected. It also means that if valid prediction is demonstrated, then the predictor attributes are not completely unreliable. Because the main interest in a prediction method is in how well it works, more importance is attached to the concept of validity than to that of reliability. This does not mean, however, that the issue of reliability is unimportant. Making predictor variables more reliable is another means of improving prediction.

Probability and Validity

Because no prediction of future behavior, achieved by any means, can be made with certainty, a statement of degree of probability is a more appropriate prediction. Predictions are properly applied to groups of persons who are similar in terms of some set of characteristics, rather than to individuals. In any prediction problem, individuals are assigned to classes, and then statements are made about the expected performance of members of the classes. For specific classifications of persons, the expected performance outcomes ought to be those that provide the most probable values for those classes.

The validity of a prediction method refers to the degree to which earlier assessments are related to later criterion classifications. The question of validity asks how well the method works. Any prediction method, however, may be thought of as having or lacking not one but many validities of varying degrees. For example, an intelligence test might provide a valid prediction of high school grades. It might provide a much less valid, but still useful, measure of expected social adjustment, and it might have no validity for the prediction of delinquency. The validity of prediction refers to the relation between an earlier assessment and a specific criterion measure, and it is dependent upon the particular criterion used. A prediction method has as many validities as there are criterion measures to be predicted.

Evidence of validity with respect to a specific criterion of interest in a particular target population obviously is necessary before any practical application of the method in that population. Nevertheless, juvenile justice agencies sometimes use risk measures developed and tested elsewhere (or simply created without following all of the steps described above) without first obtaining this crucial validation. The practice of not validating a method in samples from a given target population should be discouraged.

Predictor Candidates

Predictor items may be obtained from self-reports of the youth concerned, reports of observers or judges, records of past performance, other elements of the youth's social history, or direct observation by the person making the prediction. Predictor items may include psychological test scores, measures of attitudes or interests, biographical items, ratings—or, indeed, any data about the subject.

In juvenile justice system agencies, predictor items commonly are data coded from social histories, complaints, or other records pertaining to youth. Many studies have found the following items to be related to typical criterion classifications such as new offenses or referrals:

- **Age.** The relation of age to later delinquency and crime is well established. The likelihood of offending increases sharply with age in the early teens, reaching a peak in the late teens or early twenties, and then declines continuously with age. Another age variable often found useful in predictions for youth populations is age at first complaint, referral to the courts, or delinquency.
- **Gender.** Because girls generally are found to be better risks than boys are (e.g., less likely to commit subsequent offenses), gender often is a useful predictor.
- **Record of complaints or offending.** Most measures of prior delinquency may be expected to be related to typical criterion classifications. These measures include number of prior complaints or court referrals, prior petitions, previous adjudications, prior probation violations, and other indicators of previous difficulty with rules violations.

- **Classifications of type of complaint.**
- **History of drug or alcohol abuse.**

Although these items are commonly found to be useful predictors, any jurisdiction that is considering using an item as a predictor should test the relation of the item to the jurisdiction's criterion classification in a sample drawn from its own target population. The usual relation may not be found, or the strength of the relation may be different from that found in other jurisdictions.

Methods of Combining Predictors

A number of predictor items may be used in establishing a single prediction score or in establishing a predictive classification of persons into groups (i.e., a "risk classification," as often used in juvenile justice). The items may be selected and combined in various ways. All or some of the predictor items may be combined without weighting, some procedure for selecting and weighting the items may be used, or some other type of classification based on the items may be adopted. How to determine which procedure works best is the primary focus of this report and the question addressed in the comparisons that follow.

Linear Additive Models

Arbitrary Weights

The most common method of constructing a prediction score is to use many predictor items and to arbitrarily assign each item the same weight. The procedure either uses dichotomous attributes as predictors or reduces variables to attributes by collapsing a continuous distribution to two categories. Each item found to be predictive is assigned one point regardless of the strength of its association with the criterion. The sum of the points assigned in this way is the total score.

Thus, the model is as follows:

$$Y = (a) + x_1 + x_2 + x_3 + \ldots + x_i$$

where Y is the prediction score related to the expected value of the criterion, a is an arbitrary constant that may or may not be included, and x_i is some predictor attribute.

This method often is referred to as the Burgess method, after its originator in parole prediction studies.[12] It is a simple and popular procedure, but it may be inefficient because it does not take into account the interrelations among the predictor items. Nevertheless, if the number of items is large, there are good arguments besides simplicity for using this procedure. If the number of predictor items is large, the effect of differential weighting that takes into account the overlaps among items tends to become unimportant. If the number of items is relatively small, however, weighting the items by one of the available empirical methods can improve the efficiency of prediction.

If all items are weighted equally—i.e., each is assigned one—the equation above shows that the total score is simply the sum of all the items. In juvenile justice risk classification studies, however, the procedure often is modified by a more complex (though still arbitrary) weighting of the predictor items. These weights are arrived at subjectively or on the basis of several arbitrary cutting scores on a continuous variable such as "number of prior referrals to the court."

Then the model is:

$$Y = (a) + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ldots + \lambda_i x_i$$

where Y and a are as before, x_i is some predictor attribute, and λ_i is some arbitrary weight. For example, a judgment may be made that some specific offenses predict reoffending, and these offenses may be subjectively assigned more weight than are others. Or, an item such as "number of prior referrals" may be classified into several groups such as "high," "medium," and "low" and then scored as 3, 2, and 1 respectively. The weights are called "arbitrary" because, unlike the procedures described in the next section, they are not based on the data—i.e., they are not derived empirically. Although often used in risk classification studies, these more complex arbitrary weighting procedures cannot be recommended; therefore, the Burgess method as presented in this study uses only unit weights (i.e., one point for each item).

The advantages of the Burgess method include its simplicity, its ability to produce an easily understood

[12] E.W. Burgess, "Factors Determining Success or Failure on Parole," in *The Workings of the Indeterminate Sentence Law and Parole in Illinois*, edited by A.A. Bruce (Springfield, IL: Illinois State Parole Board, 1928).

scale that has face validity, and the fact that the resulting instrument is easy to score and interpret.[13] There are, however, several disadvantages: the intercorrelations—i.e., "overlaps" among the items—are ignored; there is no basis for assuming that the arbitrary weight of one will provide optimal discrimination of the criterion groups; and there is no way to tell which items are actually redundant.

Methods of Weighting Items

The second set of methods in common use is based on some procedure for determining statistically how to weight the predictor items to arrive at more efficient prediction. There are three often-used procedures, each with some advantages and disadvantages. These procedures are called multiple linear regression, discriminant analysis, and logistic regression.

Multiple Linear Regression

This procedure provides a linear equation (a set of weighted items added together) with weights assigned to minimize the squared deviations of observed and expected criterion values. It takes into account the intercorrelations among predictor variables as well as the relation of each predictor to the criterion. This method is most appropriate when the criterion measure is composed of continuous scores on an interval scale—e.g., scores providing a measure of overall social adjustment. It often is used with a dichotomous criterion such as "success" or "failure," but such use violates some statistical assumptions and has disadvantages as discussed below.

The model is as follows:

$$Y = (a) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ \ldots + \beta_i x_i$$

where Y is the predicted value of the criterion, a is a constant (the intercept of the regression line), x_i is a predictor variable and β_i is its coefficient (weight) estimated from the data.

One advantage of this procedure is that it results in a theoretically optimal weighting of the predictors when they are used in combination with each other. Another advantage of the procedure is that it makes it possible to estimate each predictor item's relative contribution to explaining variability in the outcome criterion—thereby facilitating selection of predictor items. Moreover, the standardized coefficients (weights) serve to indicate the relative importance of each item in the context of all others in the equation. Another possible advantage of this procedure is that a method exists for eliminating the effect of unwanted "invidious" predictors, as discussed later in this report. Some classifications that result from this procedure may have substantial face validity, but others may not (i.e., it may not be apparent why the variables are weighted as they are).[14]

Disadvantages of this procedure include its assumptions that (1) predictor variables are linearly related to the criterion measure and (2) variables are normally distributed and measured on an interval scale. In addition, multiple linear regression usually develops weights that are derived from the correlation matrix (for the predictor candidates and the criterion) using the total sample of subjects; by doing so, there is an implicit assumption that the relations are the same for all subgroups of youth, which may be false. Another disadvantage arises when, as is usual for risk classification in juvenile justice applications, the criterion classification is not a continuous measure but instead is composed of two or more nominal classes (i.e., groups with numbers assigned solely to give them names). In that case, logistic regression (discussed below), or a related model, would be theoretically superior to multiple linear regression. Another disadvantage, arising from the violation of statistical assumptions underlying multiple linear regression, is that expected values resulting from the analysis may be greater than one or less than zero, although the criterion is scored either zero or one. In addition, it is unreasonable to assume, as the model requires, that the errors are normally distributed.

[13] Moreover, a simple, equally weighted linear additive model has been found to be as good as, and in some ways better than, multiple linear regression. See H. Wainer, "Estimating Coefficients in Linear Models: It Don't Make No Nevermind," *Psychological Bulletin* 83 (1976).

[14] Even linear models that use random regression weights have been found to perform substantially better in predictions than humans do. See R.M. Dawes and B. Corrigan, "Linear Models in Decision-Making," *Psychological Bulletin* 81 (1974).

Discriminant Analysis

This procedure is closely related to the multiple linear regression procedure, but the weights are determined in such a way that the criterion classification groups are maximally separated in terms of the average values for the prediction scores in relation to their pooled standard deviations. If the criterion classifications provide only two groups (e.g., scored zero and one), then this method is mathematically equivalent to multiple linear regression.[15] It can, however, be used with more than two groups, making it possible to find several equations (up to one less than the number of criterion groups) that optimally separate the groups. The advantages and disadvantages of discriminant analysis parallel those of multiple linear regression.

Logistic Regression

When the criterion is composed of only two groups, as is usual for risk classification in juvenile justice, logistic regression is the theoretically best approach. The multiple linear regression and discriminant analysis procedures discussed above present statistical difficulties when the criterion can take on only two values (such as "success" or "failure")—the assumptions that underlie the use of these methods are violated in such applications. The logistic regression procedure requires fewer assumptions but provides an estimate of the likelihood that an event will occur. The weights assigned are those that make these estimates most likely. The equation can be written to show how the odds that an event will occur change with changes in each variable included. This study uses logistic regression and bases the resulting classification tool on the weights that generate the probabilities.[16]

This procedure seems more complex, but the underlying rationale is straightforward. Estimates of the probability of an event occurring are made directly. Whereas in the multiple linear regression model discussed above, weights for predictors are estimated on the basis of "least squares," in logistic regression they are estimated by a method of "maximum likelihood"—i.e., the coefficients selected are those that make the observed results most likely. The weights can be interpreted as a change in the logarithm of the odds ratio associated with a one-unit change in any predictor. It is easier to think in terms of odds rather than log odds, though, and the equation can be written in terms of odds, resulting in an additive model.

The model is as follows:

$$\Omega = e^z/(1 + e^z)$$

where Ω is the probability of the event, e is the base of the natural logarithms (about 2.718), and z is the linear combination

$$Z = (a) + B_1x_1 + B_2x_2 + \ldots + B_ix_i$$

where a is a constant, the Bs are coefficients estimated from the data, and xs are the values of the predictors.[17]

The equation may be written in terms of the log of the odds ratio, which is called the logit:

log (Probability of event/Probability of non-event) = log $(p/(1-p)) = Z$

The B coefficients show how the odds change with each predictor variable. If expressed in terms of *odds*, rather *log odds*, the equation may be written as follows:

$$Z = e^{(B + B_1x_1 + B_2x_2 + B_3x_3 \ \cdots \ + B_ix_i)}$$

Raising e to the power B for any predictor variable shows the factor by which the odds change when that predictor variable is raised by one unit.

The main advantage of logistic regression is that few statistical assumptions are required for its use. In addition, it generates probability values that, unlike those generated by regression with a dichotomous criterion, are constrained between zero and one. A disadvantage of logistic regression is that persons unfamiliar with statistics may find it more difficult to interpret. The resulting classification may or may not have much face validity for administrators. Another disadvantage

---

[15] For this reason, discriminant analysis is not included in this report's comparisons. See, for example, O.R. Porebski, "On the Interrelated Nature of the Multivariate Statistics Used in Discrimination Analysis," *British Journal of Mathematical and Statistical Psychology* (1966).

[16] For the classification proposed, the probabilities are not needed.

[17] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (New York, NY: John Wiley and Sons, 1989).

is that complete data for each subject are needed to calculate weights; thus, missing data cause great problems.

Configural or Clustering Methods

A variety of configured or clustering methods for grouping youth according to combinations of predictor items also have been used in delinquency prediction. The most common have been those called "predictive attribute analysis," "association analysis," and "configural analysis."[18] This study uses a predictive attribute analysis, which is the clustering method most directly related to the prediction problem and also is the configural procedure that generally has shown the best results in related studies.[19]

Predictive attribute analysis is a hierarchical, divisive clustering method. First, the single best predictor among all the available (attribute) candidates is found. Then, after dividing the whole sample into two subgroups on the basis of that predictor, the single best predictor is found within each of the subgroups. The groups are again divided, and the process is repeated for four subgroups. This process continues until some "stopping rule" ends it. A common rule is to stop when no further discriminators are found at a predetermined level of confidence, such as .01 or .05. If the sample studied is large, this process may generate a large number of subgroups; thus, the procedure may be stopped when identification of further subgroups does not seem operationally useful (even though this requires a subjective judgment).

Advantages of predictive attribute analysis are that it requires few statistical assumptions and usually produces results that have considerable face validity. Another possible advantage is the discovery of any interactions of predictors that may affect the outcome of the prediction. Another advantage is that, unlike other methods, predictive attribute analysis takes into account the heterogeneity of the subjects and, in fact, builds on these subgroup differences.

One small disadvantage of predictive attribute analysis is that computer analysis programs for the entire procedure are not readily available (although they are for the other methods discussed).[20] Another disadvantage may be that, in predictive attribute analysis (unlike other methods), the final classifications for operational use are found directly rather than by the use of cutting scores for a continuous distribution—a factor that limits the extent to which group sizes for differential supervision programming may be manipulated for administrative advantages. Another issue is that termination of the procedure depends on arbitrary stopping rules, including the level of significance required for continuation of the process. Yet another concern is that attributes may be very close (or even equal) in apparent predictive power except, for example, in the third decimal place of the measure of the relation to the criterion. In such a case, the choice between two attributes is actually quite arbitrary—usually the item thought to be more reliable is selected.

"Bootstrap" Methods

"You can't lift yourself up by your own bootstraps" is an adage sometimes referred to by statisticians. They mean that no statistical analysis can improve inadequate data (i.e., "GIGO, garbage in, garbage out") and also that no amount of extensive statistical manipulation can find results not present in the data ("GISGO, garbage in, sophisticated garbage out").

Nevertheless, there are sound reasons to expect that a statistical approach sometimes called "bootstrapping" might improve the procedures discussed above. If a population contains subgroups that differ from one another in the way the predictor candidates are correlated or in their relations to the criterion, the prediction method may benefit from a separate analysis of those subgroups. This report explores the use of one such bootstrap procedure.[21]

[18] L.T. Wilkins and P. MacNaughton-Smith, "New Prediction and Classification Methods in Criminology," *Journal of Research in Crime and Delinquency* 1(1954):19–32. For a description of other clustering methods, see T. Brennan, "Classification Methods in Criminology," in D.M. Gottfredson and M. Tonry (1987), *supra*, note 1, 201–248.

[19] See, for example, S.D. Gottfredson and D.M. Gottfredson, (1979), *supra*, note 7.

[20] All analyses described in this report were completed using Statistical Package for the Social Sciences (SPSS) 6.0; see M.J. Norusis, *SPSS for Windows* (Chicago, IL: SPSS, Inc., 1993).

[21] The term "bootstrap" is sometimes used in other ways, e.g., in reference to analyses of repeated random samples to determine stable estimates.

Data Used

Sample

This study was based on the records of 9,476 youth ages 8–15 who were referred to the Juvenile Court of Maricopa County, AZ, in calendar year 1990. These records were provided by the court to the National Juvenile Court Data Archive which is maintained by the National Center for Juvenile Justice. Records of 16- and 17-year-olds were excluded because these youth would "age out" of the juvenile court jurisdiction (18) before the end of the study's 2-year followup period.[22, 23]

Construction and Validation Samples

Approximately a 20-percent random sample of records (1,924) was selected to compose the construction sample—i.e., the sample on which the prediction instruments were developed. The remaining 80 percent (7,552) were reserved for a validation sample—the sample to be used for verification.

The selection of the construction sample size was based on several considerations. The larger the construction study sample, the greater the stability to be expected upon validation; however, many smaller juvenile courts have relatively small samples available for study, and this study sought to simulate the application of methods that might be meaningful to all. Experience shows that when the linear models are used, construction samples of about 500 to 1,000 subjects usually result in adequate stability when tested on validation samples. Larger samples are needed, however, for predictive attribute analysis (in which samples quickly become smaller as the subdivision process advances) and for the bootstrap method. Thus, the selection of the 20-percent construction sample seemed a reasonable compromise for taking into account the need for a sufficiently large sample, the limitations likely to be encountered by smaller courts, and the problem of sample size reduction for some of the methods to be assessed.

Jurisdictions with smaller populations available for sample selection might find it advantageous to split the sample randomly in half to provide construction and validation samples. This is a common procedure. However, because prediction implies looking to the future and because conditions may change over time, a better approach is to use all available cases in one year for the construction sample and cases from a later period for the validation sample.

Criterion

The criterion for this study was defined as "return to the juvenile court with a new complaint within 2 years." This criterion was selected because it is commonly used in juvenile courts, typically has adequate discrimination in samples, and is a reasonable indicant of relevant youth behavior subsequent to a referral to the court. In this study, 42.8 percent of youth in the construction sample and 44.7 percent of youth in the validation sample had at least one new referral within 2 years.

Attributes

Attributes were scored as zero or one. Some items—such as whether there were any prior referrals to the court—are natural attributes. Others result from collapsing continuous variables—such as the number

---

[22] The age variable usually proves helpful in prediction, but this study's exclusion of 16- and 17-year-olds limits its variability. Thus, the study may find only a weak relation of age to the criterion, even though the relation would be strong were the full age range considered.

[23] It is desirable that all sample members have equivalent exposure to risk of the unfavorable outcome. Although it is technologically feasible to follow youth through any detention, commitment to State confinement, and charges in adult courts, such tracking rarely has been accomplished in the juvenile courts to date. The lack of the kind of data that would result from such tracking reduces the reliability and validity of the criterion classifications, thereby reducing the efficiency of prediction.

of prior referrals—into only two categories.[24] Because some analyses require attribute data, all variables were made into attributes (even though they could be used alternatively as continuous scores when the analytic procedures permitted that). Selected attribute predictor candidates are listed in table 1, which also shows the proportion of the construction sample possessing each attribute, the percentage of new referrals among youth possessing the attribute, and the attribute's relation to the criterion. (The last is measured by the phi coefficient, an estimate of correlation for fourfold tables—i.e., two rows and two columns.)[25] Table 1 lists the items in order of the magnitude of the phi coefficients.

Fourfold tables can, of course, be produced for each listed attribute from the data in table 1. Table 2 shows an example of a fourfold table for the first listed attribute—"any prior referrals."

Variables

Table 3 shows selected items that are distributed continuously, along with a measure of their relations to the criterion: the point biserial correlation.[26] The distributions tend to be markedly skewed to the right.

[24] An example of creating an attribute from a continuously distributed variable is this study's use of the classification "three or more complaints." The distribution of the variable "number of complaints" was dichotomized in this way because subjects who had more than two complaints experienced new referrals within 2 years at rates higher than the base rate (i.e., the overall rate of new referrals for the sample). It is common practice to examine the distributions for a continuous variable for the outcome groups (in the construction sample) and then to make a judgment about classifying the variable into an attribute. In doing so, however, the risk of capitalizing on chance variation in selecting the cutting score should be kept in mind.

[25] Phi = the square root of the quantity (chi-square divided by the number of cases), when there is only one degree of freedom.

[26] The point biserial correlation is an estimate of the correlation coefficient for a continuous interval level variable and a dichotomous dependent one. For simplicity, this study considered all variables listed as having equal intervals.

Table 1: Selected Attributes Included for Study, With Relation to Criterion (*N* = 1,924)

Attribute	Percent of All Youth in Sample With Attribute	Percent With Attribute Who Had New Referrals Within 2 Years	Phi Coefficient	Significance
Any prior referrals	41%	64%	.35	< .001
At-risk record *	41	64	.35	< .001
Any prior delinquent offense	37	65	.34	< .001
Three or more complaints in history**	24	73	.34	< .001
Any prior informal dispositions	37	64	.32	< .001
Any prior theft offense referrals	28	68	.32	< .001
Any prior formal dispositions	22	71	.30	< .001
Any prior adjudications	18	74	.29	< .001
Petition filed	23	61	.20	< .001
Male	70	49	.19	< .001
Any prior status offense referral	15	65	.19	< .001
Any prior adjustment	11	69	.19	< .001
Any prior cases with detention	9	73	.19	< .001
At-risk youth***	13	65	.17	< .001
Not white	39	52	.16	< .001
Grade 7 or higher	70	47	.14	< .001
Any prior violent offense referrals	4	76	.13	< .001
Mexican American	26	53	.12	< .001
Age 13 or older	78	46	.11	< .001
Auto theft or person offense	17	55	.11	< .001
Age 14 or older	60	46	.09	< .001
Referral not from law enforcement	6	60	.09	< .001
Theft offense referral	48	38	–.09	< .001
Present referral for auto theft	5	61	.09	< .001
Violence offense, present referral	4	62	.08	< .001
Person offense	12	53	.07	< .001
Detained	8	54	.07	< .003
Any prior drug offense referrals	2	67	.07	< .002
Not in school, but should be	8	55	.06	< .008
Three or more counts at referral	5	56	.06	< .004
Referral for delinquency	81	42	–.03	ns
In school	86	43	.02	ns
Status offense referral	19	45	.02	ns
Referral for drug offense	2	49	.02	ns
Parents married	22	43	.00	ns

* At-risk record: any record of referral for a drug offense or violence offense or for delinquency or status.
** Three or more complaints in history: reflects the sum of present and prior complaints.
*** At-risk youth: not in school, referred by source other than law enforcement agency (e.g., school, parents), or prior informal adjustment of alleged offense.

Table 2: New Referral Within 2 Years, Analyzed by Whether There Were Any Prior Referrals

Any Prior Referrals	New Referral Within 2 Years: No Number	No Percent	Yes Number	Yes Percent	Total Number	Total Percent
No	812	71.9%	317	28.1%	1,129	58.7%
Yes	289	36.4	506	63.6	795	41.3
Total	1,101	57.2	823	42.8	1,924	100.0

Phi coefficient = .354
Chi-square = 241
df = 1
$P \leq .001$

Table 3: Selected Variables Included for Study, With Relation to Criterion (N = 1,924)

Variable	Range	Mean	Standard Deviation	Median	Mode	Point Biserial Correlation Coefficient	Significance
Prior delinquency referrals	0–18	.96	1.90	.00	.00	.33	<.001
Number of complaints*	1–22	2.23	2.32	1.00	1.00	.32	<.001
Number of prior referrals	0–21	1.23	2.32	.00	.00	.32	<.001
Prior informal dispositions	0–17	.78	1.50	.00	.00	.30	<.001
Prior theft offense referrals	0–14	.56	1.26	.00	.00	.30	<.001
Prior formal dispositions	0–9	.45	1.06	.00	.00	.29	<.001
Prior adjudications	0–8	.29	.76	.00	.00	.27	<.001
Length of delinquency**	0–7	.65	1.26	.00	.00	.27	<.001
Prior cases with detention	0–7	.14	.53	.00	.00	.17	<.001
Prior status offenses referrals	0–8	.26	.78	.00	.00	.16	<.001
Age	8–15	13.50	1.58	14.00	15.00	.14	<.001
Prior violent offense referrals	0–3	.04	.23	.00	.00	.12	<.001
Grade in school***	2–13	7.68	1.73	8.00	8.00	.08	<.003
Prior drug offense referrals	0–3	.02	.18	.00	.00	.07	<.017
Number of current counts	1–23	1.27	.88	1.00	1.00	.05	<.011
Age at first referral	8–15	12.50	1.77	13.00	14.00	–.07	<.007

* Number of complaints = total number of complaints in history, including any present complaints.
** Length of delinquency = difference in years between present referral and first referral.
*** Grade in school was missing in 226 cases (12 percent of the sample).

Development and Validation of Operational Tools

Each prediction method under consideration is used to devise a simple form, in a process intended to simulate what might be done in a juvenile court that wishes to develop a classification tool for use in its everyday operations. Courts generally prefer a simple procedure that permits the classification of all youth into a few categories, with each category differing in the proportion of youth expected to return to the court on a new complaint. Thus, this study sought to produce forms that are easily completed and that result in the classification of youth into five groups.

In devising these classification procedures, the distributions of scores in the construction sample were examined in order to determine cutting scores that would allow placement of all cases into five groups. The cutting scores were determined by selecting points that would result in groups with approximately equal numbers of cases. The proportions of cases with each score and the percentages with at least one new referral were also examined. When distributions were markedly skewed, the five groups were far from equal in their portions of total cases classified. When cutting scores are based on apparently optimal discrimination of outcomes in the construction sample, there exists the risk of capitalizing on chance variation; the use of a validation sample, however, provides a check on this.

In an operational situation, cutting scores should be determined with a view to the objectives of the intended use of the classification instrument. If, for example, the objective is to identify a group with low probabilities of new referrals in order to place youth in a minimal supervision program, the decision about the cutting score might take into account a number of concerns: the shape of the distribution, the degree of validity of the instrument, the proportions selected at a given point, the percentage of youth with favorable or unfavorable outcomes, and the judged tolerance for risk.

In any decision of this sort, two types of error must be taken into account: misclassifying a youth as being likely to return to court with a new referral, when the youth actually "succeeds," and misclassifying a youth as being a likely "success," when the youth actually experiences a new referral.[27] Both types of error have important consequences that should be assessed in program planning when tools such as those discussed in this report are to be used.

For each method, the following sections summarize the results of the analysis in the construction sample and also provide a measure of the predictive validity found when the five-group classification "operational tools" devised from the methods were applied to the validation sample.[28] For a simple measure of validity, Cramèr's statistic is presented to describe the relation between the classification model and the criterion.[29]

---

[27] For discussion of such errors and consequences for predictive decisions, see S.D. Gottfredson and D.M. Gottfredson, "Behavioral Prediction and the Problem of Incapacitation," *Criminology* 32, no. 3 (1994):441–474.

[28] Validities of the prediction methods (rather than of the classifications) are summarized later in this report (see table 14).

[29] Cramèr's statistic is an estimate of correlation calculated from chi-square for any k by l table. This statistic is used because the classification is only rank ordered. Several other measures of the efficiency of prediction have been developed, and the question of which measure to use is complex. Other indices would be preferred if the samples compared differed in base rates or in the number of categories in the classification. For discussions of the limitations of the measures used here and of other measures of predictive efficiency, see S.D. Gottfredson and D.M. Gottfredson, "The Accuracy of Prediction Models," in *Research in Criminal Careers and "Career Criminals,"* vol. 2, edited by A. Blumstein et al. (Washington, DC: National Academy Press, 1986). For a briefer discussion, see S.D. Gottfredson, "Prediction: Methodological Issues," in D.M. Gottfredson and M. Tonry, (1987), *supra*, note 1, 29–33.

Burgess Method (Equal Weight Linear Model)

Various Burgess-type models were devised. All of the models followed the traditional method, with one point added for each predictor variable, but varying in the number of attributes included in the scale. The optimal number of items for this type of scale is not well established, although it may be assumed that as the number of positively correlated variates increases, the correlations between any two sets of weighted scores approach one and the effect of weighting the items tends to disappear. Experience shows that usually little predictive efficiency is gained after 8 or 10 items, and adding more items probably is redundant and may contribute more to face validity than empirical validity.[30] The issue of number of items is not necessarily a trivial one, because the acceptance and use of the instrument may be at stake. On the other hand, adding items has some operational cost and at some point fails to increase (or may actually decrease) predictive efficiency.

This study somewhat arbitrarily presents two scales: one based on only 9 items, the other based on 15. The validity of the five-group classification derived from application of each scale is examined.

Figure 1 shows the Burgess 9-item scale. In the construction sample, the Cramèr's Statistic was .39 (for the ungrouped scores). In the validation sample, it was .38 (for the grouped scores), which means that the grouping and testing in the different sample resulted in little shrinkage (loss of predictive efficiency).[31] Table 4 summarizes results from application of the Burgess

Figure 1: Burgess 9-Item Scale

Add one point for each item:

Any prior referral	___
Any prior adjudication	___
Any prior formal disposition	___
Any prior informal disposition	___
Any prior theft offense referrals	___
Three or more complaints in history	___
Any prior delinquency referral	___
Petition filed	___
Prior referral for drugs, delinquency, violence, or status	___

Total score: ☐

Classify youth:

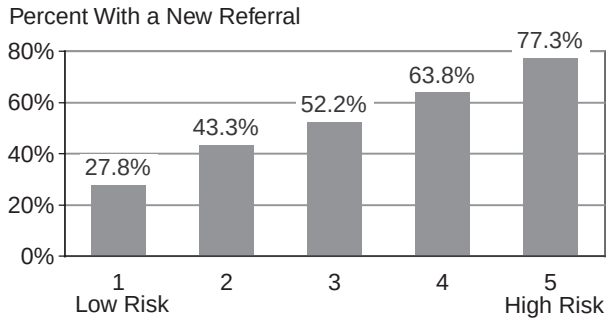
Score:	Group:	
0	1 Lowest risk	
1–3	2	Group ☐
4–5	3	
6–7	4	
8–9	5 Highest risk	

Table 4: Classification by the Burgess 9-Item Scale, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral Number	New Referral Percent
1	3,724	49.3%	1,034	27.8%
2	756	10.0	327	43.3
3	975	12.9	509	52.2
4	872	11.5	556	63.8
5	1,225	16.2	947	77.3
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .383

Figure 2: Classification Groups From the Burgess 9-Item Scale, Validation Sample



[30] The effect of adding more items was illustrated with the data in this study when Burgess scales made up of 6, 9, 12, 15, and 20 items were devised and tested. The coefficients measuring the relation of scores to the criterion (i.e., the Cramèr's statistics) were as follows: 6 items, .359 (in the construction sample) and .349 (in the validation sample); 9 items, .392 and .387; 12 items, .405 and .394; 15 items, .417 and .399; 20 items, .401 and .389.

[31] "Shrinkage" refers to the apparent loss in predictive power from the construction sample to the validation sample. Some shrinkage is normally expected and usually results from overfitting of the device or equation on the construction sample. For the comparisons in this report, shrinkage may be due to such loss, to rounding, or to the reduction of scales for the purpose of providing the five-group classifications.

9-item scale to the validation sample. Figure 2 shows the percentage of youth with at least one new referral to juvenile court during the 2-year followup period, for each of the five classification groups. The group classification, like the distribution of the Burgess scores, is markedly skewed to the right, and the classification places half of the youth in the lowest risk category.[32] Groups 2–5 have new referrals above the base rate of 43 percent.

Figure 3 shows the Burgess 15-item scale. Table 5 summarizes validation results from the 15-item scale, and figure 4 shows the percentage of youth in each classification group with at least one new referral during the 2-year followup.

Despite the significantly larger number of items in the 15-item scale, the validity coefficient (Cramèr's statistic) is only slightly larger than in the 9-item scale. Nevertheless, the classification based on the 15-item scale might be more useful administratively, because it provides better discrimination at the "low risk" end of the scale. In addition, the 15-item method offers more possibilities for the choice of cutting scores to establish the classification groups. Depending on the objectives of the classification decision, the greater choice of cutting scores may provide an operational advantage in that the five-group classification could be formulated in many ways.[33]

Multiple Linear Regression

Table 6 presents a summary of the multiple linear regression analysis, including the attributes and variables included in the resulting scale. Because the β coefficients show the relative contributions to prediction of each variable in the context of all other variables, it can be seen that the attribute "any prior referrals" is found to be the best predictor. Other variables contributing to the prediction are "number of prior referrals," "whether male," "whether Caucasian," and

Figure 3: Burgess 15-Item Scale

Add one point for each item:

Any prior referral	____
Any prior adjudication	____
Any prior formal disposition	____
Any prior informal disposition	____
Any prior theft offense referrals	____
Three or more complaints in history	____
Any prior delinquency referral	____
Petition filed	____
Prior referral for drugs, delinquency, violence, or status	____
Any prior detention	____
Male	____
Any prior adjustment	____
Any Prior status offense referral	____
At risk youth	____
Not Caucasian	____

Total score: [ ]

Classify youth:

Score:	Group:	
0	1	Lowest risk
1	2	
2–5	3	
6–9	4	
10 or more	5	Highest risk

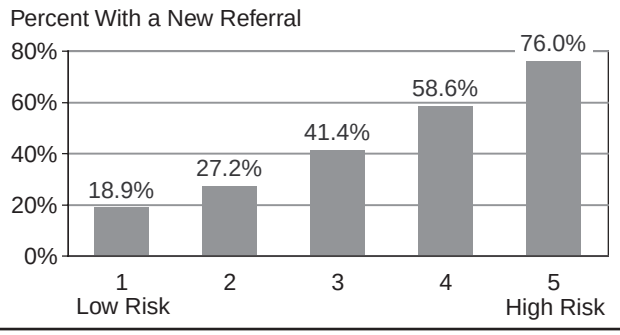
Group [ ]

Table 5: Classification by the Burgess 15-Item Scale, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral Number	New Referral Percent
1	830	11.0%	157	18.9%
2	2,103	27.8	573	27.2
3	1,758	23.3	728	41.4
4	1,489	19.7	872	58.6
5	1,372	18.2	1,043	76.0
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .390

Figure 4: Classification Groups From the Burgess 15 Item Scale, Validation Sample



[32] Because half of the sample had scores of zero (no prior referrals, no more than two complaints, and no petition filed), it was not possible to classify youth into approximately equal groups.

[33] This issue is discussed further, with examples from the results of the present study, in a later section (pages 27–28).

"age at complaint."[34] The other variables examined did not contribute to the apparent predictive value of the regression equation when the variables shown in table 6 were also included.

To provide a simple form for scoring purposes, the unstandardized regression coefficients were rounded to two decimal places and multiplied by 100. (The constant was ignored.) The resulting form, shown in figure 5, is similar to one that might be used operationally.

Table 7 summarizes results from application of the regression analysis-based classification to the validation sample. Figure 6 shows the percentage of youth with at least one new referral to juvenile court during the 2-year followup period, for each of the five classification groups. The validity and operational utility of the regression classification seem similar to those of the classification based on the 15-item Burgess scale. The overall relations of the classifications to the criterion (as measured by Cramèr's statistic) are practically identical; and, as is the case with the 15-item Burgess method, the distribution of scores in the regression classification offers many possibilities for selections of cutting scores to fit the objectives intended for the use of this tool.

Table 6: Multiple Linear Regression of New Referrals on Selected Items

Item (Attribute or Variable)	Unstandardized Coefficient (*B*)	Standard Error of *B*	Standardized Coefficient (β)	Significance
Male	.161	.023	.150	<.001
Caucasian	–.107	.021	–.106	<.001
Age at complaint	.018	.007	.058	<.001
Number prior referrals	.029	.006	.136	<.001
Any prior referrals	.224	.027	.223	<.001
Constant	.008	.093		.934

R = .420
R^2 = .176

[34] The issue of removing the effect of unwanted variables is discussed in a later section (page 30).

Figure 5: Classification of Youth Based on Multiple Linear Regression Analysis

If

Not Caucasian	**Add**	11	_____
Any prior referrals	**Add**	22	_____
Male	**Add**	16	_____

And add

2 times the age at the time of the complaint	_____
3 times the number of prior referrals	_____

Total score: ☐

Classify youth:

Score:	Group:	
20–30	1	Lowest risk
31–40	2	
41–58	3	
59–73	4	
74 or higher	5	Highest risk

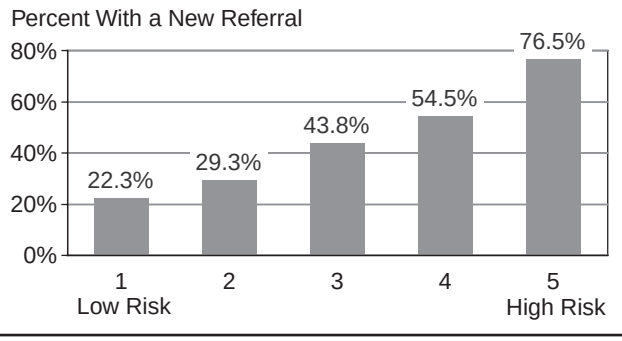
Group ☐

Table 7: Classification by the Multiple Linear Regression Scale, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral Number	New Referral Percent
1	1,537	20.4	343	22.3
2	1,818	24.1	533	29.3
3	1,242	16.4	544	43.8
4	1,398	18.5	762	54.5
5	1,557	20.6	1191	76.5
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .395

Figure 6: Classification Groups From the Multiple Linear Regression Scale, Validation Sample



Logistic Regression

Figure 7 shows the calculation for the risk scale and the method for classification into five groups, using logistic regression with the weights rounded and multiplied by 10. This scale is based on results from the construction sample, shown in table 8. The results are similar to those for multiple linear regression, discussed previously. The same items were selected, except that "number of prior delinquency referrals" was included rather than "number of prior referrals."

Table 9 summarizes results from application of the logistic regression analysis-based classificiation in the validation sample. Figure 8 shows the percentage of youth with at least one new referral to juvenile court during the 2-year followup period, for each of the five classification groups. The validity coefficients for the logistic regression method are a little lower than those found with the other methods discussed thus far. Operational utility appears likely to be on a par with the Burgess 15-item scale and with the scale derived from the multiple linear regression analysis.

Figure 7: Classification of Youth Based on Logistic Regression Analysis

If

Any prior referrals	**Add**	8.9	_____
Male	**Add**	5.9	_____

And add

2.6 times the number of prior delinquency referrals			_____
	Subscore		_____

Then subtract

If Caucasian		5.5	_____
0.7 times the age at the time of the complaint			_____
	Subscore		_____
	Total Score		

Classify youth:

Score:	Group:	
9.90 or lower	1	Lowest risk
9.91–12.30	2	
12.31–17.50	3	
17.51–19.80	4	
19.81 or higher	5	Highest risk

Group

Table 8: Logistic Regression of New Referrals on Selected Items

Item (Attribute or Variable)	Unstandardized Coefficient (*B*)	Standard Error of *B*	Exponent (β)	Wald*	*P*
Any prior referrals	.892	.136	2.440	42.907	<.001
Male	.586	.109	1.796	28.646	<.001
Caucasian	–.551	.103	.562	28.554	<.001
Age at complaint	–.072	.009	.931	66.831	<.001
Number of prior delinquency referrals	.263	.048	1.301	30.093	<.001

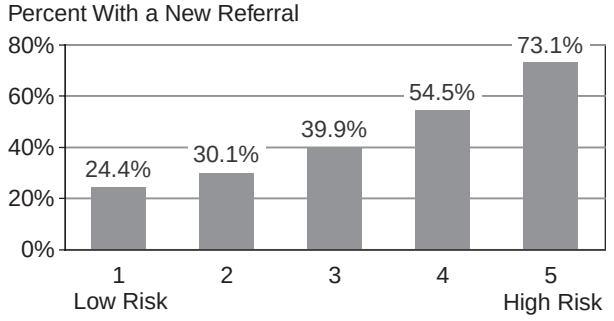
* The Wald statistic, which has a chi-square distribution, was used to test whether the observed coefficient differs from zero. In each case, the degrees of freedom = 1.

Table 9: Classification by the Logistic Regression Scale, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral	
			Number	Percent
1	1,596	21.1	389	24.4
2	1,343	17.3	393	30.1
3	1,609	21.4	645	39.9
4	1,435	19.3	794	54.5
5	1,569	20.9	1,152	73.1
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .357

Figure 8: Classification Groups From the Logistic Regression Scale, Validation Sample



Predictive Attribute Analysis

The predictive attribute analysis was carried out on the construction sample with the 1-percent level of significance (for chi-square with one degree of freedom) taken as the stopping rule. Candidates for the best predictor for each subgroup identified were all attributes shown in table 1 (not already used in the subdivision process).

This process resulted in the identification of eight groups, as shown in figure 9. The classification is simple, the procedure has considerable face validity, and the eight-group classification probably would work well in operations.

The number of groups, however, was reduced to five, to permit comparisons with the other five-group classification procedures tested. The five groups were obtained by combining several groups with similar rates of new referrals, as shown in figure 10.[35] When the classification in the validation sample was carried out, the results, shown in table 10 and figure 11, were obtained.

[35] The following groups from figure 9 were combined: 1 and 2, 4 and 5, and 7 and 8.

Figure 9: Predictive Attribute Analysis, Construction Sample

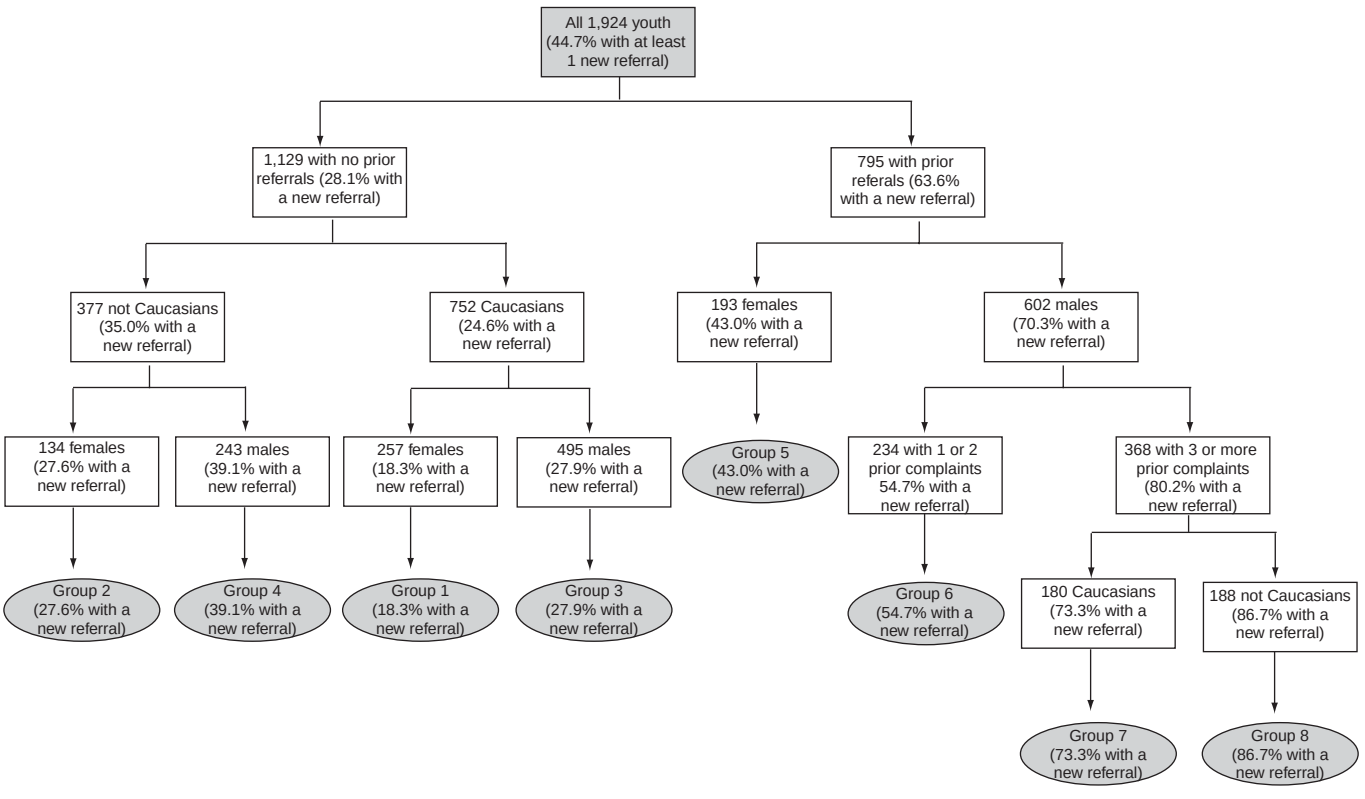
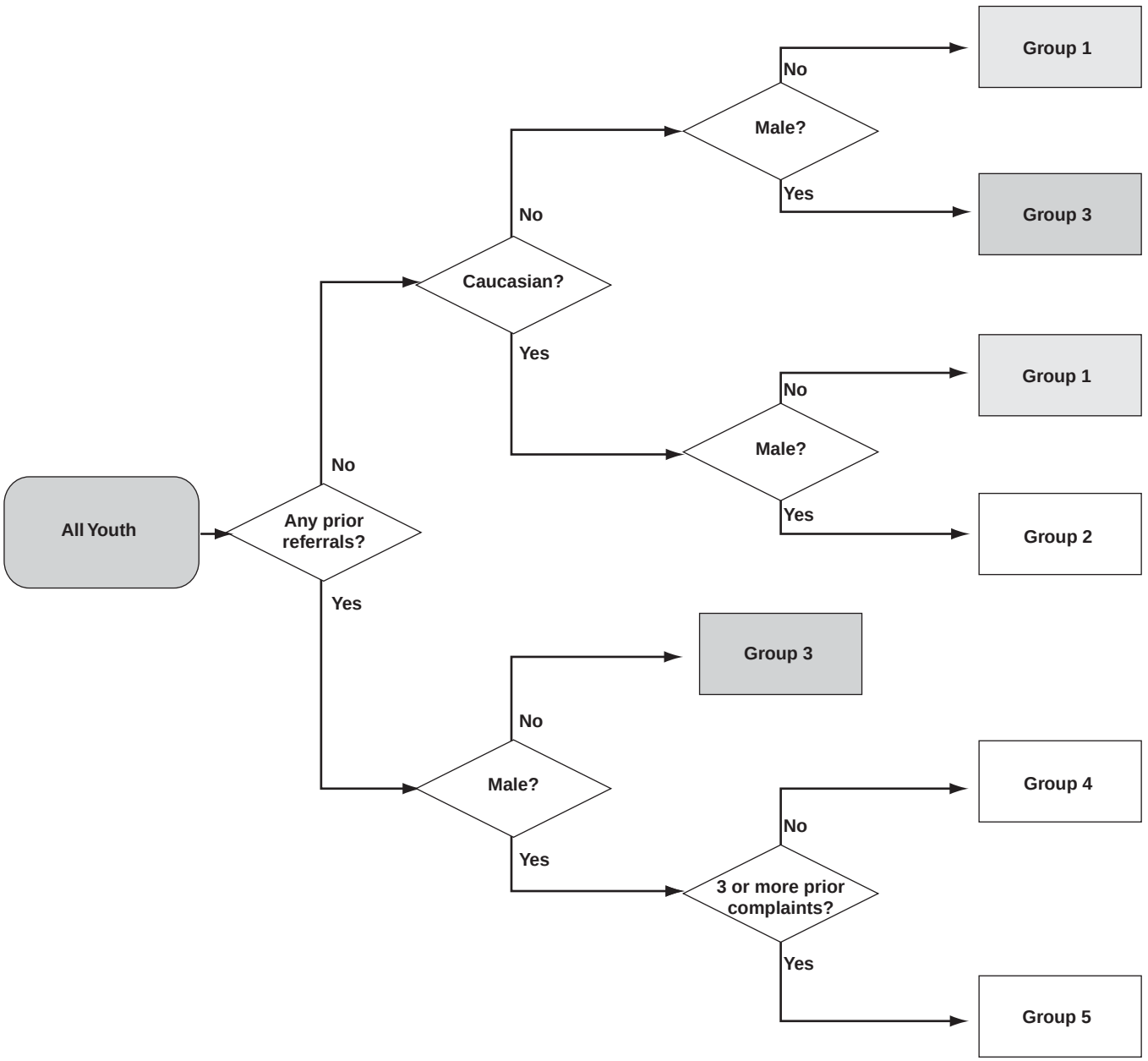


Figure 10: Classification Into Five Groups by Predictive Attribute Analysis



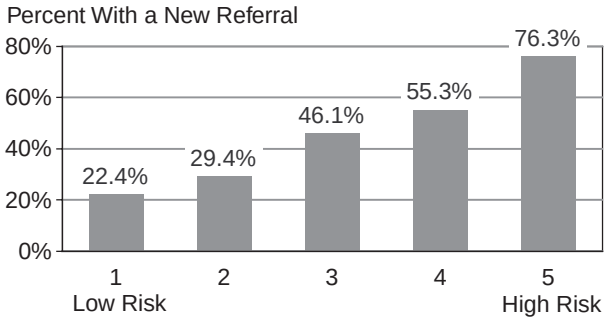
Other grouping procedures might be more useful operationally, depending upon the objectives for the intended use of the classification. For example, going back up the tree of subdivision in figure 9, five groups may be defined. Ignoring the Caucasian attribute among youth with prior referrals results in combining groups 7 and 8. Ignoring the gender attribute among youth with no prior referrals results in combining groups 2 and 4 and 1 and 3.

Table 10: Classification by Predictive Attribute Analysis, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral	
			Number	Percent
1	1,484	19.7%	332	22.4%
2	1,821	24.1	536	29.4
3	1,765	23.4	813	46.1
4	965	12.8	534	55.3
5	1,517	20.1	1,158	76.3
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .387

Figure 11: Classification Groups From the Predictive Attribute Analysis, Validation Sample



Bootstrap Analysis

A potential advantage of a bootstrap procedure is that it takes into account the possibility that the weights developed on the basis of a sample as a whole may be wrong for subgroups if intercorrelations differ for them. Disadvantages may be that sample sizes are reduced, resulting in less stability of the weights and an increased likelihood of capitalizing on chance variation. One result may be that greater shrinkage is found on validation.

To explore a bootstrap method with the Maricopa County data, the authors conducted separate regression analyses on the data for youth with and without any prior referral to the juvenile court. Tools simulating the operational use of the results in practice were devised (see figure 12) and the relation of the five-group classifications with the criterion was measured in the validation sample. The results are summarized in tables 11 and 12 and figures 13 and 14.

Next, a classification based on both equations was devised to compare the results with those found for the other methods. The two groups (with and without prior referrals) were combined into one five-group classification on the basis of the percentages with at least one new referral; again, equal groups were sought to the extent possible.[36] Table 13 and figure 15 show results from application of this classification in the validation sample. The lower validity coefficients for the two subgroup classifications are not surprising, because variability is reduced in both the predictors and the criterion. The validity coefficient was higher when all youth are classified by the combining procedure, but it was similar to that found for the other methods.

[36] For cases with no prior referrals, group 1 (22 percent with a new referral) was scored 1, groups 2, 3, and 4 (26, 33, and 33 percent respectively) were scored 2, and group 5 (48 percent) was scored 3. For cases with prior referrals, groups 1 and 2 (39 and 52 percent respectively) were scored 3, groups 3 and 4 (53 and 58 percent) were scored 4, and group 5 (78 percent) was scored 5.

Figure 12: Classification of Youth With and Without Prior Referrals

No Prior Referrals:
Calculate score if no prior referrals and

If

Male	**Add**	11	_____
Not Caucasian	**Add**	12	_____
2 times age at time of complaint			_____
	Subtotal		
Theft complaint	**Subtract**	8	_____
	Total Score		

Classify youth with no prior referrals:

Score:	Group:	
0–28	1	Lowest risk
29–33	2	
34–39	3	
40–43	4	
44 or higher	5	Highest risk

Group

Prior Referrals:
Calculate score if prior referrals and

If

3 or more prior complaints	**Add**	13	_____
Male	**Add**	25	_____
Not Caucasian	**Add**	10	_____
Petition filed	**Add**	8	_____
3 times the number of prior theft referrals			_____
All cases	**Add**	20	_____
	Subtotal		
Not referred by law enforcement agency	**Subtract**	20	_____
	Total Score		

Classify youth with prior referrals:

Score:	Group:	
0–10	1	Lowest risk
11–25	2	
26–35	3	
36–43	4	
44 or higher	5	Highest risk

Group

Table 11: Classification of Youth With No Prior Referrals, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral Number	New Referral Percent
1	1,270	29.5%	278	21.9%
2	1,104	25.6	293	26.5
3	749	17.4	247	33.0
4	657	15.3	218	33.2
5	527	12.2	252	47.8
Total	4,307	100.0	1,288	29.9

Cramèr's statistic = .175

Figure 13: Classification of Youth With No Prior Referrals Into Five Groups

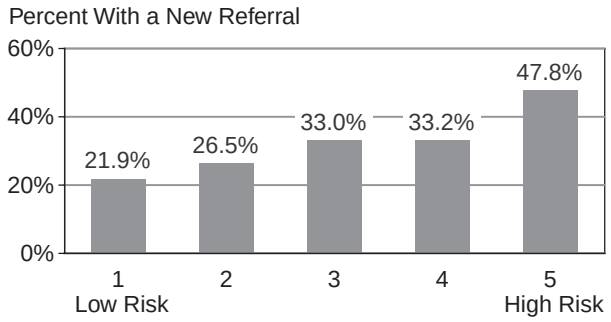


Table 12: Classification of Youth With Prior Referrals, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral	
			Number	Percent
1	204	6.3%	80	39.2%
2	502	15.5	260	51.8
3	543	16.7	288	53.0
4	503	15.5	293	58.3
5	1,493	46.0	1,164	78.0
Total	3,245	100.0	2,085	64.3

Cramèr's statistic = .277

Table 13: Classification of All Youth by Bootstrap Method, Validation Sample

Group Classification	Number in Group	Percent of Total	New Referral	
			Number	Percent
1	1,270	16.8	278	21.9
2	2,510	33.2	758	30.2
3	1,233	16.3	592	48.0
4	1,046	13.9	581	55.5
5	1,493	19.8	1,164	78.0
Total	7,552	100.0	3,373	44.7

Cramèr's statistic = .399

Figure 14: Classification of Youth With Prior Referrals Into Five Groups

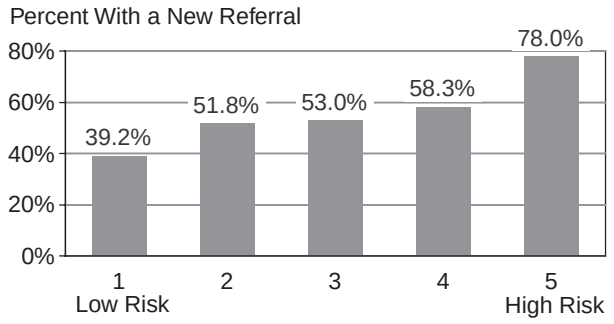
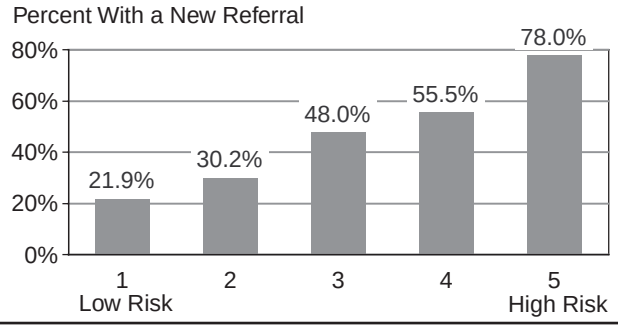


Figure 15: Classification of All Youth Into Five Groups by Bootstrap Method



Comparison of Methods

Predictive Efficiency and Operational Utility

The methods presented may be compared with respect to concepts of predictive efficiency and operational utility. Concepts concerning prediction—empirical validity and shrinkage—may be measured objectively. Concepts concerning operational utility—simplicity, face validity, and flexibility—require a more subjective judgment.

Empirical Validity

The concept of empirical validity refers to how well a classification method works when applied to samples other than the one used for its construction. In examining validity, this study emphasizes the ability of five-group classifications to predict recidivism, rather than using the methods' full range of scores or classifications. (The study also reports the validity of the full range of scores or classifications, but the findings are relevant to the prediction method itself, rather than to the classifications based on the methods.) Some loss in predictive validity is expected as a result of using groupings of risk scores (i.e., risk categories) instead of the full range of risk scores.

The validity of the prediction equations or procedures before grouping youth on the basis of scores or categories to provide the five-group classifications may be seen in table 14, which shows the correlations of scores or original classifications in the construction and validation samples. The validity coefficients (the correlations of scores with the outcomes in the validation sample) all are comparable to validities reported in other studies from methods based on similar data, and those based on the full sample do not vary markedly among the different methods. Some shrinkage was found for each prediction method.

Table 14: Correlation of Prediction Scores With Outcomes, Construction and Validation Samples

Prediction Method	Point Biserial Correlation, Construction Sample (*N* = 1,924)*	Point Biserial Correlation, Validation Sample (*N* = 7,552)*	Shrinkage (Percent)
Burgess 9-Item	.388	.378	1.0
Burgess 15-Item	.404	.392	1.2
Multiple Linear Regression	.419	.397	2.2
Logistic Regression	.387	.368	1.9
Predictive Attribute Analysis **	.415	.387	2.8
Regression Based on Sample With No Prior Referrals	.287	.250	3.7
Regression Based on Sample With Prior Referrals	.398	.356	4.2
Bootstrap (Combined Regressions, Prior Referrals/No Prior Referrals) ***	.475	.421	5.4

* Note that sample sizes were reduced for regressions based on samples of youth with and without prior referrals.

** Groups scored 1–8 according to percents with at least one new referral in the construction sample and treated as a continuous score.

*** Correlations calculated from the analysis of variance in outcomes.

Shrinkage

A smaller amount of shrinkage might give greater confidence that the validity of the prediction method, and the classification procedure derived from it, will hold up on repeated applications to validation samples. The shrinkage shown in table 14 can be thought of as a result of "overfitting" the model in the construction sample. (The shrinkage shown in table 15 under "Summary of Results" is a combination of this source of loss of validity with that attributable to grouping to provide the simpler five-category classifications.)

Simplicity

It may be assumed that a simple classification method that requires little staff time to complete will be better accepted by staff and will therefore be used with greater reliability and completeness. Simplicity results from including fewer items, ensuring that the meaning of the items is clear, and requiring minimal arithmetic to complete the form.

Face Validity

If the items included, and any weights assigned, appear reasonable and seem to square with the experience of the user, it may be assumed that the method will be better accepted and more effectively used. With weighted items, the weights assigned may not appear intuitively correct (i.e., it may be difficult to see why a particular weight was assigned), and staff without a statistical background may have difficulty following explanations of intercorrelations, least squares, or maximum likelihood.

Flexibility

Some methods, such as those that provide continuous scores with numerous opportunities for assigning cutting scores in devising the classification, offer greater flexibility for operational use. Figures 16 through 19 may clarify this concept by showing examples of

Figure 16: Frequency Distribution of Scores for the Burgess 15-Item Scale Scores for Youth With and Without at Least One New Referral (Construction Sample)

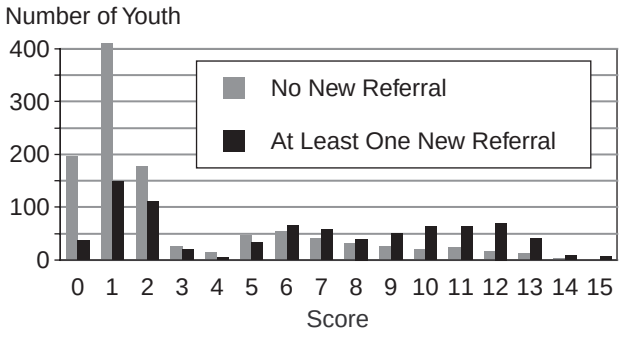


Figure 17: Percent Distribution of Scores for the Burgess 15-Item Scale Scores (Ungrouped) for Youth With at Least One New Referral (Construction Sample)

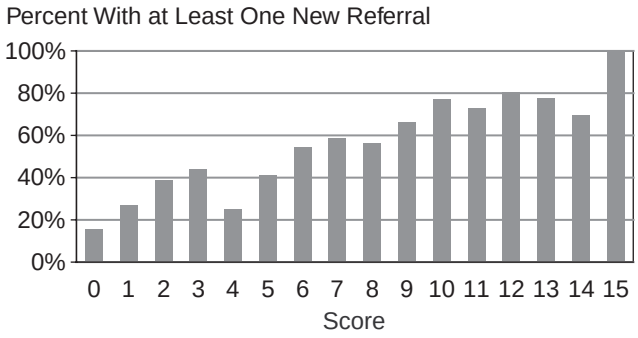


Figure 18: Frequency Distribution of Scores for the Multiple Linear Regression Scale Scores (in 15 Approximately Equal Groups) for Youth With and Without at Least One New Referral (Construction Sample)

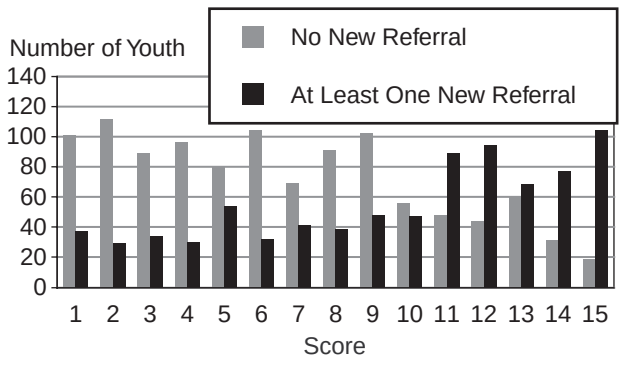
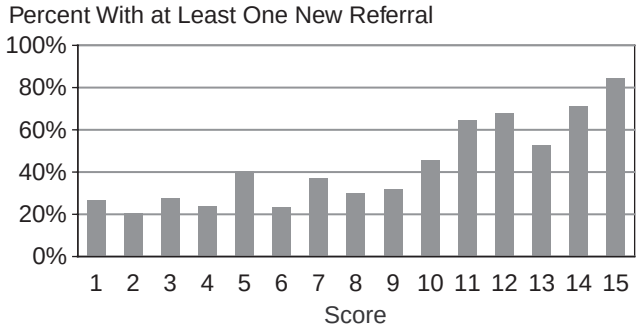


Figure 19: Percent Distribution of Scores for the Multiple Linear Regression Scale Scores for Youth With at Least One New Referral (Construction Sample)



the frequency and percent distributions of scores. As can be seen, there are many opportunities for cutting scores to provide group classifications different from those used in this study. The groupings to provide classifications for operational use should be based on objectives for the intended goals of the classification method as well as the distributions of scores.

Summary of Results

Tables 15 and 16 summarize the validation results for the five-group classifications, listed according to the prediction methods on which they were based. All of the classification tools (five-group classifications) show a useful degree of validity when applied to the validation sample of 7,552 youth. Table 16 ranks the methods by validity and shrinkage and also offers the authors' ratings for the more subjective criteria. The observed validities provided the basis for the ranking, but it should be borne in mind that all of the differences are quite small. As indicated in table 17, the classification procedures described are substantially related.

Table 15: Criterion Variance Explained, Construction and Validation Samples, Five Classification Methods

Prediction Method Providing the Basis for Classification	Construction Sample (N = 1,924) Correlation*	Construction Sample (N = 1,924) Variance Explained**	Validation Sample (N = 7,552) Correlation*	Validation Sample (N = 7,552) Variance Explained**	Shrinkage (Percent of Explained Variance)
Burgess 9-Item	.390	.152	.383	.147	3.3%
Burgess 15-Item	.412	.170	.390	.152	10.6
Multiple Linear Regression	.419	.175	.395	.156	10.8
Logistic Regression	.374	.140	.357	.128	8.6
Predictive Attribute Analysis	.421	.178	.387	.150	2.8
Bootstrap (all cases, combined)	.432	.187	.399	.159	15.0

Prior Referrals or Not (Multiple Linear Regression)	Number	Construction Sample Number	Construction Sample R**	Construction Sample R^2**	Validation Sample Number	Validation Sample R**	Validation Sample R^2**	Shrinkage (Percent of Explained Variance)
No prior referrals	5,436	1,129	.208	.043	4,307	.175	.031	27.9%
Prior referrals	4,040	795	.352	.124	3,245	.277	.077	37.9
Bootstrap	9,476	1,924	.440	.194	7,552	.406	.165	14.9

* Cramèr's statistic.
** Calculated from the analysis of variance. All are significant at the .001 level of confidence by F tests.

Table 16: Rankings and Authors' Ratings of Six Classification Procedures

Method Used as a Basis for Classification	Validity Observed (Rank Order)	Simplicity	Face Validity	Flexibility	Prediction Score Shrinkage (Rank Order)	Classification Score Shrinkage (Rank Order)
Burgess 9-Item	5	High	High	Medium	1	2
Burgess 15-Item	3	High	High	High	2	4
Multiple Linear Regression	2	Medium	Medium	High	4	5
Logistic Regression	6	Low	Medium	High	3	3
Predictive Attribute Analysis	4	High	High	Low	5	1
Bootstrap	1	Low	Medium	High	6	6

Table 17: Correlations of Classification Procedures Based on Various Prediction Methods, Validation Sample (*N* = 7,552)

Prediction Methods on which Classifications Were Based	Burgess 15-Item	Multiple Linear Regression	Logistic Regression	Predictive Attribute Analysis	Bootstrap
Burgess 9-Item	.901	.845	.763	.862	.878
Burgess 15-Item		.920	.865	.924	.755
Multiple Linear Regression			.880	.947	.931
Logistic Regression				.909	.843
Predictive Attribute Analysis					.932

* All correlations are significant at the .001 level of confidence

Removing Invidious Predictors

Depending on the objectives of a classification's intended use, some variables (such as race or gender) commonly included in most efficient predictors may be unwanted for a variety of ethical and practical reasons. An advantage of using multiple linear regression as the basis for classification is that a procedure exists for removing the effects of these "invidious" variables when this method is used.[37]

It is not possible to remove the effects of invidious predictors if other methods are used. Simply leaving out an unwanted variable does not fully accomplish the purpose of removing its effect. Because these variables usually "overlap" with others used in the procedure, the effect of the unwanted item remains, even though it may be somewhat hidden. For example, in the Burgess 15-item scale, the race and gender items could be left out, leaving a 13-item scale with little loss in validity. Effects of race and gender, however, would remain, primarily because of the correlations of these items with others among the remaining 13.

When unwanted variables are included as predictors in a multiple linear regression, their effect can be removed statistically before the risk classification instrument is finalized. For example, by incorporating a race variable early in the development of a regression equation, the relationship between race and recidivism will be captured by the race factor in the equation. Then as other factors are added to the equation, their beta weights will reflect the relationships between each variable and recidivism independent of race. Once the analysis is completed, the race factor in the equation is replaced with a constant equal to the race factor's beta weight multiplied by the code for "White." Using the resulting regression equation as the basis for the risk instrument, users can be confident that their scale has been "purged" of the influence of race on the classification. Although validity may be somewhat reduced, the classification procedure should be more acceptable than it would have been had the variable remained in the equation, and still be useful in practice. Thus, the multiple linear regression method is particularly well suited to classification when invidious predictors present a problem for application.

Which Method?

Validity measures were about the same for all of the methods tested in this study. The tools based on the various methods were substantially correlated. All exhibited validities are similar to those reported in the literature for classification instruments in use by juvenile justice agencies. A court might find any of the methods tested useful in developing a risk classification procedure. A Burgess-type scale with a dozen or so items could be preferred, based on its validity in this test sample, its face validity and simplicity, and evidence from earlier studies. In addition, the predictive attribute analysis method could be recommended for its simplicity, face validity, and apparent empirical validity in this study and prior research.

If the effects of invidious variables (e.g., race) need to be removed, however, the authors' strong preference would be for the multiple linear regression method, despite the violation of statistical assumptions underlying its use. The removal of the direct and indirect effect of race and ethnicity from the tools used by juvenile justice decisionmakers serves a key component of OJJDP's Disproportionate Minority Contact mandate. Risk scale developers must confront the possibility directly that their products may add unwanted (and unknown) biases to the decisionmaking processes within the juvenile justice system. Multiple linear regression methods give developers the capacity to remove the independent effect of race from their risk scales. Race (or ethnicity) must play a part

[37] S.D. Gottfredson and D.M. Gottfredson, "Risk Measures for Operational Use: Removing Invidious Predictors," in *Juvenile Justice with Eyes Open*, edited by D.M. Gottfredson (Pittsburgh, PA: National Center for Juvenile Justice, 2000). See F. Fisher and J.B. Kadane, "Empirically Based Sentencing Guidelines and Ethical Considerations," in *Research on Sentencing: The Search for Reform*, edited by A. Blumstein, J. Cohen, S.E. Martin, and M. Tonry (Washington, DC: National Academy Press, 1983). Although the procedures described in these reports remove a substantial proportion of the effects of invidious variables, the procedures are incomplete. A procedure for devising completely unbiased models with little loss in predictive utility is described and demonstrated in S.D. Gottfredson and G.R. Jarjoura, "Race, Gender, and Guidelines Decision Making," *Journal of Research in Crime and Delinquency* 33, no. 1 (1996):49–69.

statistically in the development of risk scales, but the developers/technicians should remove the direct and indirect effects of race from the risk scale before it is provided to the line staff to use and (subsequently) the scores are provided to decisionmakers to interpret and apply. This means that while race should be a variable in the regression equations used to develop and test risk scales, a youth's race should not be a predictive factor on any risk scale. It is the responsibility of the technicians to do what they can to minimize bias in the juvenile justice system—and careful application of proven statistical methods can go a long way to remove unintended biases from risk scales.

One final point must be made for technicians and decisionmakers alike. The more theoretically sophisticated prediction methods may work better when the scope and reliability of data are improved. One avenue for improving prediction may be the collection of additional data with hypothesized relations to the criterion, particularly for youth referred to the courts for the first time. Therefore, given the current "state of the art," the need for data improvement may be even more important than the need for increased statistical sophistication.