Improved Techniques for Assessing the Accuracy of Recidivism Prediction Scales[a]

by

Jacqueline Cohen[b]
Sherwood E. Zimmerman[c]

February 15, 1990

[a]This research was supported by grant #86-IJ-CX-0039 from the National Institute of Justice, U. S. Department of Justice. The views expressed here are those of the authors and do not necessarily reflect the opinions of the funding agency.
[b]School of Urban and Public Affairs, Carnegie Mellon University
[c]Department of Criminology, Indiana University of Pennsylvania

INTRODUCTION

Prediction has a lengthy and respected place in the history of Criminology. The long-standing interest in statistical or actuarial prediction arises from two sources. First, from the positivist tradition, successful prediction is the ultimate scientific test of criminological theory. Secondly, statistical prediction devices have been adopted to provide decision support information at important stages of the criminal justice process through explicit predictions about offenders' "future expected behavior" (Gottfredson and Gottfredson, 1980).

The development of quantitative instruments to predict criminal behavior began in the 1920's with attempts to predict recidivism by individuals being considered for probation and parole (Gottfredson and Gottfredson, 1980 p. 214). More recently, concern for public protection, combined with mounting pressures on criminal justice system (CJS) resources, have operated to intensify efforts to develop and implement prediction scales that will more effectively allocate scarce CJS resources among offenders (e.g., Greenwood and Abrahamses, 1982).

Two steps are typically involved in producing useful actuarial prediction scales for case processing decisions. First, a "construction sample" is obtained that is representative of the population of interest. Using the characteristics of the defendants or convicted offenders in this sample, an empirical scale is developed whose cut-points classify sample members into subgroups. The subgroups are identified by the differing levels

of risk individuals pose with respect to the relevant outcome (e.g., failure to appear, recidivism, career criminal). While the technology for developing prediction scales has improved considerably over the years, and while further incremental improvements are likely in the future, major advances in classification technology continue to be constrained by ineffective measurements of the variables used in these scales (Wilkins, 1969).

The second step in producing a usable prediction instrument involves assessing the results when the classification scale is prospectively applied to a separate "validation sample" obtained from the target population. Using the magnitude by which prediction accuracy deteriorates from the construction to the validation sample as an index of scale accuracy was an important advance in validating prediction instruments (Wilkins, 1969). The magnitude of this deterioration ("shrinkage"), when combined with information about the direction and magnitude of the resulting prediction errors, provides information about the expected performance of a specific scale in a single population.

The performance of a prediction instrument relative to the performance of some random prediction process is another approach for assessing the utility of prediction instruments. The frequencies expected under the random process are defined in a manner similar to that used in calculating the Chi Square statistic (Meehl and Rosen, 1955). In the 2 x 2 matrix formed by cross-classifying binary predicted outcomes with the actual case outcomes in a sample, the proportion of the sample predicted to have the

criterion outcome constitutes the scale's selection rate, while the proportion of actual criterion outcomes reflects the base rate within the sample. The expected frequencies are computed from products of the selection rate and base rate. Predictive efficiency is assessed by comparing the number of correct predictions expected by the chance process with the number of correct predictions made by a prediction scale. (Wiggins, 1973).

A limitation of all these measures is that they do not permit comparisons of the utility of a prediction scale relative to other scales, or with other populations of offenders. It seems reasonable that an effective evaluation of the Type I and Type II error[1] associated with a scale would involve comparisons with the error generated by similar scales. The ability to make those comparisons, however, has been limited by the lack of procedures for simultaneously controlling differences in the selection rates of prediction instruments and in the base rates resulting from the

---

[1] In the prediction context, Type I error occurs when the criterion attribute is predicted not to be present for cases that actually possess the criterion attribute. These errors are also referred to as "False Negative errors." When such errors are made in the criminal justice system, the consequences of being wrong are usually highly visible and can subject the relevant decision process to serious criticism, as for example, when an offender predicted to be an acceptable parole risk commits a serious crime while on parole release from prison. Type II error occurs when the criterion attribute is predicted to be present in cases that do not possess the attribute. In the criminal justice system these "False Positive errors" are especially insidious because their low visibility combines with unwarranted severe consequences for offenders. For example, the prediction that some inmates pose too high a risk for release on parole can lead to unnecessary extended incarceration of inmates who would not commit further crimes if they were released, and in this situation such errors will go undetected.

attributes or composition of the different study samples. The "Random Improvement Over Chance" (RIOC) measure, which is neither scale nor a data dependent, has been suggested as a tool capable of providing information for such assessments (Loeber and Dishion, 1983). Despite its apparent value, this measure has been infrequently reported in criminological research literature.

The properties of the RIOC measure are examined in this study through use of four prediction scales, each of which was designed to predict different outcomes, and four datasets that differ from the populations that were originally used to construct the scales. This study was specifically designed to assess the robustness of the four scales in a variety of applications. The specific purposes of the research reported here are:

1. to assess the accuracy of some existing scales that are of general interest, and
2. to examine the RIOC measure as an indicator of prediction accuracy that is not dependent on the base rate and selection rate of individual samples.

THE SCALES

The four prediction scales used in this research were created for different purposes. The INSLAW scale (Rhodes et. al., 1982) and the RAND scale (Greenwood, 1982) were somewhat similar in their purposes. The INSLAW scale was constructed on groups of arrestees in Washington, D.C. with the purpose of more effectively allocating prosecutorial resources by identifying "career criminals." The RAND scale, designed to extend the incarceration terms of inmates

who were predicted to commit crimes at high rates, was developed using a sample of inmates from 3 states (California, Michigan and Texas). Both scales were designed to prospectively identify offenders who posed substantial threats to society. Relying on "time to rearrest" as the dependent variable, the INSLAW scale sought to identify those individuals who had a substantial probability of committing a subsequent crime quickly. Using self-reported crime commission rates, the RAND scale was designed to identify those offenders who commit crimes frequently.

The SFS81 scale was developed by the Federal Parole Commission as an index of the "salient factors" that are used to assess the risk of recidivism posed by inmates who are eligible for release on parole from Federal prisons (Hoffman, 1983; U.S. Parole Commission, 1985). The third revision of the Salient Factor scale was constructed from the post-release recidivism experience of a sample of Federal offenders, and is currently being used by the Parole Commission in making parole decisions. The final scale, the CGR scale, was developed by the Center for Governmental Research as a model scale for making pretrial release decisions in New York state jurisdictions other than New York City (Center for Governmental Research, 1982/83). This scale was constructed using a sample of defendants who were awaiting trial in selected New York State jurisdictions, some of whom were on pretrial release and others who were held in pretrial detention.

As seen in Table 1, the scales all include a variable reflecting prior Adult Criminal Record, as well as a Juvenile

Criminal Record variable. Indicators of Drug Use and/or Alcohol Use are used in all but the CGR scale. The INSLAW and SFS81 scales include variables indicating the Current Age of the offender, while the CGR scale includes Education. Two of the scales, RAND and CGR, also include variables concerning the offenders' recent Employment History. With the exception of Employment History, some measures of all these variables are available in each of the datasets.

Table 1, About Here

THE DATASETS

Four datasets were selected to reflect, as much as possible, different geographical areas, as well as a mix of case processing stages in the criminal justice system (arrest, conviction, incarceration). Although there were differences in the nature and quality of information available, a series of data recodes were undertaken to operationalize the scale items as consistently as possible across the datasets.

The general research strategy was to partition an offender's longitudinal history of criminal justice involvement into "prior record" and "follow-up" data, as in Figure 1. All the datasets contain longitudinal information on individual offending as indicated by criminal justice interventions (arrest or filed charges), as well as other individual attributes. A criminal justice intervention as an adult (e.g., first adult arrest) was designated as the "target event" which would trigger the

application of the prediction scales for all sample members. All data on attributes prior to the target event were used to measure the background characteristics that entered an individual's scale score, and offending after the target event was used to define the follow-up outcome variables.

FIGURE 1, About Here

The general characteristics of the four datasets are described in Table 2[2]. The Department of Labor (DOL) data were collected originally by the VERA Institute of Justice in New York City as part of an experimental evaluation of a jobs training program implemented in Albuquerque, Miami, and New York City (Sadd, et. al., 1983). From the larger sample of persons identified as "high risk youth" between the ages of 16 and 21, we selected for analysis the subset of 746 program participants who had an arrest sometime prior to their referral into the program. This group constitutes approximately one-third of all the cases in each of the three program sites. The arrest immediately preceding program participation was used as the target event for application of the prediction instruments. The mean age at this target arrest was 17.3 years and sample members were followed for an average of 1.8

---

[2] All the data used in this study were originally collected for other purposes, and we are grateful to the researchers involved for making them available to us.

years after their target arrest. During this follow-up period, 19% of those in the analysis sample were arrested for an index property offense[3], 12% for robbery, and 7% for an index violent offense other than robbery.[4]

Table 2, About Here

The remaining three samples all came from California. The prison and probation (P&P) data were collected by the RAND Corporation, and contain matched samples of convicted felons who were sentenced either to prison or to felony probation (Petersilia and Turner, 1986). The offenders in these samples were convicted in Alameda and Los Angeles counties and comprised about one-third of California's total felony convictions in 1980. The arrest associated with this 1980 conviction was used as the target event for applying the prediction instruments. The prison and probation samples in the P&P dataset did not differ significantly in terms of subsequent recidivism when crime type of subsequent arrests and prediction scale scores were statistically controlled.

---

[3]Using the FBI Uniform Crime Report definitions for index offenses, index property offenses are: burglary, larceny-theft, and motor vehicle theft.

[4]Using the FBI Uniform Crime Report definitions for index offenses, index violent offenses are: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. Because it combines elements of both violent and property offenses, robbery is treated separately in this analysis.

The combined P&P data contains 1,022 individuals and includes the oldest offenders among the four analysis datasets, averaging almost 27 years of age at the target event. Sample members were followed for an average of 2.6 years, including at least 24 months following release to the community from any incarceration resulting from the target event. Despite the fact that the P&P sample was comprised of convicted felons and the DOL sample was based on arrestees, the two groups were quite similar in their recidivism rates; 25% of the P&P offenders were rearrested for an index property offense, 8% for robbery, and 5% for a violent index offense (excluding robbery). The similarities are remarkable given the differences between the two groups of offenders and the relatively short follow-up periods for which data were available.

The final two samples were based on three studies of juvenile offenders who were incarcerated in three California Youth Authority (CYA) institutions during the 1960's and 1970's. The data were brought together as part of a long-term study of criminal careers by the CYA (Haapanen and Jesness, 1982; Haapanen, 1988). The 99.5% of the male juveniles in the original CYA study who were subsequently arrested as adults (i.e., sometime after their 18th birthday) were used in the analysis reported here. The first adult arrest was used as the target event, and the samples were followed for an additional 8 to 11 years after the target event. The follow-up for these CYA data was much longer than is typically available, and the greater time-at-risk at least partly accounts

for the much higher recidivism rates observed in Table 2.[5]

A generalized least squares (GLS) procedure was used to determine whether the various subsamples within each of four major datasets could be combined, or whether they should be analyzed as separate datasets. At issue was whether the relationship between scale scores and follow-up arrests was essentially the same for the various subsamples. In the CYA data, for example, subgroups of offenders were identified based on the sampled CYA institution, their scale score, and crime type of follow-up arrests. Mean outcomes were obtained within each subgroup for each of four alternative outcome variables (RECID, NUMARR, FREQ, ENDGAR).[6] To

---

[5]While the longer times at risk certainly contribute to the high recidivism rates. Other factors could have systematic effects. The literature suggests that an early entrance into a criminal career is a particularly important correlate of the incidence and frequency of adult arrests (Blumstein et al, 1986). However, it is not possible to distinguish the relative contributions of time at risk and juvenile justice system involvement in the current data, as all the offenders were incarcerated as juveniles and most had early involvement with the juvenile justice system.

[6] The four outcome variables are defined as follows:

$$RECID_i = \left\{ \begin{array}{l} 0 \text{ if no follow-up arrests for crime type } i \\ \\ 1 \text{ if any follow-up arrests for crime type } i \end{array} \right\}$$

$NUMARR_i$ = Total number of follow-up arrests for crime type i

$$FREQ_i = \frac{NUMARR_i}{RISKTIME - ENDGAP_i}$$

is the rate of arrests for crime type i during time free in the follow-up, where $ENDGAP_i$ is the length of time free between the last follow-up arrest for crime

(continued...)

assess differences across the subsamples, the GLS analysis (weighted by subgroup size) regressed the mean outcome within subgroups on the scale score, institution, and crime type of the subgroups (Table 3).

TABLE 3, About Here

Although not reported in Table 3, the various outcomes differed significantly over crime types. In contrast to the level of activity in the residual category of "Other" crime types, which is reflected in the constant term of the regression, the various outcomes were all lower for violent offenses, robbery, property offenses, or drug offenses. The other control variable, Scale Score, is reported in Table 3. At the level of subgroups, Scale Scores are strongly related to the various outcome variables ($p \leq .05$

---

[6](...continued)
type i and the end of the follow-up period. If the denominator is 6 months or less, reflecting a very short criminally active period, FREQ is treated as missing.

$$ENDGAR_i = \frac{ENDGAP_i}{RISKTIME}$$

is the proportion of all time free that follows the last arrest for crime type i. This variable is intended as an indicator of termination of offending. When $ENDGAR_i$ is close to zero, the offender is more likely to have remained active in crime type i throughout the follow-up. As $ENDGAR_i$ gets larger with values close to 1, the offender is more likely to have terminated offending in crime type i before the end of the follow-up.

in all but one of the analyses). For example, a higher mean RAND score for a subgroup is associated with a higher proportion of recidivists, an average of more arrests per offender, a higher average arrest frequency by active offenders, and longer average periods of continued offending (i.e., shorter ENDGAR) by members of that subgroup.

After controlling for crime type and the background differences reflected in scale scores, Table 3 indicates that numerous significant differences persisted between outcomes in the Preston subsample and those in the Fricot subsample (which is reflected in the constant term). Outcomes in the YCRP subsample, by contrast, are generally more similar to Fricot. Based on these results, we decided to combine the YCRP and Fricot subsamples into the YCOT analysis sample, but to maintain a separate Preston analysis sample.[7] This procedure combines in the YCOT sample the sub-samples of offenders who were younger when institutionalized as juveniles, and who were exposed to various experimental treatment options. The youths who were older when incarcerated, who had more extensive prior records, and who were committed to a more traditional juvenile training school at Preston were analyzed separately in the PRES sample.

The combined YCOT sample contains 1,079 former CYA wards who

[7] Similar GLS regressions were performed to detect subgroup differences for the three cities in the DOL data (Albuquerque, Miami, and New York City), and between prisoners and probationers in the P&P convicted sample. No other strong subgroup differences were found. Thus, the three cities were combined to form a single DOL dataset, as were prisoners and probationers to form a single P&P dataset.

were arrested after their 18th birthdays (see Table 2). The 16% of the YCOT sample who were incarcerated at Fricot Ranch between 1960 and 1963 began their criminal careers early; the median age at that incarceration was 10.9 years. The remaining 84% of the YCOT sample participated in experimental studies of the effectiveness of transactional analysis (O.H. Close Institution) and behavior modification (Karl Holton Institution) between 1969 and 1971. Their median age while incarcerated as juveniles was 16.6 years. The mean age at the target arrest for the combined YCOT sample was 18.7 years. Some 69% of the individuals in the YCOT sample were subsequently arrested for index property offenses, 27% for robbery, and 35% were arrested for an index violent offense (excluding robbery) during a follow-up period that averaged 7.7 years.

The PRES sample (described in Table 2) consisted of the 1,596 former CYA wards at the Preston institution who had target arrests after their 18th birthday. The median age of individuals when they were incarcerated at Preston in 1966-67 was 17.6 years, and the mean age at the target arrest was 18.6 years. These offenders were subsequently followed for an average of 10.8 years after the target arrest. The follow-up period was somewhat longer than that available in the YCOT sample and the recidivism rates were somewhat higher: 75% subsequently arrested for property offenses, 36% for robbery, and 45% for violent offenses (excluding robbery). It is important to reiterate that both the PRES and YCOT samples had substantially longer follow-up periods, which are likely to be a

factor in their substantially higher recidivism rates when compared to the DOL and P&P samples.

The four datasets share several features that were critical to the research reported here. They all involve large samples of offenders, which is essential for developing crime-specific estimates of offending patterns. Even the large initial samples available here quickly dwindle in size as one focuses only on the active offenders recidivating within specific crime categories. The datasets were also rich in the background variables needed for calculating individuals' scores on the four prediction scales, although no dataset perfectly supported all of the scales. Finally, all the samples included sufficient follow-up periods to reasonably observe subsequent offending--operationalized by arrests--if it occurred.

Consistent with our intent when selecting the data, the different samples varied considerably in their background variables. Table 4 contrasts the samples on two key variables used in the prediction scales. The level of prior drug or alcohol problems ranged from only 5% of the serious juvenile offenders who had been in special CYA treatment facilities (YCOT) to 39% of the convicted sample (P&P). The extent of prior arrests also varied considerably. It was lowest for the young adult arrestees from the jobs program samples (DOL) and highest, again, for the sample of convicted offenders (P&P). To some extent, the more extensive prior problems found in the convicted sample reflects their older age at the time of the target event, and thus their longer time at

risk of arrest and of drug or alcohol abuse before the target event.

TABLE 4, About Here

Comparisons across the datasets (Figure 2) illustrate the large variation in the predictions made by the RAND and CGR scales. The SFS and INSLAW scales made more similar predictions across the four datasets. There is also considerable variability in predictions across scales applied to the same dataset. In YCOT, for example, predicted high risk offenders ranged from 10% for the RAND scale to 51% for the CGR scale.

FIGURE 2, About Here

THE ANALYSIS

The traditional measures of prediction accuracy are well established: Total Prediction Accuracy Rate (TPAR), and three indicators of the frequency of errors among predictions. These error-based measures of prediction accuracy are the Total Error Rate among all predictions (TER = 1-TPAR), the False Positive Rate (FPR) among predicted successes, and the False Negative Rate (FNR) among predicted failures. Relying on these traditional measures to assess the accuracy of prediction scales poses a number of problems, especially when comparing the performance of different

scales on the same dataset, or when examining the performance of a single scale across different populations.

First, these measures of accuracy are all affected by the Base Rate, BR (the rate at which truly high risk individuals are present in a sample) and by the Selection Rate or Selection Rate, SR (the rate at which high risk individuals are predicted to be present in the sample.[8] In general, as Base Rates increase and Selection Rates decline, False Positive Rates (FPR's--of truly low-risk offenders who are incorrectly classified as high-risk by a scale) decline, while False Negative Rates (FNR's--of truly high-risk offenders who are incorrectly classified as low-risk) increase. The accuracy measures that are obtained in any application of prediction scales are thus highly dependent on specific features of the data. Furthermore, since the two types of prediction errors move in opposite directions, it is difficult to develop a prediction instrument that simultaneously minimizes both types of error.

The inadequacies of traditional prediction accuracy measures are particularly salient when different empirical scales are used to make predictions about the risk posed by criminal offenders.

---

[8] In the scale construction phase, the selection rate is a decision variable that is freely determined by the analyst. The selection rate may be small, with high-risk offenders classified as being restricted to only a small portion of the sample. Conversely, the selection rate may be large, with increasing fractions of the sample being classified as high-risk individuals. Once a scale cut-point has been designated, and the analysis of the scale moves beyond the construction phase to applications in new datasets, where the selection rate (like the base rate) is exogenously determined by sample characteristics.

The sample and scale dependence of these measures undermines attempts to make comparisons among scales. When the same scale is applied to different data samples, variation in the base rates can make it difficult to compare scale performance across the samples. Similar problems occur when assessing predictions from multiple scales in a single dataset. The classification rules for different scales (especially the criterion variables and the cut-points used to identify distinct risk groups) may lead to widely varying selection rates for the different scales.

Figure 3 illustrates the variability in accuracy measures for the four study scales in the four datasets. Each line plot presents the error rates -- either False Positive or False Negative -- that result when the four scales are used to predict rearrest violent index offenses in a single dataset. The circles along the solid lines represent the four scales' False Positive errors and the X's along the broken lines represent their False Negative errors. When the four scales were applied to the same dataset with a single base rate (comparing points <u>within</u> any single plot) the lines were relatively flat, indicating there was little difference in either the False Positive Rate or in the False Negative Rate as the Selection Rate varies among the scales.

FIGURE 3, About Here

The differences in error rates are large across datasets (reading vertically across the different lines in Figure 3). All

the scales were substantially in error when identifying recidivists in violent offenses in the P&P or DOL datasets, making about 9 in 10 False Positive predictions in these low base rate datasets. Alternatively, when the four scales are applied to the PRES (Preston) dataset, there are relatively fewer errors in classifying recidivists (about 45% FPR's). Less than 10% of the offenders recidivated to violent offenses in the P&P or DOL samples, while 45% recidivated in the PRES sample. This illustrates the well-known relationship between False Positive Rates and the Base Rate of an outcome: low Base Rates result in larger FPR's, while high Base Rates result in smaller FPR's. Similarly, the inverse relationship between False Positive and False Negative classification rates, described above, can be observed in Figure 3. Datasets with high False Positive Rates tend to have low False Negative Rates, and conversely.

Most existing criminological prediction efforts have focused on minimizing false positive errors when identifying recidivists because the consequences are so severe for individuals who are incorrectly classified. From this perspective, all the scales seem better suited for application to the PRES dataset (where the FPR's are about 45%), than to the P&P or DOL datasets (where the FPR's

are about 90%).[9]

As concern for protecting public safety has increased in the face of increasingly scarce prison resources, more emphasis has been placed on the objective of minimizing false negative errors. Under this alternative criterion, the scales all perform best on the P&P and DOL data sets (with FNR's of 6% for recidivism in violent offenses). Likewise, the scales perform poorly with respect to FNR's on the PRES dataset (FNR's of 40% for violent recidivism).

The results in Figure 3 illustrate the problems that are created by sample and scale dependence in traditional measures of prediction error. In choosing among prediction scales, or in comparing the performance of one scale across datasets, it would be desirable to remove the differences in accuracy that are due to changes in the Base Rate or in the Selection Rate.

TRADITIONAL MEASURES OF ERROR

In moving to sample-independent measures of accuracy, we first identify any constraints on the possible range of accuracy. This requires examining the algebra of the relationships among error rates, base rates, and selection rates. For a dichotomous outcome variable, predictions can be partitioned using a simple 2 X 2 table

[9] The same general pattern across datasets also exists for recidivism rates in other crime types. FPR's are always highest for offenders in the P&P and DOL datasets (at around 65% for property offenses, 85% for robbery, and 40% over all offenses). They are lower for offenders in the PRES dataset (at 20% for property offenses, 55% for robbery, and under 5% over all offenses).

with a total of three degrees of freedom (Figure 4). The sample characteristics found in the margins (SR and BR) reflect two degrees of freedom and result in constraints on the number of errors that are possible in the sample. In particular, false positive and false negative errors can be no larger than the smallest marginal for each cell.

As illustrated in Figure 4, the upper constraint on the number of false positive's is determined either by the Selection Rate (SR) or by (1-BR), whichever is smaller. Since the False Positive Rate (FPR) is defined relative to the entries in the SR column, the maximum FPR is either 1 when SR≤(1-BR), or the fraction (1-BR)/SR. Similarly, the False Negative Rate is defined relative to the entries in the (1-SR) column, and the maximum FNR is either 1 when (1-SR)≤BR, or the fraction BR/(1-SR) otherwise.

FIGURE 4, About Here

The number of errors is also constrained from below. When SR>BR, it is impossible to have no False Positive errors because that would require a total of [SR*N] True Positives, and this exceeds the row marginal. Thus, the number of False Positive errors can not be smaller than [(SR - BR)*N] when SR > BR. Likewise, when BR > SR, the smallest number of False Negative errors is [(BR - SR)*N]. Dividing each of these errors by the appropriate column marginals yields the minimum bounds on the two types of error rates.

The various constraints identified affect the range of possible values for the observed error rates of prediction scales. Figure 5 illustrates these constraints on the ranges of possible error rates when predicting subsequent Index Property arrests for the convicted felons in the P&P sample. Even though the observed FNR's are low across all 4 scales, these error rates are sharply constrained and cannot fall outside of a very low and narrow range of values. When the INSLAW scale was applied to this sample, for example, the FNR could never exceed 31%, and it was also constrained from below so that it could never be smaller than 7%.

FIGURE 5, About Here

Such low and narrow bands of potential False Negative error are comforting in criminal justice policy applications where concern for public protection makes these errors highly visible as well as politically salient. Identifying the range of potential False Negative errors may also be useful in scale construction. A limited range of possible FNR's increases confidence in the expected accuracy of prospective predictions about future offending in operational populations that are characterized by similar BR's. It is important to note, however, that FNR's are low and narrowly constrained only when BR's are also low.

The range of possible False Positive errors in Figure 5 is larger, but it is also constrained in some cases[10]. When the SFS81 scale was used to predict Property arrests in the P&P data, for example, the FPR could never fall below 30%. Even if the SFS81 scale were perfect in classifying truly high-risk offenders (i.e., the True Positives were all correctly identified as high-rate offenders by the scale's selection rate criterion), the fact that SR>BR will result in an excess of incorrectly predicted high-rate offenders (i.e., the True Positives are not the only offenders identified by the Scale Selection Rate criterion).

From this discussion it is apparent that prediction error can be minimized only within a mathematically determined constraint. In most operational settings that constraint will mean that some prediction error will occur, as such error can only be eliminated when BR=SR, and this is a highly unlikely outcome. This mathematical relationship underlies the frequent observation that it is desirable to maximize the congruence between a sample BR and the SR of a prediction scale.

ALTERNATIVE METHODS OF ERROR

Aside from the constraints on minimum and maximum errors, Figure 5 also contrasts observed accuracy with the random accuracy that occurs when a scale is independent of recidivism, and contributes no information beyond that found in the sample's Base

---

[10]The pattern of error ranges in Figure 5 will be reversed when BR's are high. Then False Positive errors will be relatively low and constrained within narrow ranges, compared with much wider ranges for False Negative errors.

Rate (BR). Among those predicted to be recidivists using a random classification rule, the fraction BR are expected to be True Positives and (1-BR) will be False Positives. Likewise, under a random classification rule, the fraction (1-BR) of predicted non-recidivists are expected to be True Negatives, while BR will be False Negatives. The range of minimum and maximum error rates, as well as random accuracy will, of course, vary as the sample-specific Base Rates and the related scale-specific Selection Rates change.

The sample dependency in the traditional measures of False Positive and False Negative errors is graphically depicted in Figures 3 and 5. This dependence makes these measures of prediction error (and the related Total Error Rate) difficult to interpret when comparing the accuracy of either different scales that vary in their Selection Rates, or in comparing the accuracy of the same scale across different datasets whose Base Rates vary (thereby creating variation in the Selection Rates). Such comparisons require a method that produces standardized information across scales and datasets.

One such standardized measure is the Relative Improvement Over Chance (RIOC) statistic (Loeber and Dishion, 1983). This statistic (eq. 1) contrasts the improvement in accuracy achieved beyond "random accuracy", relative to the full potential for such improvement. The RIOC statistic calibrates the observed improvement above random accuracy with respect to the range of possible improvement to maximum accuracy. By explicitly taking account of

the roles of Base Rate and Selection Rate in determining random and maximum accuracy, the RIOC effectively avoids the sample and scale dependence of traditional accuracy measures.

$$RIOC = \frac{\text{Observed Accuracy} - \text{Random Accuracy}}{\text{Maximum Accuracy} - \text{Random Accuracy}} \quad (1)$$

One criticism of the RIOC statistic has been that by focusing on total errors, the statistic places equal weight on the False Positive and False Negative errors that enter into this total (Farrington and Tarling, 1985). To allow for potentially different levels of concern for the two types of error, we have computed variants of the RIOC separately for FPR's and FNR's. The $RIOC_+$ (FPR) and the $RIOC_-$ (FPN) statistics separately standardize the observed accuracy among those classified as high risks and those classified as low risks, respectively. Each statistic ranges between 0.0 and 1.0, and can be interpreted as the proportional or percentage improvement in accuracy toward the maximum possible accuracy.

RESULTS

All four scales were applied to each of the four datasets, and the resulting FPR, FNR, and RIOC's for various crime types are reported in the Appendix. The separate $RIOC_+$ and $RIOC_-$ statistics for predictions of recidivism in property offenses in the P&P dataset are presented for illustrative purposes in Table 5.

Table 5, About Here

The the $RIOC_+$'s and the $RIOC_-$'s are identical for each Scale, despite the large differences in their potential ranges (Figure 3).[11] The accuracy achieved for False Positives and False Negatives represents the same relative level of improvement within their respective ranges. While the FPR and FNR thus appear very different in their absolute magnitudes, the two error rates are actually very similar relative to chance accuracy and maximum possible accuracy in a dataset (see Figure 5). In Table 6 we

---

[11] As reported in note C of the Appendix, the $RIOC_+$ and the $RIOC_-$ statistics are analytically identical. This symmetry for positive and negative predictions is consistent with the finding in Farrington and Loeber (1989) that the ordinary unweighted RIOC is equal to a weighted RIOC obtained when errors and correct predictions are weighted to reflect different levels of concern for the various types of classifications. A weighted RIOC, for example, would permit varying levels of concern for False Positive and False Negative errors, as well as for correct predictions of true positives (i.e., recidivists) and true negatives (i.e., non-recidivists).

The symmetry for positive and negative predictions results from the limits on the degrees of freedom available in a 2 x 2 table. With three degrees of freedom possible, and two of these used by the Base Rate (BR) and the Selection Rate (SR) observed in a dataset, only one degree of freedom remains within the 2 x 2 table. In Figure 4, for example, knowing only the number of False Positives (FP's) among predicted recidivists along with the Base Rate and the Selection Rate, it is sufficient to fully specify all other entries in the table, including those for predicted non-recidivists. Thus, given the Base Rate the Selection Rate for a 2 x 2 table, the errors in predicting non-recidivists can be directly inferred from the errors in predicting recidivists, or conversely. This dependency across the two predictions yields the symmetry between positive (recidivist) and negative (non-recidivist) predictions in the $RIOC_+$ and $RIOC_-$ statistics.

examine a case where for any single scale there are large variations in the magnitude of the FPR (typically ranging from about 20% to 65%). After standardizing for the underlying differences in the possible FPR range, the corresponding RIOC statistics are much more similar across datasets. These results contradict the conventional wisdom that it is impossible simultaneously to improve both False Positive and False Negative errors. While the absolute magnitudes of FP and FN error rates move in opposite directions, so do the constrained ranges of possible errors. As a result, both measures experience identical relative improvement in accuracy.

TABLE 6

In Table 6, the accuracy of recidivism predictions generally exceeds chance for all of the scales when applied to convictees in the P&P sample and to inmates in the PRES and YCOT samples.[12] By contrast, it was very difficult to predict rearrest recidivism with any significant accuracy among the young adult arrestees in the DOL

---

[12]Farrington and Loeber (1989) provide an estimate of the variance of an RIOC statistic. (The expression for the variance is found in note d of the Appendix.) The ratio of the RIOC statistic to the resulting standard error (obtained from the square root of the variance) is distributed as a standard normal variable. This ratio can be used directly to assess the likelihood that the difference between a RIOC statistic and zero could have occurred by chance. The variance estimator also provides the basis for assessing the significance of the difference between any two RIOC values.

sample, largely because of the inappropriateness of the statistical test when the number of recidivists is very small (See appendix note g). Only the CGR Scale--which was originally developed on arrestee samples--ever exceeds random accuracy with the DOL arrestee data.

The very low recidivism rate of the DOL arrestee sample, limits our ability to compare the relative accuracy of each scale across a range of alternative samples. Nevertheless, the finding that the CGR scale was the only scale that was suitable over the full range of samples examined here suggests a tentative hypothesis that requires further testing on other arrestee samples. On the one hand, it appears that scales developed on more broadly representative offender samples, like arrestees, can be usefully applied to more highly selected samples like inmates. Scales that are developed on more selected samples, on the other hand, appear to be less suited for application to more general offender samples.

CONCLUSIONS AND IMPLICATIONS

It has long been recognized that raw False Positive and False Negative error rates are inappropriate bases for determining the accuracy of classification instruments. These error rates are highly sample dependent, varying in magnitude with changes in the Base Rate and Selection Rate in any particular sample. One especially noteworthy problem with FPR and FNR error rates has been their movement in opposite directions as the SR or BR change. For example, FPR decreases while FNR increases as the BR increases

and/or the SR decreases. This has led to the inference that it is difficult to improve both types of error simultaneously.

The main insight provided by the analysis in this paper is that the BR and SR implied by a data sample not only affect the absolute magnitude of the FPR and FNR; they also place constraints on the possible range of FPR or FNR error rates. In particular, the limiting range may, and usually does, fall short of the natural boundaries between zero and one.

The constraints on the possible values of FPR and FNR error rates mean that observed FPR's and FNR's may differ widely in magnitude without reflecting substantive differences in accuracy over different scales or varying datasets. Furthermore, the constraints on FPR and FNR are determined by the sample Base Rate and Selection Rate. Hence, as the BR or SR for a data sample change, so do the constraints on the error rates.

Shifts in the constraints mean that changes in opposite directions in the magnitudes of FPR and FNR error rates do not reflect opposite changes in accuracy. On the contrary, a reduction in the magnitude of FPR that is accompanied by an increase in the magnitude of the FNR can represent improvement in accuracy for both Positive and Negative predictions. While the absolute magnitude of the FNR may have increased, the associated constraints on possible values of the FNR also have shifted, resulting in the same relative improvement in accuracy within the respective ranges of FPR and FNR error rates. Alternatively, both error rates may have gotten worse, despite the apparent decrease in the FPR. The actual

direction of changes in accuracy will be determined by the nature of the changes in constraints on the two errors.

This finding runs counter to the common wisdom concerning the interpetation of False Negative and False Positive error. However, it highlights the sensitivity of these traditional measures to misinterpretations related to the network of relationships between a sample base rate and scale selection rate. From these findings we conclude that the RIOC measure, which is data and scale independent, should be used to supplement decision making concerning all criminal justice prediction scales. Use of this measure will also provide a mechanism by which the performance of a variety of prediction instruments, applied to datasets with widely diverging case rates, can be compared.

The RIOC statistic has already been described in criminological literature, and is easily accessible to the research community. It provides a measure of accuracy that is standardized relative to the varying constraints on accuracy. Being free of such data dependencies, the RIOC is a powerful indicator of relative accuracy for both recidivist and non-recidivist predictions.

References

Blumstein, A., J. Cohen, J. Roth, and C. A. Visher (eds.)

1986 Criminal Careers and "Career Criminals", Vol I. Washington, D.C.: National Academy Press.

Center for Governmental Research

1982/3 An empirical and policy examination of the future of pretrial release services in New York State, Vols. II and III. Report prepared for the New York State Division of Criminal Justice Services by the Center for Governmental Research Inc., 37 South Washington Street, Rochester, NY 14608.

Capas, J. and Loeber, R.

1989 The statistical properties of the index: Relative improvement over chance (RIOC). British Journal of Mathematical and Statistical Psychology (in press).

Farrington, D.P. and Loeber, R.

1989 Relative improvement over chance (RIOC) and "phi" as measures of predictive efficiency and strength of association in 2 x 2 tables. Journal Quantitative Criminology 5: 201-213.

Farrington, D.P. and Tarling, R.

1985 Criminological prediction: the way forward. In D. P. Farrington and R. Tarling (eds.) Prediction in Criminology (pp. 258-269). Albany, NY: State University of New York Press.

Gottfredson M.R. and Gottfredson, D.M.

1980 Decisionmaking in Criminal Justice: Toward the Rational Exercise of Discretion. Cambridge, Mass.: Ballinger.

Greenwood, P. with A. Abrahamses

1982 Selective Incapacitation. Santa Monica, CA: The RAND Corporation.

Haapanen, R. and Jesness, C.F.

1982 Early identification of the chronic offender. Report prepared for the National Institute of Justice, U.S. Department of Justice by the California Department of Youth Authority, Sacramento, CA.

Haapanen, R.

1988 Selective incapacitation and the serious offender: A longitudinal study of criminal career patterns. Report prepared for the National Institute of Justice, U.S. Department of Justice by the California Department of Youth Authority, Program Research and Review Division, 4241 Williamsbourgh Dr., Sacramento, CA 95823.

Hoffman, P.B.

1983 Screening for risk: A revised salient factor score. Journal of Criminal Justice 11: 539-547.

Loeber, R. and Dishion, T.

1983 Early predictors of male delinquency: A review. Psychological Bulletin 94: 68-99.

Meehl, P.E. and Rosen, A.

1955 Antecedent probability and the efficiencies of psychometric signs, patterns, or cutting scores. Psychological Bulletin 52: 194-216.

Petersilia, J. and Turner, S. with Peterson, J.

1986 Prison versus Probation in California: Implications for Crime and Offender Recidivism, Report #R-3323-NIJ prepared for the National Institute of Justice, U. S. Department of Justice. Santa Monica, CA: The RAND Corporation.

Rhodes, W., Tyson, H., Weekley, J., Conly, C., and Powell, G.

1982 Developing criteria for identifying career criminals. Report to the Department of Justice. INSLAW Inc., Washington, D.C.

Sadd, S., Kotkin, M., and Friedman, S.R.

1983 Alternative youth employment strategies project: Final report. Report prepared for the Employment and Training Administration, U.S. Department of Labor by Vera Institute of Justice, 377 Broadway, New York, NY 10013.

U.S. Parole Commission

1985 Parole Commission Rules (28 C.F.R. 2.1-2.63). November 4, 1985, U.S. Parole Commission, U.S. Department of Justice.

Wiggins, J.S.

1973 Personality and Prediction: Principles of Personality Assessment. Reading, Mass: Addison-Wesley.

Wilkins, L.T.

1969 Evaluation of Penal Measures. New York: Ramdon House.

Table 1

Characteristics of the Four Scales
Used for the Prediction Analysis

Characteristic	RAND	INSLAW	SFS81	CGR
Adult Criminal Record	+	+	+	+
Juvenile Criminal Record	+	+	+	+
Drug/Alcohol Use	+	+	+	
Age at Target Arrest		+	+	
Educational Attainment				+
Employment History	+*			+*

*Not available in all datasets

Table 2
Characteristics of the Four Datasets
Used for the Prediction Analysis

Dataset	Sample Characteristics	Target Event	Mean Age	Follow-Up (Years)	Recidivism by by Crime Type *
DOL (n=746)	Referrals to Jobs Programs	Arrest Before Referral	17.3	1.8	Property = 19% Robbery = 12% Violent = 7% Drugs = 6%
P&P (n=1,022)	Convicted Felons	Arrest Leading to Sampled Conviction	26.7	2.6	Property = 25% Robbery = 8% Violent = 5% Drugs = 10%
YCOT (n=1,079)	Serious Juvenile Offenders	First Arrest as an Adult	18.7	7.7	Property = 69% Robbery = 27% Violent = 35% Drugs = 35%
PRES (n=1,596)	Serious Juvenile Offenders	First Arrest as an Adult	18.6	10.8	Property = 75% Robbery = 36% Violent = 45% Drugs = 47%

*The FBI Uniform Crime Report definitions for index offenses were employed in classifying the offense crime types. The index property offenses are: burglary, larceny-theft, and motor vehicle theft. The index violent offenses are: murder and nonnegligent manslaughter, forcible rape, and aggravated assault. Robbery, an index violent offense, and drug sale or use were treated separately in this analysis.

Table 3

Direction and Significance of Coefficients from the Generalized Least Squares (GLS) Analysis of the CYA Data to Detect Differences Across the Subsamples[a]

Outcome Variable	Scale	Scale Score	Preston	YCRP	R^2
RECID	RAND	+***	+**	+*	.76
	INSLAW	+***	+***	+*	.89
	SFS81	-***	+***	+*	.84
	CGR	+***	+***	+	.83
NUMARR	RAND	+***	+**	+*	.89
	INSLAW	+*	+***	-	.89
	SFS81	-***	+***	+	.85
	CGR	+***	+***	-	.89
FREQ	RAND	+**	+*	-	.53
	INSLAW	0	-	+	.55
	SFS81	-**	-	+	.60
	CGR	+*	-*	-	.68
ENDGAR	RAND	-***	-*	-*	.70
	INSLAW	-***	-***	-	.89
	SFS81	+***	-***	-*	.82
	CGR	-***	-***	-	.86

[a]The functional form of the GLS regressions applied to subgroup data employed for each scale type is:

MEAN PREDICTED OUTCOME = b_0 + b_1 SCALE SCORE + b_2 PRESTON + b_3 YCRP + b_4 ROBBERY + b_5 DRUGS + b_6 VIOLENT + b_7 PROPERTY

In addition to other unspecified factors, the constant b_0 reflects the combined effects of the FRICOT institution and the residual category of "Other" crime types.

[b]Significance of regression coefficients using a 2-tailed test:

* $p \leq .05$
** $p \leq .01$
*** $p \leq .001$

Table 4

Selected Background Characteristics of the Datasets Used in the Prediction Analyses

Dataset	Sample	Prior Drug/ Alcohol Use	Average Number of Prior Arrests
DOL (n=746)	Referrals to Jobs Programs	17%	0.5
P&P (n = 1,022)	Convicted Felons	39%	3.6
PRESTON (n = 1,596	Serious Juvenile Offenders	5%	2.1
YCOT (n = 1,079)	Serious Juvenile Offenders	31%	2.4

Table 5

Accuracy of Alternative Scales in Predicting Rearrest Recidivism for Property Crimes in the P&P Dataset

Scale	FPR	$RIOC_+$	FNR	$RIOC_-$
RAND	58.5%	.222	21.8%	.222
INSLAW	67.9%	.097	23.1%	.097
SFS81	64.1%	.244	18.8%	.244
CGR	66.2%	.131	21.4%	.131

"Inmate" Scales

"Arrestee" Scales

Identical

Table 6

Accuracy of Each Scale in Predicting Rearrest Recidivism for Property Crimes Across the Four Datasets: RIOC (FPR)

	SCALE			
Sample	RAND	SFS81	INSLAW	CGR
DOL	.215[b] (64%)	-1.000[a,b] (100%)	.589[b] (33%)	.160** (68%)
P&P	.222*** (59%)	.244*** (64%)	.097** (68%)	.131*** (66%)
YCOT	.357** (20%)	.343*** (21%)	.060 (30%)	.116* (27%)
PRES	.258*** (19%)	.201** (20%)	.213** (20%)	.165** (22%)
Original Construction Sample	.345*** (48%)	.239*** (54%)	.743*** (15%)	.181*** (64%)

Note: Significance levels in a one-tailed z-test:

* $p \leq .05$
** $p \leq .01$
*** $p \leq .001$

[a]See note e in the Appendix for a discussion of negative RIOC's.

[b]See note g in the Appendix for a discussion of RIOC when cell frequencies are very small.

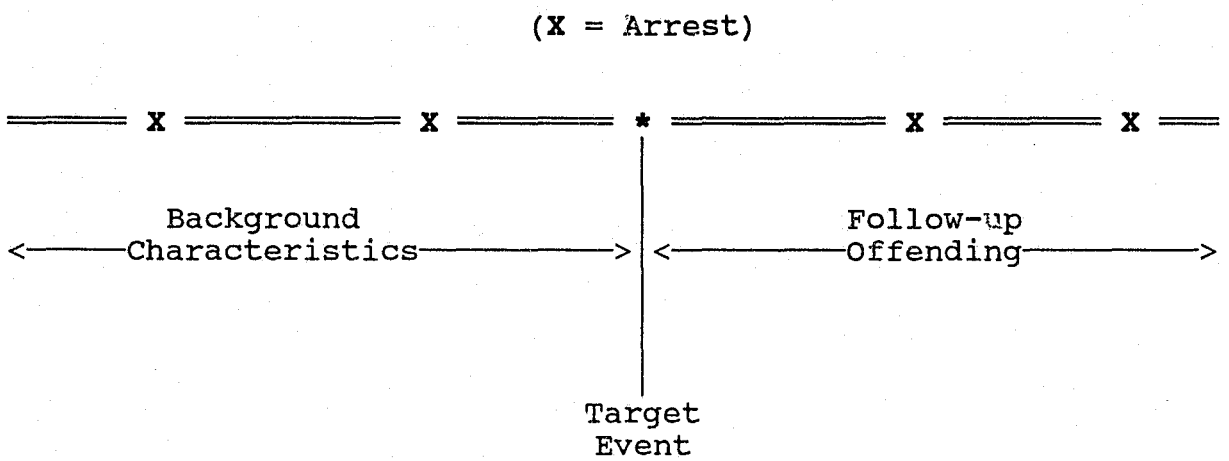


Figure 1. Longitudinal View of a Criminal Career

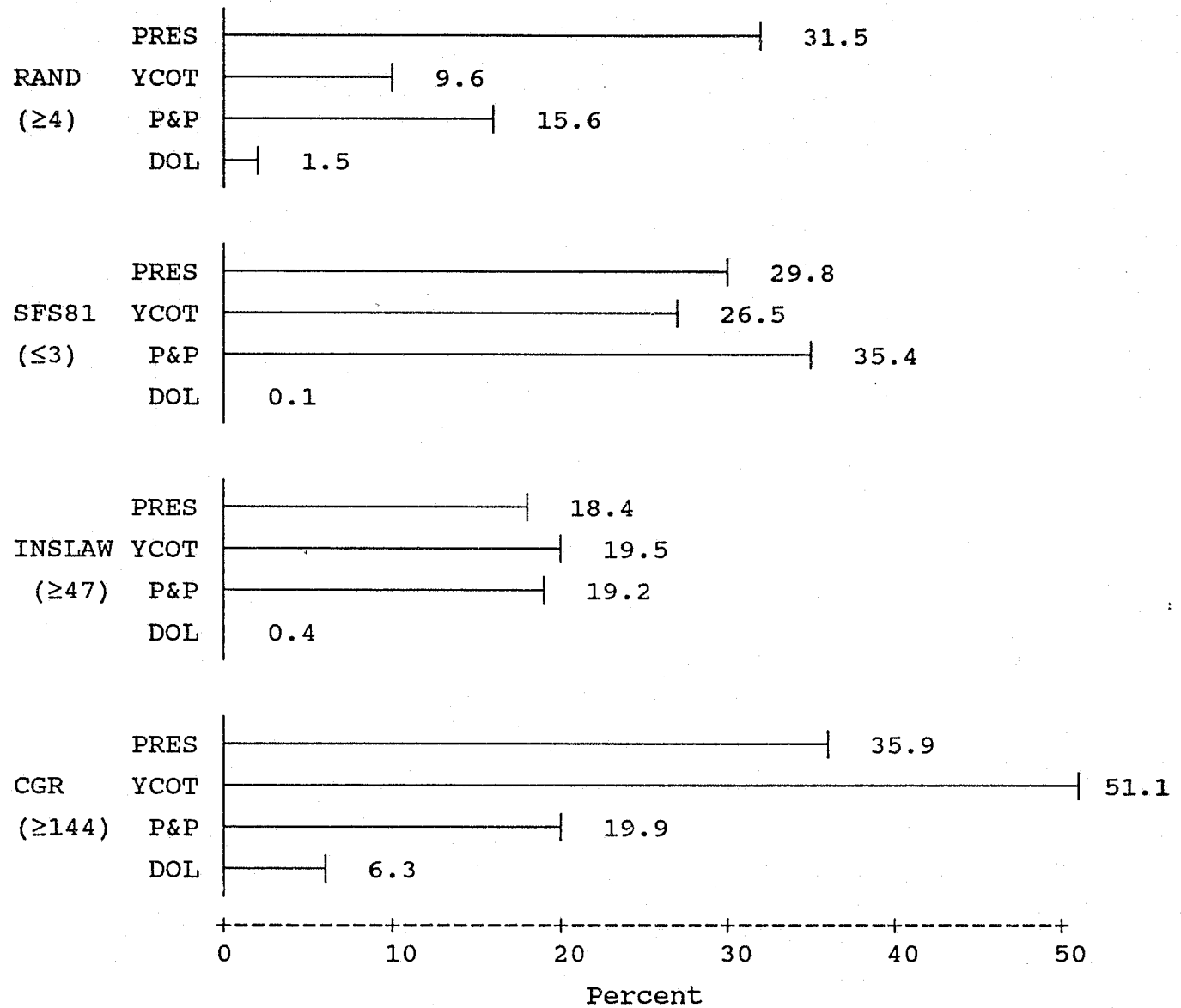
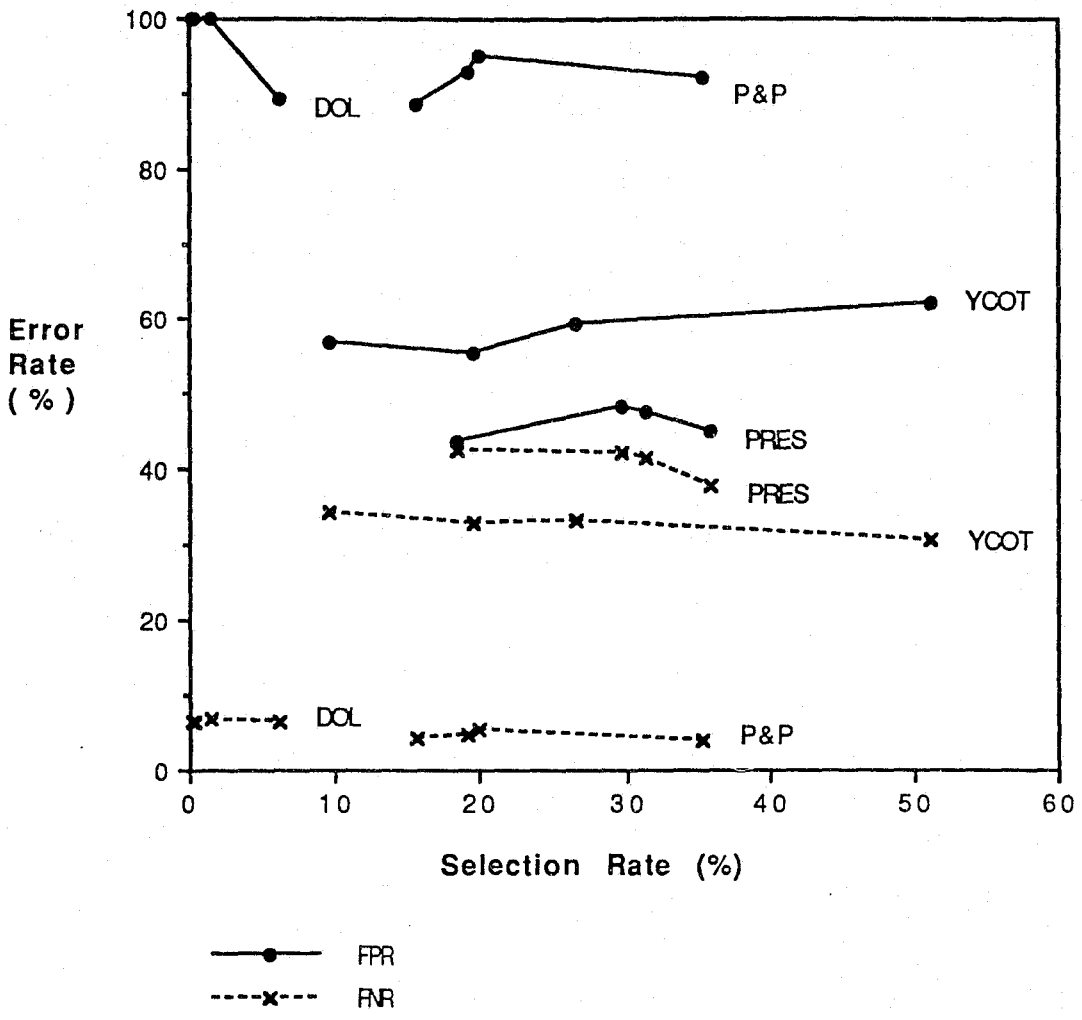


Figure 2. Percent of Each Sample Classified as High Risk by each Scale



Base rates:	
PRESTON	74.5%
YCOT	68.6%
P&P	24.9%
DOL	18.9%

Figure 3. Accuracy of Scales in Predicting Rearrest for a Violent Index Offense: Variability with Base Rates of Alternative Data Sets and With Selection Rate of Alternative Scales.

		Predicted: Recidivist	Predicted: Non-Recidivist	
Actual	Recidivist	TP	FN	BR=(TP+FN)/N
	Non-Recidivist	FP	TN	(1 - BR)
		SR = (TP+FP)/N	(1 - SR)	N

<u>False Positive Rate (FPR)</u>

$$FPR = FP/(SR \cdot N)$$

$$FPR_{MAX} \leq \begin{cases} 1 \text{ if } SR \leq (1 - BR) \\ \dfrac{1 - BR}{SR} \text{ otherwise} \end{cases}$$

$$FPR_{MIN} \geq \begin{cases} 0 \text{ if } SR \leq BR \\ \dfrac{SR - BR}{SR} \text{ otherwise} \end{cases}$$

<u>False Negative Rate (FNR)</u>

$$FNR = FN/[(1 - SR)N]$$

$$FNR_{MAX} \leq \begin{cases} 1 \text{ if } (1 - SR) \leq BR \\ \dfrac{BR}{1 - SR} \text{ otherwise} \end{cases}$$

$$FNR_{MIN} \geq \begin{cases} 0 \text{ if } BR \leq SR \\ \dfrac{BR - SR}{1 - SR} \text{ otherwise} \end{cases}$$

Figure 4. Relationships Among Error Rates, Base Rates (BR), and Selection Rates (SR)

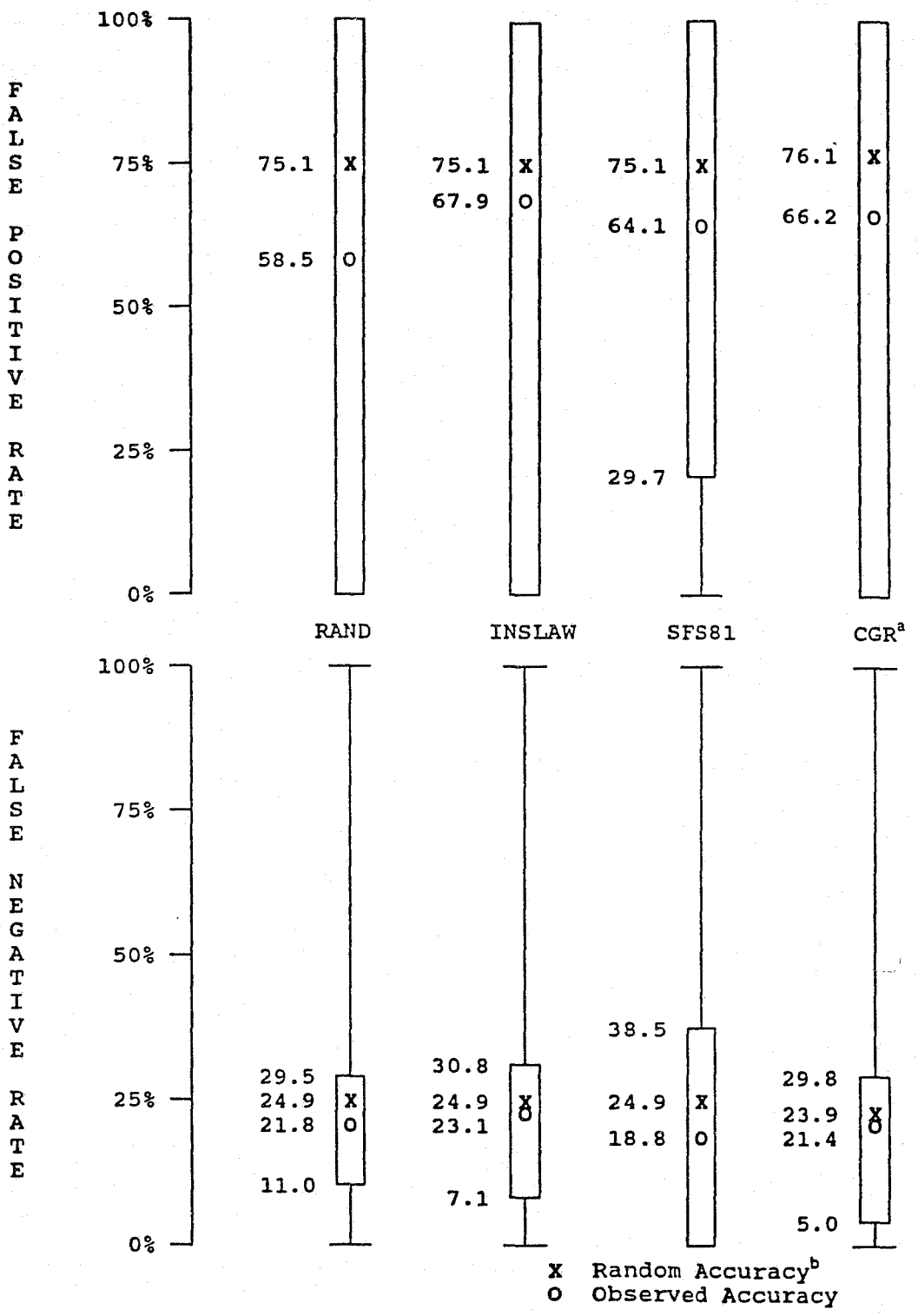


Figure 5. Constraints on Scale Accuracy in Predicting Recidivism (Rearrest) in Index Property Offenses by the P&P Sample (N=1002)

(Notes Continued on Next Page)

Notes To Figure 5.

[a]When the CGR scale was applied to these data, scale scores could be calculated for only 979 cases. This resulted in a slightly different BR, and consequently different random accuracy rates.

[b]For FP rate X = 1-BR; for FN rate X = BR.

Appendix

Predictive Accuracy by Scale and Dataset for Recidivism in Various Crime Types

Table A1. Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Violent Offenses (Murder, Rape, or Aggravated Assault)

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
RAND:				
(N)	(1596)	(1079)	(1022)	(746)
BR[b]	45.0	35.1	5.3	6.6
SR[b]	31.5	9.6	15.6	1.5
FPR[b]	47.6	56.7	88.7	100.0
MINFPR	0.0	0.0	66.6	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR[b]	41.6	34.3	4.2	6.7
MINFNR	19.7	28.2	0.0	5.2
MAXFNR	65.6	38.9	6.3	6.7
RIOC[c]	.135***	.126*	.211***	-1.000[e,g]
(S.E.)[d]	(.033)	(.069)	(.057)	(1.129)
INSLAW:				
(N)	(1596)	(1079)	(1022)	(746)
BR	45.0	35.1	5.3	6.6
SR	18.4	19.5	19.2	0.4
FPR	43.5	55.2	92.9	100.0
MINFPR	0.0	0.0	72.4	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	42.4	32.8	4.8	6.6
MINFNR	32.6	19.4	0.0	6.2
MAXFNR	55.1	43.6	6.5	6.6
RIOC	.209***	.149***	.083	-1.000[g]
(S.E.)	(.048)	(.046)	(.065)	(2.173)
SFS81:				
(N)	(1596)	(1079)	(1022)	(746)
BR	45.0	35.1	5.3	6.6
SR	29.8	26.5	35.4	0.1
FPR	48.1	59.4	92.0	100.0
MINFPR	0.0	0.0	85.1	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	42.1	33.2	3.8	6.6
MINFNR	21.6	11.7	0.0	6.4
MAXFNR	64.1	47.8	8.2	6.6
RIOC	.125***	.084*	.283**	-1.000[g]
(S.E.)	(.035)	(.037)	(.098)	(3.769)

Table A1. (Continued) Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Violent Offenses (Murder, Rape, or Aggravated Assault)

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
CGR:[f]				
(N)	(1056)	(830)	(979)	(746)
BR	43.9	34.3	5.3	6.6
SR	35.9	51.1	19.9	6.3
FPR	45.1	62.3	94.9	89.4
MINFPR	0.0	32.8	73.3	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	37.8	30.8	5.4	6.3
MINFNR	12.6	0.0	0.0	0.3
MAXFNR	68.5	70.2	6.6	7.0
RIOC	.195***	.103*	-.035	.044
(S.E.)	(.036)	(.049)	(.271)	(.037)

See notes at end of Appendix tables.

Table A2. Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Robbery

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
RAND:				
(N)	(1596)	(1079)	(1022)	(746)
BR[b]	35.8	27.2	8.2	11.5
SR[b]	31.5	9.6	15.6	1.5
FPR[b]	53.0	65.4	84.9	90.9
MINFPR	0.0	0.0	47.2	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR[b]	30.7	26.5	7.0	11.6
MINFNR	6.4	19.5	0.0	10.2
MAXFNR	52.3***	30.2*	9.7***	11.7
RIOC[c]	.174***	.101*	.154***	-.211[e,g]
(S.E.)[d]	(.028)	(.057)	(.045)	(.829)
INSLAW:				
(N)	(1596)	(1079)	(1022)	(746)
BR	35.8	27.2	8.2	11.5
SR	18.4	19.5	19.2	0.4
FPR	50.7	66.2	88.8	33.3
MINFPR	0.0	0.0	57.1	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	32.8	25.7	7.5	11.3
MINFNR	21.4	9.7	0.0	11.2
MAXFNR	43.9***	33.8**	10.2	11.6
RIOC	.210***	.090**	.087*	.623[g]
(S.E.)	(.039)	(.038)	(.051)	(.208)
SFS81:				
(N)	(1596)	(1079)	(1022)	(746)
BR	35.8	27.2	8.2	11.5
SR	29.8	26.5	35.4	0.1
FPR	56.1	65.7	87.6	100.0
MINFPR	0.0	0.0	76.8	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	32.4	24.7	5.9	11.5
MINFNR	8.6	1.0	0.0	11.4
MAXFNR	51.1***	37.1***	12.7***	11.5
RIOC	.126***	.096***	.281***	-1.000[g]
(S.E.)	(.029)	(.031)	(.077)	(2.768)

Table A2. (Continued) Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Robbery

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
CGR:[f]				
(N)	(1056)	(830)	(979)	(746)
BR	35.0	26.9	8.3	11.5
SR	35.9	51.1	19.9	6.3
FPR	56.5	68.9	92.3	83.0
MINFPR	2.4	47.4	58.5	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	30.3	22.4	8.4	11.2
MINFNR	0.0	0.0	0.0	5.6
MAXFNR	54.7	54.9	10.3	12.3
RIOC	.136***	.166**	-.070	.062
(S.E.)	(.031)	(.059)	(.213)	(.051)

See notes at end of Appendix tables.

Table A3. Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Property Offenses (Burglary, Larceny, or Auto Theft)

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
RAND:				
(N)	(1596)	(1079)	(1022)	(746)
BR[b]	74.5	68.6	24.9	18.9
SR[b]	31.5	9.6	15.6	1.5
FPR[b]	18.9	20.2	58.5	63.6
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	81.1	100.0	100.0	100.0
FNR[b]	71.5	67.4	21.8	18.6
MINFNR	62.8	65.2	11.0	17.7
MAXFNR	100.0***	75.9**	29.4***	19.2
RIOC[c]	.258***	.357**	.222***	.215[g]
(S.E.)[d]	(.063)	(.138)	(.042)	(.144)
INSLAW:				
(N)	(1596)	(1079)	(1022)	(746)
BR	74.5	68.6	24.9	18.9
SR	18.4	19.5	19.2	0.4
FPR	20.1	29.5	67.9	33.3
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	73.3	68.1	23.1	18.7
MINFNR	68.7	61.0	7.0	18.6
MAXFNR	91.3*	85.2	30.8**	19.0
RIOC	.213*	.060	.097**	.589[g]
(S.E.)	(.090)	(.091)	(.037)	(.278)
SFS81:				
(N)	(1596)	(1079)	(1022)	(746)
BR	74.5	68.6	24.9	18.9
SR	29.8	26.5	35.4	0.1
FPR	20.4	20.6	64.1	100.0
MINFPR	0.0	0.0	29.8	0.0
MAXFPR	85.5	100.0	100.0	100.0
FNR	72.3	64.7	18.8	18.9
MINFNR	63.7	57.3	0.0	18.8
MAXFNR	100.0**	93.3***	38.5***	18.9
RIOC	.201**	.343***	.244***	-1.000[e,g]
(S.E.)	(.066)	(.075)	(.040)	(2.070)

Table A3. (Continued) Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Property Offenses (Burglary, Larceny, or Auto Theft)

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
CGR:[f]				
(N)	(1056)	(830)	(979)	(746)
BR	73.8	69.0	23.9	18.9
SR	35.9	51.1	19.9	6.3
FPR	21.9	27.4	66.2	68.1
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	73.1	60.6	100.0	100.0
FNR	71.3	65.3	21.4	18.0
MINFNR	59.1	36.7	5.0	13.4
MAXFNR	100.0	100.0	29.8	20.2
RIOC	.165**	.116*	.131***	.160**
(S.E.)	(.069)	(.051)	(.036)	(.068)

See notes at end of Appendix tables.

Table A4. Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Drugs

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
RAND:				
(N)	(1596)	(1079)	(1022)	(746)
BR[b]	47.2	35.3	9.6	6.2
SR[b]	31.5	9.6	15.6	1.5
FPR[b]	45.4	57.7	86.8	90.9
MINFPR	0.0	0.0	38.4	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR[b]	43.9	34.6	8.9	6.1
MINFNR	23.0	28.4	0.0	4.8
MAXFNR	68.9	39.1	11.4	6.3
RIOC[c]	.139***	.108	.070*	.031[g]
(S.E.)[d]	(.035)	(.069)	(.041)	(.077)
INSLAW:				
(N)	(1596)	(1079)	(1022)	(746)
BR	47.2	35.3	9.6	6.2
SR	18.4	19.5	19.2	0.4
FPR	43.5	66.2	84.7	66.7
MINFPR	0.0	0.0	50.0	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	45.2	35.7	8.2	6.1
MINFNR	35.3	19.7	0.0	5.8
MAXFNR	57.9	43.8	11.9	6.2
RIOC	.175***	-.043[e]	.141**	.290[g]
(S.E.)	(.050)	(.084)	(.047)	(.148)
SFS81:				
(N)	(1596)	(1079)	(1022)	(746)
BR	47.2	35.3	9.6	6.2
SR	29.8	26.5	35.4	0.1
FPR	51.9	64.3	87.0	100.0
MINFPR	0.0	0.0	72.9	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR	46.9	35.2	7.7	6.2
MINFNR	24.8	12.0	0.0	6.0
MAXFNR	67.3	48.0	14.8	6.2
RIOC	.016	.005	.194**	-1.000[g]
(S.E.)	(.036)	(.037)	(.071)	(3.898)

Table A4. (Continued) Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Drugs

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
CGR:[f]				
(N)	(1056)	(830)	(979)	(746)
BR	50.8	35.5	9.3	6.2
SR	35.9	51.1	19.9	6.3
FPR	41.7	62.7	85.6	89.4
MINFPR	0.0	30.4	53.3	2.1
MAXFPR	100.0	100.0	100.0	100.0
FNR	46.5	33.7	8.0	5.9
MINFNR	23.2	0.0	0.0	0.0
MAXFNR	79.2	72.7	11.6	6.6
RIOC	.153***	.051	.135**	.049
(S.E.)	(.042)	(.048)	(.050)	(.037)

See notes at end of Appendix tables.

Table A5. Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Any Offense (Total)

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
RAND:				
(N)	(1596)	(1079)	(1022)	(746)
BR[b]	93.2	92.3	45.6	46.9
SR[b]	31.5	9.6	15.6	1.5
FPR[b]	3.2	3.8	30.8	45.5
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	21.5	79.8	100.0	100.0
FNR[b]	91.6	91.9	41.3	46.8
MINFNR	90.1	91.5	35.6	41.6
MAXFNR	100.0	100.0	54.0	47.6
RIOC[c]	.529***	.500[g]	.434***	.144
(S.E.)[d]	(.137)	(.323)	(.067)	(.281)
INSLAW:				
(N)	(1596)	(1079)	(1022)	(746)
BR	93.2	92.3	45.6	46.9
SR	18.4	19.5	19.2	0.4
FPR	2.0	4.8	40.3	0.0
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	36.7	39.5	100.0	100.0
FNR	92.2	91.6	42.3	46.7
MINFNR	91.7	90.4	32.7	46.7
MAXFNR	100.0	100.0	56.4	47.1
RIOC	.698***	.381*	.295***	1.000[g]
(S.E.)	(.196)	(.215)	(.059)	(.542)
SFS81:				
(N)	(1596)	(1079)	(1022)	(746)
BR	93.2	92.3	45.6	46.9
SR	29.8	26.5	35.4	0.1
FPR	5.9	3.8	37.6	100.0
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	22.7	29.0	100.0	100.0
FNR	92.9	90.9	36.4	47.0
MINFNR	90.4	89.5	15.8	46.8
MAXFNR	100.0	100.0	70.6	47.0
RIOC	.131	.500**	.309***	-1.000[e,g]
(S.E.)	(.143)	(.176)	(.039)	(1.063)

Table A5. (Continued) Predictive Accuracy by Scale[a] and Data Set for Recidivism (Rearrest) in Any Offense Total

SCALE STATISTICS	Data Set (Source) PRESTON (CYA)	YCOT (CYA)	P&P (RAND)	DOL (VERA)
CGR:[f]				
(N)	(1056)	(830)	(979)	(746)
BR	93.7	92.4	44.7	46.9
SR	35.9	51.1	19.9	6.3
FPR	1.8	4.7	46.2	29.8
MINFPR	0.0	0.0	0.0	0.0
MAXFPR	17.7	14.9	100.0	100.0
FNR	91.1	89.4	42.5	45.4
MINFNR	90.1	84.5	31.0	43.3
MAXFNR	100.0***	100.0***	55.9**	50.1***
RIOC	.709***	.379***	.165**	.439***
(S.E.)	(.158)	(.119)	(.058)	(.133)

See notes at end of Appendix tables.

Table A6. Predictive Accuracy for Scale[a] for Recidivism in Scale Construction Samples

Scale Statistics	RAND (Greenwood and Abrahamses, 1982)	INSLAW (Rhodes et al, 1982)	SFS81 (Hoffman, 1983)	CGR (Center for Governmental Research, 1982/3)
(N)	(781)	(1,708)	(3,955)	(1,557)
BR[b]	28.0	41.6	31.5	22.5
SR[b]	28.9	11.7	34.2	19.0
FPR[b]	48.2	15.0	54.0	63.5
MINFPR	3.1	0.0	7.9	0.0
MAXFPR	100.0	100.0	100.0	100.0
FNR[b]	18.4	35.9	24.0	19.2
MINFNR	0.0	33.9	0.0	4.3
MAXFNR	39.5	47.2	100.0	27.8
RIOC[c]	.345***	.743***	.239***	.181***
(S.E.)[d]	(.037)	(.056)	(.017)	(.028)

See notes at end of Appendix tables.

Note: Significance levels in a one-tailed z-test:

*p ≤ .05
**p ≤ .01
***p ≤ .001

[a]The designed cutpoints were used when applying the four scales to each dataset. Predicted recidivists included those individuals with the following scale scores:

RAND	≥	4
SFS81	≤	3
INSLAW	≥	47
CGR	≥	144

[b]The various statistics are defined as follows:

BR = Base Rate = Percent Recidivists in Total N

SR = Selection Rate = Percent Predicted Recidivists in Total N

FPR = False Positive Rate = Percent Non-Recidivists Among Predicted Recidivists

TPR = True Positive Rate = Percent Actual Recidivists Among Predicted Recidivists (TPR=1-FPR)

FNR = False Negative Rate = Percent of Recidivists Among Predicted Non-Recidivists

TNR = True Positive Rate = Percent Actual Non-Recidivists Among Predicted Non-Recidivists (TNR=1-FNR)

[c]RIOC, or "relative improvement over chance", is a measure of predictive accuracy defined in Loeber and Dishion (1983) as:

$$RIOC = \frac{\text{Observed Accuracy} - \text{Random Accuracy}}{\text{Maximum Accuracy} - \text{Random Accuracy}}$$

In a 2 x 2 table of predicted and actual recidivist outcomes, (like that in Figure 4), the FPR expected from a random classification is (1-BR) and the random FNR is BR.

When computed separately for predicted recidivists (+) and for predicted non-recidivists (-), it is found that:

$$RIOC_+ = RIOC_- = RIOC = \begin{cases} \dfrac{TPR - BR}{1 - BR} & \text{when } TPR \geq BR \text{ and } BR \geq SR \\ \dfrac{(TPR-BR)SR}{BR(1-SR)} & \text{when } TPR \geq BR \text{ and } SR > BR \end{cases}$$

[d]Farrington and Loeber (1989) derives the variance of the RIOC statistic from the standard normal sampling distribution of the ASR statistic (Haberman, 1973). This variance is transformed here in terms of sample BR and SR as:

$$VAR(R) = \begin{cases} \dfrac{BR(1-SR)}{SR(1-BR)} \cdot \dfrac{1}{N} & \text{when } TPR \geq BR \text{ and } BR \geq SR \\ \dfrac{SR(1-BR)}{BR(1-SR)} \cdot \dfrac{1}{N} & \text{when } TPR \geq BR \text{ and } SR > BR \end{cases}$$

for N = total cases in the sample. The standard error (S.E.) of the statistic is just the square root of this variance, and the ratio of RIOC divided by its standard error is distributed as a standard normal variable.

[e]When observed accuracy falls below random accuracy, the scale performs worse than random accuracy and observed accuracy moves in the direction of the minimum possible accuracy. In this case the RIOC is negative, varying between 0 and -1, and is calculated as:

$$\frac{\text{Observed Accuracy} - \text{Random Accuracy}}{\text{Random Accuracy} - \text{Minimum Accuracy}}$$

When computed separately for predicted recidivists (+) and predicted non-recidivists (-), it is found that:

$$RIOC_{+} = RIOC_{-} = \begin{cases} \dfrac{TPR - BR}{BR} & \text{when } TPR < BR \text{ and } 1-BR \geq SR \\ \dfrac{(TPR-BR)SR}{(1-SR)(1-BR)} & \text{when } TPR < BR \text{ and } 1-BR < SR \end{cases}$$

The variance of this negative RIOC statistic is derived from the variance of the standard normal ASR statistic as:

$$VAR(R) = \begin{cases} \dfrac{(1-SR)(1-BR)}{SR \cdot BR \cdot N} & \text{when } TPR < BR \text{ and } 1-BR \geq SR \\ \dfrac{SR \cdot BR}{(1-SR)(1-BR)N} & \text{when } TPR < BR \text{ and } 1-BR < SR \end{cases}$$

[f]The CGR scale invokes a variable on educational achievement. Data to support this variable were sometimes missing for individuals in the various datasets, and this accounts for the reduction in sample size compared to other scales.

[g]Copas and Loeber (1989) warn that the large sample properties invoked in the significance tests for the RIOC statistic, particularly the symmetric normal distribution for sample estimates of the RIOC, do not apply when any of the cell frequencies in the 2 x 2 table do not exceed 5. Cell frequencies of five or less occur in this case, and thus no significance levels are reported.