Accuracy of Recidivism Prediction Scales: A Technical Description of Data and Research Methods

by

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

August, 1995

[a]H. John Heinz III School of Public Policy and Management Carnegie Mellon University
[b]Department of Criminology, Indiana University of Pennsylvania

ACKNOWLEDGMENT

The research reported here was funded by grant #IJ-CX-86-0039 from the National Institute of Justice, U.S. Department of Justice. Points of view or opinions expressed are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice. The prediction scales used in this study were constructed by criminal justice researchers with the purpose of informing a series of operational decisions about individual defendants or convicted offenders. All the data used in this study were originally collected by other researchers and for other purposes. We are grateful to these individuals for making their work and their data available to us, and for their advice and assistance which greatly improved our research.

INTRODUCTION

Prediction has a lengthy and respected tradition in the history of criminology. This 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. Second, statistical prediction instruments have been adopted to inform decisions at important stages of the criminal justice process by making explicit predictions about offenders' "future expected behavior" (Gottfredson and Gottfredson, 1980).

The development of quantitative instruments for predicting 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, heightened concern for public protection, combined with mounting pressures on criminal justice system (CJS) resources, have resulted in intensified efforts to develop and implement prediction scales that will more effectively allocate scarce CJS resources among offenders (e.g., Greenwood and Abrahamses, 1982).

The purpose of the current study is to examine the transportability of operational criminal justice prediction scales across populations and applications. It is important to know whether scales developed for predicting offender behavior in one population can be employed in another population without substantial degradation in performance, and

whether scales developed for use at one criminal justice processing stage (decision context) can be used in another. If substantial prediction integrity can be maintained when prediction scales are transported across populations and decision contexts, it would mean that jurisdictions could adopt prediction scales developed elsewhere and thereby avoid the considerable costs associated with local scale development. To date, no rigorous large-scale studies have tested the transportability of criminal justice prediction scales.

The conventional wisdom concerning this possibility has been that a scale developed for one population and decision context will not have the same predictive integrity in a different population or for a different decision context. Given the current inefficiency of criminal justice prediction scales, differences in base rates and population characteristics usually are presumed to produce so many prediction errors that transporting scales is inadvisable. The study reported here investigates this presumption by examining the robustness of several criminal justice prediction scales in a variety of applications.

This study uses four prediction scales, originally designed to predict different outcomes. The vehicle for the analysis is four data sets from four different populations. The data sets are from populations different than those originally used to construct the scales. The analysis employs traditional measures of predictive efficiency (percent of correct predictions, FNR and FPR), the RIOC measure, and two measures we developed to facilitate this comparative analysis (MinTER* and CORIOC).

Table 1.

CHARACTERISTICS OF THE FOUR SCALES USED FOR THE PREDICTION ANALYSIS

(a) Intended Application of Each Scale

	CGR	INSLAW	RAND	SFS81
Intended Decision	Pretrial Release	Prosecution	Sentencing	Parole Release
Construction Population	State Arrestees	Federal Prisoners/ Probationers	State Prisoners	Federal Parolees
Criterion Outcome	Reappearance/ Rearrest	Rearrest	Reoffending (Self-Report)	Parole Revocation/ Reconviction

(b) Types of Offender Attributes in Each Scale

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

---

* Employment data are not available in some data sets

post-release recidivism experience of a sample of inmates released from Federal prisons, and currently is being used by the Parole Commission in making parole decisions.

The final scale (CGR) was developed as a model for pretrial release decisions in New York state jurisdictions outside New York City (Center for Governmental Research, 1982/83). This scale was constructed using a sample of defendants awaiting trial in selected New York State jurisdictions, some of whom were on pretrial release and others who were held in pretrial detention. Defendants were classified in terms of their subsequent court appearance and their rearrest experiences.

The types of offender attributes included in each scale are shown in Table 1(b). All the scales include a variable reflecting prior Adult Criminal Record, as well as a Juvenile Criminal Record variable. Indicators of Drug Use and 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 reflecting the offenders' recent Employment History. With the exception of Employment History, there is some basis for measuring all scale variables in each data set.

The scales were selected to be sufficiently diverse for examining the question of transportability. All scales rely on adult and juvenile criminal behavior, but otherwise invoke somewhat different information. Heterogeneity was further enhanced by the

different variable definitions and weighting schemes the scales employed. As seen in Appendix A, only the RAND scale uses simple Burgess weights (scores of 0 or 1 for each variable). The INSLAW and SFS81 scales employ integer weights that were derived from regression-based analyses, while the CGR scale uses the actual analysis weights derived from a logistic regression analysis. Finally, the scales vary in their intended application (they are targeted at different decision contexts and different populations).

THE DATA SETS

Four data sets 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). In addition, these data sets were selected because they adequately supported the prediction scales identified for study (see "Scale/Data Set Fit," below). Figure 1 shows the general strategy used in analyzing these data. All the data sets contain longitudinal information on individual offending as indicated by criminal justice interventions (arrest or filed charges), as well as other individual attributes. An adult criminal justice intervention was defined as an arrest or the disposition of an arrest of persons who have passed their 18th birthday or reached the age of majority.[1] The arrest associated with the first adult intervention was designated as the "target event" and was used to trigger the application of the prediction scales for qualifying sample members. Data on attributes prior to the target event were used as measures of the background

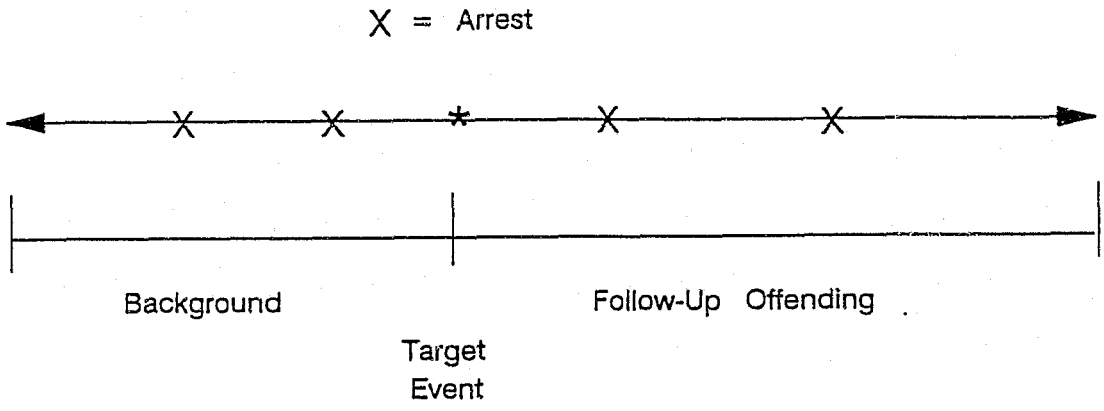
[1] The age of majority for the purposes of criminal prosecution was 18 in all jurisdictions except New York where persons aged 16 or older are normally prosecuted in adult criminal court.

characteristics that entered an individual's scale score, and offenses occurring after the target event were used to construct the follow-up outcome variables.

The general characteristics of the four data sets are described in Table 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 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. Sample members were followed for an average of 1.8 years after their target arrest and, during this follow-up period, 19% of those in the analysis sample were arrested for an index property offense, 12% for robbery, and 7% for an index violent offense other than robbery.

The remaining 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 during 1980 in Alameda and Los Angeles counties. These counties contributed about one-third of California's total felony

Figure 1. Longitudinal View of a Criminal History



convictions that year. The arrest associated with this 1980 conviction was used as the target event for applying the prediction scales.

The combined P&P data contains 1,022 individuals and includes the oldest offenders among the four analysis data sets, averaging age 27 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 made up of convicted felons and DOL was a sample of arrestees, in the relatively short follow-up periods for which data were available, 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 final data came from three studies of juvenile offenders who were incarcerated in 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 2,675 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 individuals were followed for an additional 8 to 11 years after this event. The follow-up period for these CYA offenders is much longer than is typically

Table 2.

CHARACTERISTICS OF THE FOUR DATA SETS USED FOR THE PREDICTION ANALYSIS

Data Set	Sample Characteristics	Target Event	Mean Age at Target	Follow-up (Years)	Recidivism by Crime Type[a]	
DOL (N=746)	Rerrals to Jobs Programs in: New York, Miami and Albuquerque	Arrest prior to referral or 1st arrest since age 18 when none prior to referral	17.3	1.8	Property Robbery Violent Total[b]	= 19% = 12% = 7% = 47%
P&P (N=1022)	Convicted felons sentenced to prison or probation in California	Sampled Convictions	26.7	2.6	Property Robbery Violent Total	= 25% = 8% = 5% = 46%
CYA-PRESTON (N=1596)	Serious juvenile offenders in California	First arrest as an adult	18.6	10.8	Property Robbery Violent Total	= 75% = 36% = 45% = 93%
CYA-YCOT (N=1079)	Serious juvenile offenders in California	First arrest as an adult	18.7	7.7	Property Robbery Violent Total	= 69% = 27% = 35% = 92%

[a]The FBI Uniform Crime Report definitions of index offenses are used in classifying crime types. Index property offenses include burglary, larceny-theft and motor vehicle theft. Index violent offenses include murder and nonnegligent manslaughter, forcible rape, and aggravated assault. Robbery -- which is classified as an index violent offense by the FBI, but shares many features with property offenses -- is treated separately throughout this analysis.

[b]The category of Total crimes includes the three enumerated index offenses along with any other follow-up arrests. Thus, Total recidivism always will exceed the sum of the enumerated offense types.

available, and the greater time-at-risk partly accounts for the higher recidivism rates observed in Table 2.

Each of the three major data sets (DOL, P&P and CYA) was itself composed of two or more subsamples of individuals from different jurisdictions or individuals who were exposed to different treatment conditions. In the CYA data, for example, subsamples of offenders were identified primarily by the institution to which they were assigned as juveniles (Preston School of Industry, YCRP experimental program, and Fricot Ranch School). Preliminary, analyses were performed to determine whether these subsamples from the same data source were sufficiently similar (in light of the prediction processes being investigated) that they could be combined into a single analysis sample from that data source.

A generalized least squares (GLS) procedure was used to evaluate whether the relationship between scale scores and follow-up arrests is the same within the various subsamples. In separate analyses by scale (RAND, INSLAW, SFS81, and CGR) and data source (DOL, P&P and CYA), we aggregated individuals who shared a common scale score and subsample membership into subgroups (e.g., all individuals from the Preston subsample whose RAND scale score was 3 formed a subgroup). The percent rearrested was determined within each of these subgroups for five distinct crime types. Weighing each observation by subgroup size, the percent rearrested in a subgroup was then

regressed on scale score, subsample, and crime type[2] as follows:

$$\%Rearrest_{ij} = a + b*ScaleScore_i + \Sigma c_i Subsample_i + \Sigma d_j Crime_j + e \qquad (1)$$

Significant coefficients, c_i, for different subsamples indicate differences in subsample outcomes beyond those reflected in scale scores and crime types. The results of this analysis for the CYA data are reported in Table 3.

Although not reported in Table 3, the percent rearrested 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 intercept, the rearrested percentages were all significantly lower for violent offenses, for robbery and for property offenses. Results for the other control variable, scale score, are reported in Table 3. At the level of subgroups, scale scores are strongly related to %REARREST ($p \leq .001$ level of significance). For example, a higher RAND score for a subgroup is associated with a higher proportion of rearrests among members of that subgroup.

After controlling for crime type and the background differences reflected in scale scores, Table 3 indicates that highly significant differences persisted between outcomes in the Fricot subsample (which is reflected in the constant term) and the Preston subsample. Outcomes in the YCRP subsample are significant at the .05 level or are non-

---

[2]Observations across different crime types for the same subgroup of individuals are treated as independent. This will be violated to the extent that arrests for some crime types lead to incarceration, and thereby reduce the time at risk of arrest for other crime types. The assumption of independence, however, is a reasonable approximation if the order in which different crime types occur during the follow-up is distributed similarly in each subgroup.

Table 3.

SUBSAMPLE DIFFERENCES IN RELATIONSHIP BETWEEN SCALE SCORES AND PERCENT REARRESTED: DIRECTION AND SIGNIFICANCE OF EFFECTS IN THE CYA DATA[a]

SCALE		SCALE SCORE[b]	PRESTON[b]	YCRP[b]	R^2
RAND	(n=80)[c]	+***	+**	+*	.76
INSLAW	(n=75)	+***	+***	+*	.89
SFS81	(n=90)	-***	+***	+*	.84
CGR	(n=75)	+***	+***	+	.83

---

[a]Generalized least squares analysis was used to detect subgroup differences. The functional form of the GLS equations estimated for each Scale and data set is:

$$\%REARREST = b_0 + b_1\ SCALE\ SCORE + b_2\ PRESTON + b_3\ YCRP + b_4\ ROBBERY + b_5\ DRUGS + b_6\ VIOLENT + b_7\ PROPERTY + e$$

Effects of subsample Fricot and of the residual crime type of all other ofenses are reflected in the intercept b_0.

[b]Significance levels of regression coefficients using a two-tailed test are:

*	.05
**	.01
***	.001

[c]Parentheses contain the number of subgroup observations available for each regression (based on 5 or 6 scale scores X3 subsamples X5 crime types). This number is distinct from the number of individuals within a subgroup that is used as the weight variable. The number of observations falls below the maximum possible in the case of the RAND scale (6X3X5=90) because there are no indivuduals for some scale scores in some subsamples.

significant (the CGR scale) indicating that this subsample is more similar to Fricot. Based on these results, we decided to combine the YCRP and Fricot subsamples, but to maintain a separate Preston subsample.[3]

This aggregation combines in the YCOT sample the two 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 the more traditional juvenile training school at Preston were analyzed separately in the PRESTON sample.

The combined YCOT sample contains 1,079 former CYA wards who 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 the time they began juvenile incarceration was 10.9 years. The remaining 84% of the YCOT sample participated in experimental studies about the effectiveness of transactional analysis (O.H. Close Institution) and behavior modification (Karl Holton Institution) between 1969 and 1971. Their median age when they entered these juvenile institutions was 16.6 years. The mean age at the target arrest for the combined YCOT sample was 18.7 years. The recidivism rates following the first adult arrest for the YCOT sample were: 69% subsequently arrested for index property offenses,

[3]Similar analyses were performed to detect subgroup differences for the three cities in the DOL data and between Prisoners and probationers in the P&P convicted sample. No other strong subgroup differences were found, and so the DOL and P&P data are each analyzed as single samples.

27% for robbery, and 35% arrested for an index violent offense (excluding robbery) during a follow-up period that averaged 7.7 years.

The PRESTON sample (described in Table 2) consists of the 1,596 former CYA wards at the Preston School of Industry who had target arrests after their 18th birthday. The median age of individuals when they entered Preston in 1966-67 was 17.6 years, and the mean age at the target arrest occurred about a year later (18.6 years). These offenders subsequently were 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 after the first arrest as an adult also were somewhat higher: 75% subsequently were arrested for property offenses, 36% for robbery, and 45% for violent offenses (excluding robbery). It is important to note that the much longer follow-up periods in both the PRESTON and the YCOT samples are likely to be a factor in their substantially higher recidivism rates when compared to the DOL and P&P samples.

The four analysis data sets share several features that are critical to the research question investigated here. They all involve large samples of offenders. The data sets also are rich in the background variables needed for calculating individuals' scores on the four prediction scales, although no data set perfectly supported all of the scales. Finally, all the samples included sufficient follow-up periods to reasonably observe subsequent offending (operationalized here by arrests).

Table 4.

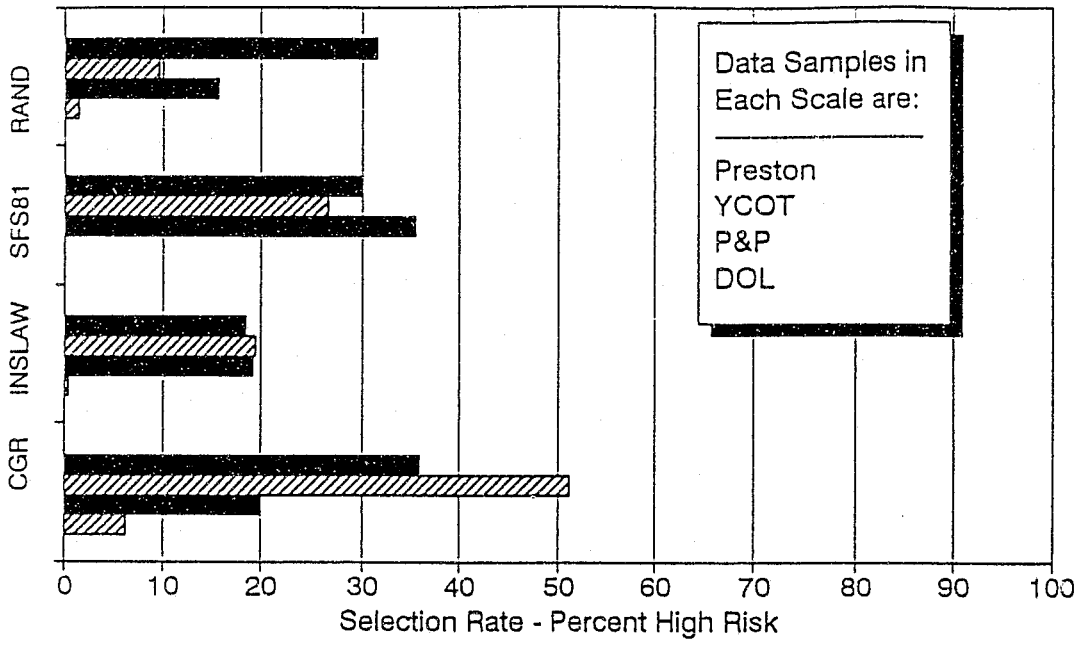
SELECTED BACKGROUND CHARACTERISTICS IN DATA USED IN THE PREDICTION ANALYSES

DATA SET	SAMPLE	PRIOR DRUG/ ALCOHOL USE	AVERAGE NUMBER OF PRIOR ARRESTS
DOL (N=746)	Referrals to Jobs Programs	17%	0.5
P&P (N=1022)	Convicted Felons	39%	3.6
PRESTON (N=1596)	Serious Juvenile Offenders	5%	2.1
YCOT (N=1079)	Serious Juvenile Offenders	31%	2.4

Consistent with our intent when selecting the data, the samples vary considerably in their background attributes. 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 previously held 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 for arrest and for drug or alcohol abuse prior to the target event.

Comparisons across the data sets (Figure 2) illustrate the variation in predictions reflected in the scale "selection rates" (SR's). A selection rate measures the percent of a sample that is classified as high risk using the score cutpoint identified when the scale was constructed. Appendix A reports the cutpoints for each scale. The SFS81 and INSLAW scales made similar predictions across the four data sets while the RAND and GCR scales varied considerably in their predictions across data sets. There also is considerable variability in predictions across scales applied to the same data set. In YCOT, for example, predicted high risk offenders ranged from 10% for the RAND scale to 51% for the GCR scale.

Figure 2 Percent Classified High Risk by Scale and Data Sample



THE FIT BETWEEN THE SCALES AND THE DATA SETS

The purpose of the analysis reported here is to examine the potential that exists for operational criminal justice scales to be transported across offender populations and jurisdictions. The transportability of a scale to other jurisdictions will depend in part on its adroitness at capturing reliable and valid relationships between predictor variables and predicted outcomes. The stronger these basic relationships are, the more broadly a scale will apply to different criminal justice populations. Another equally important factor in transportability is a scale's ability to be applied widely to new data, which in turn depends on how effectively new data can support the scale variables.

We approached the problem of assessing transportability by testing each of the four operational scales on four preexisting data sets that come from the kinds of administrative data likely to be maintained at different CJS agencies. There are two components to our analysis. In this section, we assesses the capacity of new data to support the scales. Then, in the sections that follow, we explore predictive accuracy when the scales are applied to data from different population samples.

The scales we selected for analysis require five general types of information: prior record of offending, substance abuse, age, employment, and education. Table 5 summarizes the scale variables and their application to the data, providing a basis for systematically examining how well each scale and outcome variable is supported by the four data sets. The scales vary in how they operationalize and score (see Appendix A) the

predictor variables. The length of time over which prior record is counted, for example, ranged from 2 years in the RAND scale to 5 years in the CGR scale (item 1a in Table 5), and the types of prior record incidents (items 1b and 1c) ranged from arrests to convictions and times served. This variability can pose significant challenges when scales are applied to new data. In general, the four data sets effectively supported the scales, but approximations of scale variables sometimes were required.

The two years of prior record information in the DOL data is less than is required by the INSLAW, SFS81 and CGR scales. None of the data sets provided information on lengths of prior incarcerations and, for the RAND, INSLAW and SFS81 scales, we estimated prior times served from the lengths of minimum or flat sentences imposed during the background period. Another approximation of prior record substituted prior convictions for prior arrests when applying the INSLAW and CGR scales to P&P data. Finally, the data did not always support scale requirements for the offense type found in prior record information. To operationalize application of the INSLAW and CGR scales to P&P data, we assumed all prior convictions involved property or non-violent felony offenses. All these approximations err on the side of understating the seriousness of prior records, and so will not contribute to false positive errors when predicting recidivists.

In operationalizing substance abuse variables in the RAND, INSLAW and SFS81 scales, drug related arrests and/or convictions were used as the primary indices of drug involvement. The underlying assumption is that individuals intersecting the criminal justice

Table 5.

CAPABILITIES OF DATA SETS TO MEET DATA REQUIREMENTS OF SCALES

VARIABLE	RAND	INSLAW	SFS81	CGR
A. SCALE PREDICTOR VARIABLES				
1. Prior Criminal Record:				
a. Length	2 Years *OK*	5 Years *OK*	3 Years *OK*	5 Years *OK*
		- - - - (except DOL, 2 years only)[a] - - - - -		
b. Time Served	- - - -(Estimated from sent length in all data sets)- - - -			
c. Type of Incidents	Convictions *OK*	Arrests, Probations *OK (P&P, use Convictions)*	Convictions *OK*	Arrests *OK (P&P, use Convictions)*
d. Crime Types	Robbery, Burglary *OK (P&P, Infer from current arrest)*	Violence, Property, Drugs, Other *OK (P&P, assume all property)*	Any Type *OK*	NYS Violent Felony, non-felony *OK (P&P, assume all non-violent)*
e. Juvenile Record	Convictions, Incarcerations *OK (P&P, use convictions)*	Arrests in past 5 years *OK*	Convictions, Incarcerations *OK (P&P, use convictions)*	Arrests in past 5 years *OK*
2. Substance Abuse:				
a. Drugs	Last 2 yrs, As juvenile *OK*	Herion *OK*	Herion, Opiates *OK*	NA[b]
b. Alcohol	NA	Heavy Alcohol *OK (Except DOL and YCOT)*	NA	NA
3. Current Age	NA	At arrest *OK*	At arrest *OK*	NA

[a] Problems in data are noted in parentheses.

[b] NA = Not applicable

Table 5. (continued)

VARIABLE	RAND	INSLAW	SFS81	CGR
4. Length of Employment	Last 2 Years *(Current job, DOL & P&P only)*[a]	NA	NA	Current job *OK (Except YCOT, PRESTON)*
5. Education	NA[b]	NA	NA	Number of Years *OK (Estimate YCOT, PRESTON)*
B. OUTCOME VARIABLES				
6. Type of Incident	Rearrest *OK*	Rearrest *OK*	Committed > 60 Days *(Rearrest only)*	Rearrest or Fail to Appear *(Rearrest only)*
	- - - - - - - - -(P&P new charges filed)- - - - - - - - - -			
7. Length of Follow-up	1-2 Years *OK*	3.5 Years *(Mean 21-31 mos. DOL, P&P)*	2 Years *OK*	< 1 year *OK*
8. Risk Measure	Rate of Reoffending	Time to Rearrest	Parole Revoke/ Reconviction	Rearrest/ Reappear
	- - - - - - - - - -(Rearrest recidivism)- - - - - - - - - -			
9. Target Population	Inmates *OK (Arestees DOL)*	Arrestees *OK (Inmates CYA)*	Inmates *OK (Arrestees DOL)*	Arrestees *OK (Inmates CYA)*
	- - - - - - - - - - - -(Convictees P&P)- - - - - - - - - - - -			

[a] Problems in data are noted in parentheses.

[b] NA = Not applicable

system for offenses involving drugs are highly likely to be drug users, an assumption that is confirmed by the 79% drug positives found in urine tests of arrestees charged with drug sales or possession in the Drug Use Forecasting (DUF) program (NIJ, 1992). Errors of omission are almost certainly larger in failing to detect histories of drug use among individuals who did not have prior drug arrests. These errors are ameliorated by self report data in the PRESTON and DOL data, and by reports of drug involvement at the current arrest in the P&P data. Only the YCOT data are limited entirely to drug arrests.

Information about alcohol use required by the INSLAW scale is totally absent in the YCOT and DOL data sets. Clinical data concerning alcohol use as a juvenile was used to support this scale variable in the PRESTON data set and arrest information about alcohol use in the instant offense was used in the P&P data set. As with prior record variables, the coding assumptions about use of drugs and alcohol almost certainly operate to systematically understate substance abuse in the RAND, INSLAW, and SFS81 scales, and will minimize false positive prediction errors.

Current employment data were required to support the RAND and CGR scales. Such data were unavailable in the PRESTON and YCOT data sets. Because there was no way to infer individual employment status, employment is dropped from the scale scores. For the RAND scale this adjustment effectively eliminates time employed as a risk factor for recidivism. In the CGR scale, dropping the employment variable penalizes those offenders who would have benefited from positive information about their employment

status, and is the only instance where the operationalization of a scale has the potential effect of inflating false positive errors among predicted recidivists.

Information in the DOL and P&P data sets was sufficient to support the employment variable in the CGR scale. The RAND scale variable which required two years of prior employment data was approximated by the most recent employment experience in the DOL data (last year's employment) and in the P&P data set (currently employed). These operationalizations probably overstate prior employment, and again will understate scale predictions of recidivism.

Only the CGR scale invoked an education variable. The last grade reported while the offender was incarcerated as a juvenile, and the highest achievement test level, also while incarcerated as a juvenile, were projected forward to estimate total years of schooling in the Preston and YCOT data, respectively. Thus , an individual completing only 6 out of 9 possible years of schooling when incarcerated as a juvenile was assumed to continue through school at the same rate, completing a total of 8 years [6 + .67(3)] by the time of their first adult arrest. This approximation probably overestimates the years of school completed, and thus will understate the risk of subsequently offending.

The data sets were least effective in supporting the CGR scale. This scale contains only four variables, and two of these (employment and education) were either unavailable or required considerable estimation in the PRESTON and YCOT data sets. Among the

data sets, P&P is the weakest at supporting the scales. Arrest information had to be inferred from conviction data, the types of prior offenses involved were not identified, and the only employment information was current employment.

A variety of assumptions were required to apply the scales to the data employed in this research. These assumptions are arguably reasonable, and indeed it is likely that operational data in jurisdictions wanting to employ these scales would not be significantly better. In almost all approximations we erred on the side of caution with respect ot recidivism predictions, reflecting a greater concern for minimizing false positive errors.[4] On balance, we believe the data are capable of supporting a meaningful analysis of the potential transportability of the prediction scales.

The outcome variables for the scales are shown at the bottom of Table 5. P&P are the only data without explicit information on subsequent arrests; follow-up arrests were inferred from the presence of criminal charges filed against an individual. This will result in an undercount of arrest events when charges are not filed by the prosecutor following an arrest, but the magnitude of these errors is not likely to be large. It also is interesting to note the differences between the length of the follow-up periods specified in the scales (about 2 to 5 years) and those available in the data sets (about 2 to 18 years). The target populations used to develop the scales and those used in this analysis also differ considerably as was intended in this study of the transportability of scales across

[4]The implications of these assumptions for predictive accuracy are discussed in note 13.

populations and decision contexts.

THE ANALYSIS PROCESS

Two steps typically are 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 and release outcomes of the defendants or convicted offenders in a sample, an empirical scale is developed, and score cut-points are chosen to classify sample members into subgroups. The subgroups are identified by differing levels of risk posed by individuals on the criterion 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 the ineffective measurements of scale variables (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," also obtained from the target population. An important advance in validating prediction instruments was measuring scale accuracy in terms of the magnitude by which prediction accuracy deteriorates from the construction to the validation sample (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 population.

Another approach for assessing the utility of prediction instruments is to compare the performance of a prediction scale relative to the performance of some random prediction process. The frequencies expected under the random process are defined like those in the Chi Square statistic (Meehl and Rosen, 1955). In a 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[5] (SR), while the proportion of actual criterion outcomes reflects the base rate (BR) within the sample. The expected frequencies are computed from products of the selection rate and base rate, and predictive efficiency is assessed by comparing the number of correct predictions made by a prediction scale to the number expected from a chance process (Wiggins, 1973).

A final set of indices traditionally used to assess the performance of prediction scales focuses on the two types of prediction errors: type I error, and type II error[6]. Using

[5]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 those classified as high-risk offenders 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 data sets, the selection rate (like the base rate) is exogenously determined by sample characteristics.

[6]In the prediction context, Type I error occurs when cases actually possessing the criterion attribute are missed by the prediction instrument. These errors also are referred to as "False Negative errors." The consequences of false negative errors in the criminal justice system are highly visible. They also 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 on release from prison. Type II error occurs when the criterion attribute is incorrectly predicted to be present. These "False Positive errors" are insidious in the criminal justice process because their low visibility combines with severe consequences for offenders. For example, the prediction that an inmate poses too high a risk to warrant parole can lead to continued, but unnecessary, incarceration when made about an inmate who will never again commit a crime, and these prediction errors are impossible to detect since the inmate is not released.

recidivism as an illustrative criterion outcome, type I error would be assessed using the false negative rate (FNR), or the proportion of recidivists who are missed by the prediction instrument, while type II error would be assessed by the false positive rate (FPR) of non-recidivists found among those who are predicted to recidivate. Since the two types of prediction errors move in opposite directions, it has proved difficult to develop prediction instruments that simultaneously reduce both types of error.

When using any error measure, it would be desirable to contrast alternative predictive scales in terms of their relative error characteristics. Such comparisons have been limited, however, by the fact that all of the error measures are highly dependent on the base rate and the selection rate that characterize the sample and the prediction scale, respectively. (Gotfredson and Gotfredson, 1980; Blumstein, et al, 1986; Loeber and Dishion, 1983) Thus, it is difficult to make comparative assessments either of alternative scales applied to the same populations or of the same scale applied to different populations.

While the BR remains fixed within a single population, the SR probably will differ across alternative scales and for different cutpoints of the same scale. It is equally difficult to make comparisons of a single scale's performance when it is applied to several different populations of offenders because of changes in both the BR and SR. Until recently, the ability to compare scales in terms of their errors has been constrained by the lack of procedures for simultaneously accommodating differences in the SR's of prediction scales

and in population BR's.

The inverse relationship between false positive and false negative error rates -- as one increases, the other decreases -- typically results in observed FPR's and FNR's that differ widely in magnitude. As described by Cohen and Zimmerman (1990), however, such differences in the absolute magnitudes of error rates do not necessarily reflect meaningful differences in the overall accuracy of a scale. This apparent paradox is the product of a set of constraints on the range of possible values of FNR and FPR errors that derive from the population BR and scale SR. While the absolute magnitude of the FNR may be low compared to the associated FPR, the constraints on possible values of the FNR and FPR also differ so that each error rate represents the same relative improvement in accuracy within the ranges of possible values for FPR and FNR error rates.

This relationship underlies the benefits of the "Relative Improvement Over Chance" (RIOC) measure as an index capable of providing information about the relative performance of prediction instruments (Loeber and Dishion, 1983). The RIOC statistic calibrates the observed improvement over random accuracy relative to the constraint on maximum accuracy. The functional form of this measure is:

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

The RIOC statistic ranges between -1.0 and 1.0, and can be interpreted as the

proportional (or percentage) improvement (or loss) in accuracy relative to the maximum possible accuracy. Despite its apparent utility, this measure has not been widely used in the criminological literature.

Bounds On Traditional Error Rates

Traditional measures of prediction accuracy are well established. They include the total prediction accuracy rate (TPAR), and three indicators of the frequency of errors: the total error rate among all predictions (TER = 1-TPAR), the false positive rate (FPR) among predicted failures, and the false negative rate (FNR) among predicted successes.

We have shown elsewhere (Cohen and Zimmerman, 1990) that the range within which false negative and false positive error can vary is constrained by the BR and SR. For any combination of BR and SR, the range within which total error must fall can be fully specified. Consider the example in Figure 3, in which BR = .7 and SR = .5.

Accurate predictions include true positives (TP) and true negatives (TN) in the diagonal cells of Figure 3a, while prediction errors are the false positives (FP) and false negatives (FN) in the off-diagonal cells. As shown in Figure 3b, the maximum proportion of cases in each cell is constrained by the BR and SR. If, for example, SR is .5, the share of cases that are FN can not exceed the smaller marginal value of (1 - SR) = .5 < BR, while the share of FP cases is limited by the smaller marginal value of (1 - BR) = .3 < SR. In

Figure 3. Constraints on Prediction Errors - An Example for BR=.7, SR=.5

a. Distribution of Correct Predictions and Prediction Errors (Fraction of Total N)

Actual \ Predicted	+	−	
+	TP	FN	BR = .7
−	FP	TN	1-BR = .3
	SR = .5	1-SR = .5	

b. Maximum Cell Values (Fraction of Total N)

Actual \ Predicted	+	−	
+	.5	.5	BR = .7
−	.3	.3	1-BR = .3
	SR = .5	1-SR = .5	

VARIABLE KEY
(All variables are fractions of total N)

TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
BR	Base Rate
SR	Selection Rate
TER	Total Error Rate

Max TER = Max FN + Max FP = .8
Min TER = Min FN + Min FP = 1 - Max TN - Max TP = .2

this case, the total error rate cannot exceed 80% (MaxTER = MaxFN + MaxFP = .8) Likewise, the minimum error rate is constrained by the maximum possible correct predictions. In the Figure 3b example, correct predictions can not exceed 80% and so the minimum total error rate is 20%.

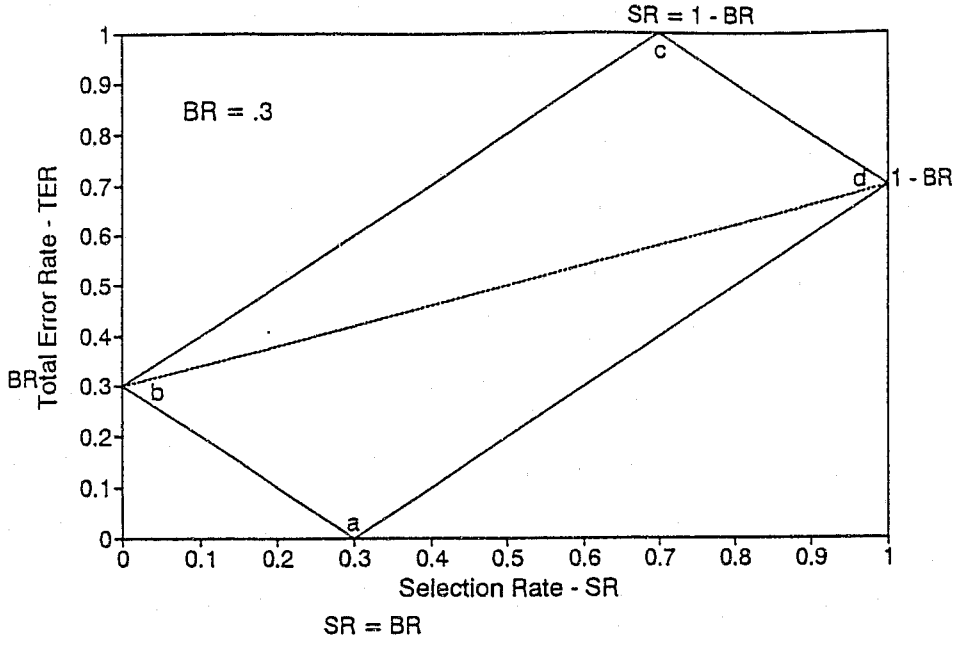
Figure 4 displays the full region of minimum and maximum bounds on total prediction errors over all possible values of SR for three BR values. When no cases are classified as possessing the criterion attribute (SR = 0), or all cases are so classified (SR = 1.0), the total error rate (TER) is determined completely by the base rate (BR). TER achieves the maximum possible value of 1.0 only when SR = (1 - BR), and achieves the minimum possible value of zero only when SR = BR.

The polygons formed by the corner points in Figure 4 completely bound the region of possible values for TER. Maximum possible error (MaxTER) is delineated by the line segments (b,c) and (c,d) and the minimum possible error (MinTER) is delineated by (b,a) and (a,d). The dotted line segment (b,d) specifies the value of random error (RandTER) at each SR [RandTER = BR(1 - SR) + SR(1-BR)]. At all values of SR other than (1 - BR) and BR, total errors are constrained to fall within a range that is smaller than 0 to 1, and sometimes that range is considerably smaller. The range of possible TER is largest when the selection rate (SR) falls between BR and (1 - BR).

The total error rate (TER) is composed of FP errors and FN errors. Figure 5

Figure 4. Region of Possible Prediction Errors - - Variation With Values of Base Rate BR and Selection Rate SR

a. Base Rate Smaller Than .5



b. Base Rate = .5

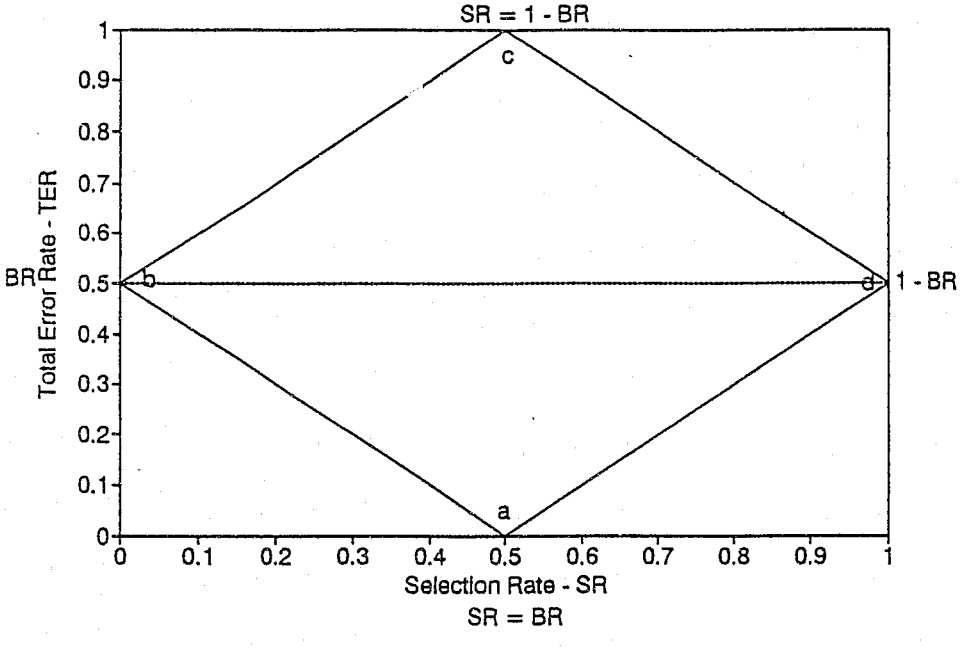
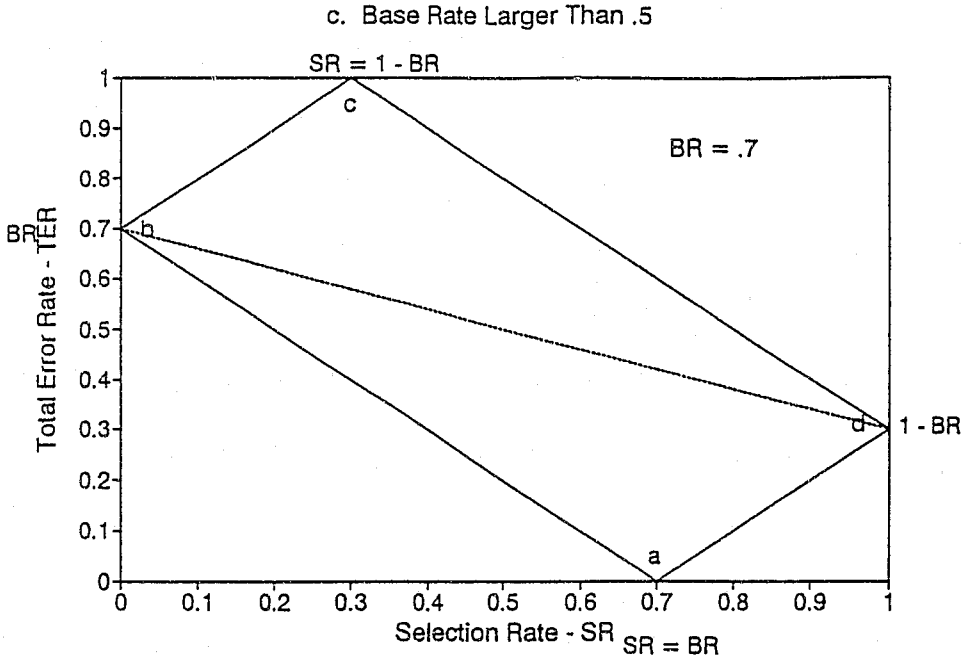


Figure 4. Region of Possible Prediction Errors - - Variation With Values of Base Rate BR and Selection Rate SR



partitions the false positive and false negative components in the maximum and minimum total error rate. Up to the point where SR = BR, false negative errors contribute more than false positive errors in both the minimum and maximum TER. At values of SR > BR, false positive errors exceed false negative errors in the total. We note further that increases in the bounds of TER are due entirely to increases in the FP component of TER. Declines in the bounds on TER correspondingly arise from decreases in the FN component. These patterns occur regardless of the value of BR.

In an operational prediction context, the relevant concern is the degree to which error can be minimized (not how large error can be). It is therefore important to focus on the minimum possible error for a prediction instrument in a particular sample. The value of MinTER can be obtained directly from:

$$MinTER^* = SR - BR \qquad \text{(Equation 2)}$$

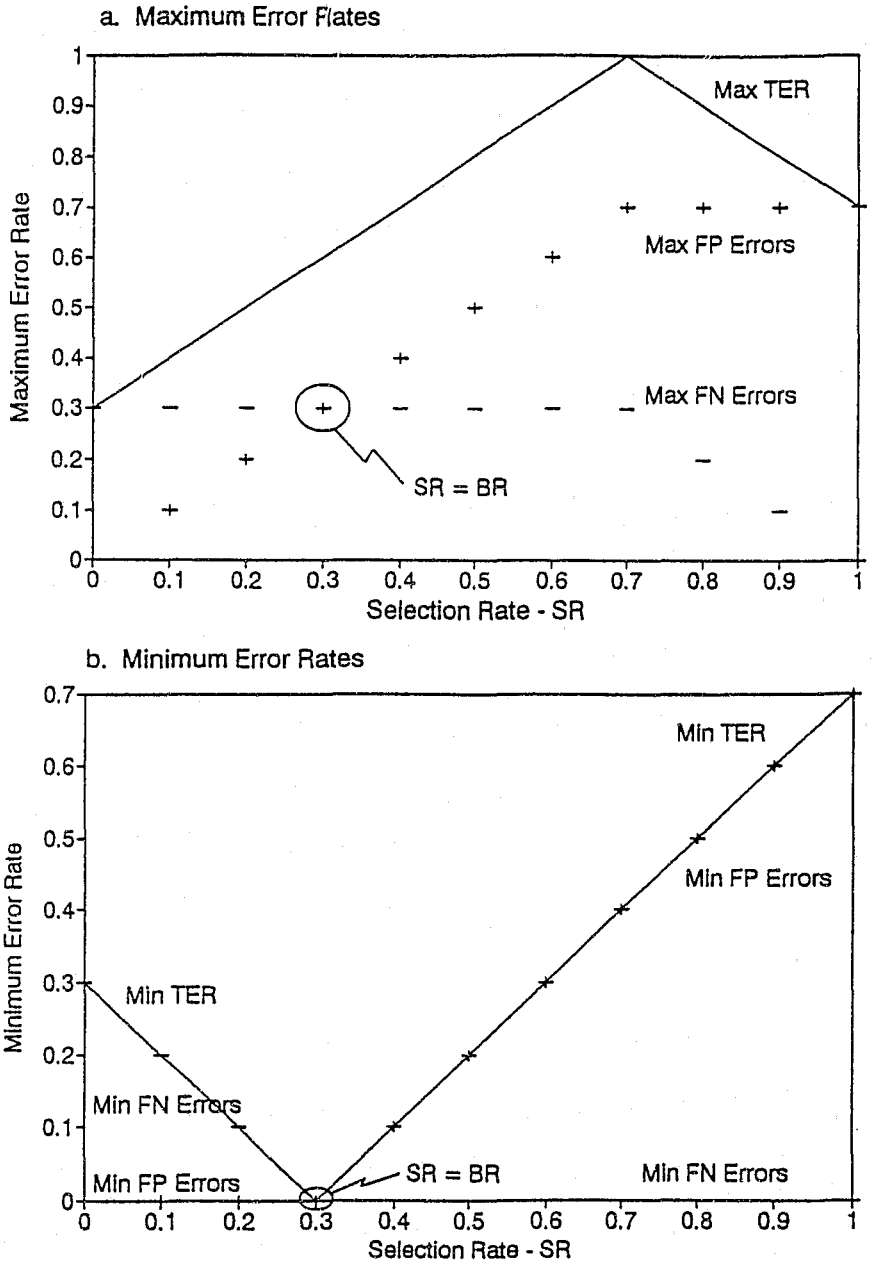
With $MinTER = |MinTER^*|$

If $MinTER^*$ is < 0, MinTER consists entirely of FN errors.

If $MinTER^*$ is > 0, MinTER consists entirely of FP errors.

The magnitude of minimum error is the absolute value of the difference, (SR - BR). The sign of this difference indicates whether the minimum value of TER is constrained by erroneous false negative predictions or by false positive errors. When SR - BR = 0 there are no structural constraints on minimizing TER, and only then is it possible to correctly classify every case. This restriction of MinTER to zero only when SR = BR underlies the desirability of maximizing congruence between sample BR's and prediction scale SR's.

Figure 5. Relative Contributions of False Posiitive (FP) and False Negative (FN) Errors to Total Error Rate: Illustration for BR = .3



NOTE: Minimum and maximum error rates are obtained from the marginal constraints in a 2-by-2 classification table as follows:

Max FP = Min(1-BR,SR) Max TP = Min(BR,SR) => Min FP = SR - Max TP

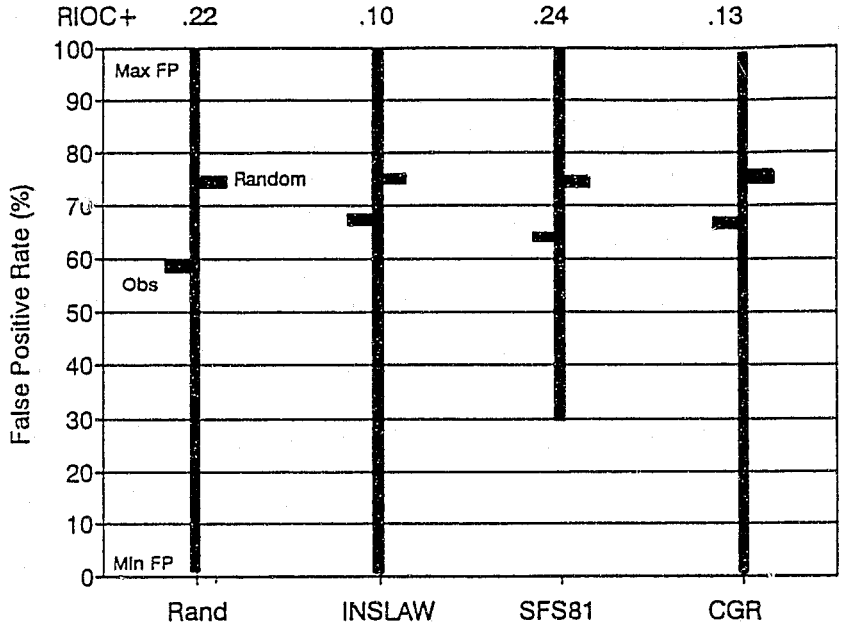
Max FN = Min(BR,1-SR) Max TN = Min(1-BR,1-SR) => Min FN = (1-SR) - Max TN

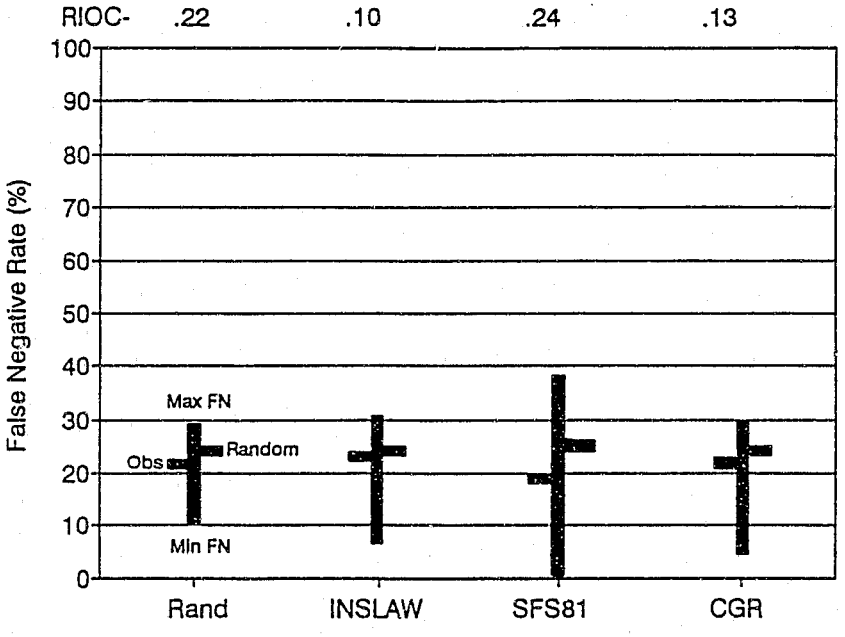
The constraints on possible errors that arise from sample-specific BR's and SR's mean that observed FPR's, FNR's and TER's may differ widely in magnitude without reflecting substantive differences in accuracy over different scales or across data sets. Thus, when choosing among prediction scales, or comparing the performance of one scale across data sets, it would be desirable to remove the dependence of accuracy on values of the Base Rate or the Selection Rate.

As suggested earlier, one such standardized measure is the Relative Improvement Over Chance (RIOC) statistic (Loeber and Dishion, 1983). This statistic (Equation 1) contrasts the improvement in accuracy achieved beyond "random accuracy", relative to the full potential for such improvement. In particular, the observed magnitude of improvement above random accuracy is calibrated by the constrained range between random accuracy and maximum accuracy. In addition, within any prediction exercise involving one scale applied to one data set, separate RIOC values calculated for FNR, FPR and TER are identical (Cohen and Zimmerman, 1990). In other words, the proportional improvement over random error is identical for false positives, false negatives and total errors, despite differences in the ranges of potential improvement for each type of error.

Figure 6 graphically depicts these proportional relationships between observed and random accuracy within the ranges of possible FPR and FNR. The RIOC statistic, which ranges between -1.0 and +1.0, is interpreted similarly to other measures of reductions in

Figure 6. Accuracy in Predicting Re-arrest for Any Index Property Offense Using Four Scales Applied to the P&P Data of Prisoners and Probationers from California Counties (n = 1,022)





error. Positive values indicate the proportional improvement that is achieved in the range from random accuracy to maximum possible accuracy (ie., toward minimum error). Negative RIOC values reflect decrements in prediction accuracy away from random accuracy and in the direction of the maximum error possible in the process.

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 the total (Farrington and Tarling, 1985). In fact, as described above, this is not a problem. Both FP and FN errors perform identically once they are calibrated relative to their respective ranges of possible values. Relative to random error, and within their constrained distributions of minimum and maximum possible error, observed accuracy for FN and FP errors is identical.

The absolute magnitude of the FNR may increase when a scale cutpoint is decreased to identify fewer individuals who are likely to recidivate. At the same time, however, the associated constraints on possible values of the FNR also may shift in response to change in the SR. The resulting FN and FP errors that are observed will represent the same relative improvement in accuracy within the respective ranges of FPR and FNR error rates. This important result is illustrated graphically in Figure 6 where the range of possible FN errors is much smaller than the range for FP errors, but the relative improvement in accuracy is identical within each of their respective ranges.

For example, false positive errors for the RAND scale can fall anywhere in the full range from 0% to 100%, while false negative errors are constrained within a much narrower range from 11.0% to 29.4%. Nevertheless, relative to random accuracy (indicated by the right sidetick mark in Figure 6), the actual relative improvement in accuracy (RIOC) is .22 for both false positive and false negative errors. Furthermore, RIOC values estimated separately for FP and FN errors are identical to the overall RIOC for all errors combined. Given this identity among these three measures, here we employ the overall RIOC to compare the accuracy of alternative scales.

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. The sample and scale dependence of these measures on base rates and selection rates, respectively, undermines attempts to make comparisons among scales. When the same scale is applied to different data samples, each with a different base rate, it is difficult to compare scale performance across the samples using FP or FN rates. Similar problems occur when assessing predictions from multiple scales applied to a single data set in that 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.

Accuracy Measures Used In This Study

There are three salient considerations when assessing the effectiveness of

alternative prediction scales or of a single scale applied to different populations. First, it is important to determine the limits on potential error. Given the invariant relationships we have identified between FP and FN errors, we focus on the total error rate (TER) and specifically on the minimum bounds for total error (MinTER). Second, information is needed about how well the prediction scales are performing in absolute terms, and for this we use the Total Percent of Correct predictions (%CORR). Finally, prediction scales are assessed relative to their optimum performance potential using the RIOC statistic.

With respect to TER, we impose an initial screening constraint on the minimum TER that is acceptable and exclude from analysis cases where the MinTER >= .33. That is, we decided a priori to exclude from consideration scales applied to samples where the combination of BR and SR structurally force the magnitude of error to exceed one-third of all cases. This relatively high threshold on minimum error would be overly generous in normal criminal justice decision making applications, but it seems appropriate for examining the transportability of scales across a substantial range of base rates. In this analysis we also chose to make no distinction based on the relative contributions of false negative and false positive errors to the minimum TER. While the composition of the minimum error would be an important operational concern, we believe it is less relevant in the context of this type of comparative analysis.

For those analyses that pass the initial screen, our final assessment is guided by jointly examining the absolute level of prediction accuracy when a scale is applied to a

data set (%CORR), and the scale's performance relative to the possible improvement from the baseline of random error (RIOC). These two accuracy measures are combined by conditioning the proportional improvement in accuracy for positive RIOC values by the proportion of correct predictions in each analysis.[7] This final measure, denoted by CORIOC, is the square root of the product,

$$CORIOC = sqrt[RIOC * (\%CORRECT/100)] \quad \text{(Equation 3)}$$

$$\text{for } RIOC >= 0;$$

$$CORIOC = 0 \text{ if } RIOC < 0.[8]$$

Taking the square root normalizes the distribution of the highly skewed product of two proportions within a range of possible values between 0.0 to 1.0.[9] The final CORIOC values are compared across the scales applied to each unique data set/crime type combination to identify the "best performing" scale in that data set/crime type.

RESULTS

All four scales (RAND, INSLAW, SFS81, CGR) are used to examine scale

---

[7]The simple product measure proposed here also can be interpreted as conditioning the absolute percent correct by relative performance on the RIOC statistic.

[8]The composite measure in equation 3 is offered here because of its intuitive appeal as a weighted RIOC (or, alternatively as a weighted proportion correct). The use of a product measure for CORIOC is not meant to imply independence between the two component measures. Indeed, independence is not likely since both %CORRECT and RIOC reflect predictive accuracy in the same data. Alternative statistics that combine %CORRECT and RIOC might offer more desirable statistical properties, especially if a known distribution of the statistic were to provide a basis for assessing the statistical significance of observed values. This issue is left for future analysis.

[9]Negative RIOC values were reset to zero since negative CORIOC values are not substantively meaningful.

performance in predicting rearrest for five crime types (Violent[10], Robbery, Property[11], Drugs, Total) in each of the four data sets (PRESTON, YCOT, P&P, DOL). The result was 80 separate analyses. The results of these analyses are reported in Appendix B.

The BR was unique within each data set/crime type and each scale had a different SR. The BR's ranged from 93.7% recidivism (for Total crime in Preston) to 5.3% recidivism (for Violent crime in the P&P sample). Adult recidivism in the two data sets containing serious juvenile offenders (Preston and YCOT) averaged over 50% across the five crime types. The two data sets of adult offenders (P&P and DOL) averaged about 18% recidivism following their first adult arrest. The differences in BR's undoubtedly reflect the more lengthy follow-up periods for the California youth samples, and also may reflect differences in the age of onset of delinquent or criminal behavior.[12] The juvenile offender samples were followed for an average of about 7 years after their first arrests as adults, compared with an approximately 2 year average follow-up for the P&P and DOL samples. The longer follow-up periods for the juvenile offender samples increased the probability that recidivism would be observed.

---

[10]Using the FBI Uniform Crime Report definitions for index offenses, index violent offenses are: murder and nonnegligent manslaughter, forcible rape, aggravated assault, and robbery (which was treated separately in this analysis and is not included among violent offenses).

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

[12]While the juvenile offender samples were all known to start offending as juveniles, juvenile records were not available for the other two samples.

The SR's ranged from 51.1% predicted recidivism (for the CGR scale applied to the YCOT sample) to 0.1% predicted recidivism (for the SFS81 scale on the DOL sample). The patterns of mean SR's were similar to the BR pattern. Predicted recidivism (across all scales) averaged around 25% for the two California juvenile offender samples and for the California adult offender sample (Preston, YCOT and P&P), but averaged only 2.1% for the individuals in the DOL sample (where inclusion was based on employment program referrals).

The First Stage Analysis.

The initial screening criterion, that the value of MinTER be <= .33, excluded scale/data set/crime type combinations in which more than one-third of the cases were structurally required to be prediction errors. This occurred in 19 (24%) of the original analyses and these analyses were excluded. The 61 analyses that met or exceeded the screening criterion were retained for further substantive examination. An inspection of Appendix B, indicates that the most striking features of the excluded analyses are their high BRs, relative to other analyses in a data set, and the large negative MinTER* values. For these analyses, therefore, substantial numbers of structural FN errors are created by

the relatively low scale SRs and the high associated BRs.[13]

The MinTER criterion excludes 8 analyses from the Preston data set, 7 from YCOT, and 4 from DOL. No P&P analyses are excluded, indicating that, in this data set, the proportion of scale predictions most closely match the proportion of those rearrested. All of the analyses for the aggregate category of "Total" crimes are excluded by the first stage screening criterion, except those associated with the P&P data set. In addition, all the analyses of recidivism in 4 Property crimes are excluded from the Preston data set and another 3 in YCOT.

The MinTER* values are negative in 48 of the 61 retained analyses (77%). Positive values of MinTER* indicate that false positive errors predominated in TER. Positive MinTER* values occur only in the P&P analyses, where 13 of the 20 analyses have SR's that exceed the sample BR's. The characteristics of offenders used in applying the prediction scales to the P&P data, therefore, not only produce a better match between SR's and BR's, but also tend to produce a substantial proportion of analyses in which minimum error limits are structurally responsive to FP errors.

---

[13]As described in the discussion of data "fit", the coding assumptions are conservative in all but one instance resulting in structural errors that depress SR's. In cases where SR < BR (77.5% of the 80 analyses), the structurally reduced SRs increase the threshold for minimum error (see Figure 5). The additional structural errors result exclusively from increased FN errors. In the remaining 22.5% of the analysis (SR > BR), the structurally reduced SRs lower minimum possible error by decreasing FP errors. For most cases, then, our coding decisions create a tough test for the scales by increasing the minimum possible error they can achieve, and the results of our analyses should be interpreted as conservative estimates of scale performance. Finally, it is important to note that there are no indications that scale and data pairs with important coding assumptions have systematically different patterns of outcomes.

The final column of Table 6 illustrates that in the analyses that were excluded the scales did not predict well. The mean percentages of correct predictions in the excluded analyses are substantially lower than in the retained analyses. Thus, the MinTER screening procedure performs as intended. Although the SR's do not differ appreciably between the retained and excluded analyses, the BR's in the excluded analyses are much higher. This again reflects the poor performance of the scales when substantial proportions of the samples are rearrested.

Second Stage Analysis.

The second stage in investigating the potential for transportability of criminal justice prediction scales involves examining the performance of the RAND, INSLAW, SFS81, and CGR scales using the analyses that are retained after the stage one screening criterion. Scale performance on populations that differ from the original construction and validation samples is assessed in terms of predicted recidivism following the target adult arrest. This assessment simultaneously considers each scale's actual predictive accuracy (operationalized as the total percentage of correct predictions) and its accuracy relative to potential performance (the RIOC statistic). As described above, we use a single index, the CORIOC, in which the relative performance (RIOC) is conditioned by the absolute proportion correct. The CORIOC is interpreted as a joint indicator of the absolute and relative predictive performance of a scale.

The mean CORIOC values for all stage two analyses, reported in Table 7(A), vary

Table 6.

MEAN PERCENTS CORRECT, BASE RATES AND SELECTION RATES
OF THE PREDICTION ANALYSES DATA SETS
(Reported in Percentages)

DATA SET		ALL ANALYSES %CORR	ALL ANALYSES BR%	ALL ANALYSES SR%	RETAINED ANALYSES %CORR	RETAINED ANALYSES BR%	RETAINED ANALYSES SR%	EXCLUDED ANALYSES %CORR	EXCLUDED ANALYSES BR%	EXCLUDED ANALYSES SR%
PRESTON	%CORR	50.0%			57.9%			38.2%		
	BR%		59.2%			42.8%			83.8%	
	SR%			28.9%			28.9%			28.9%
		(k=20)			(K=12)			(K=8)		
YCOT	%CORR	51.5%			59.9%			35.9%		
	BR%		51.7%			35.3%			82.8%	
	SR%			26.7%			28.6%			23.2%
		(K=20)			(K=13)			(K=7)		
P&P	%CORR	70.6%			70.6%			-----		
	BR%		18.6%			18.6%			-----	
	SR%			22.5%			22.5%			----
		(k=20)			(k=20)			(k=0)		
DOL	%CORR	81.0%			87.8%			53.8%		
	BR%		18.0%			10.8%			46.9%	
	SR%			2.1%			2.1%			2.1%
		(k=20)			(k=16)			(k=4)		

---

k = the number of Crime type x Scale analyses involved for each Data set

between .214 (SFS81) and .333 (INSLAW). These results indicate that the best performing scale across the range of data sets and crime types was INSLAW. The low CORIOC values for the SFS81 scale reflects the substantial number of negative RIOC's, which also are reflected in the negative mean RIOC values in Table 7(A). (See also Appendix B).

When only the analyses containing significant RIOC's are used in calculating the CORIOC means, the scale means are generally larger and variation across scales is reduced from a low of .299 (CGR) to a high of .357 (SFS81). Inspection of Appendix B shows that excluded analyses with insignificant RIOC's tended to have high total percentages of correct predictions and largely negative RIOC values, particularly in the DOL data. Thus, the mean %CORRECT is lower for each scale in Table 7(B) than Table 7(A) and the corresponding mean RIOC values are higher. The changes in the mean %CORRECT are proportionally smaller than those in mean RIOC's and thus, in this analysis of means, CORIOC varies more closely with the measure of relative improvement in accuracy.

The specific data set/crime type/scale analyses with significant RIOC's that are retained are delineated in Table 7(C). There are no large differences across the scales in the types of crimes involved in each data set. Differences are evident across the data sets for each scale. For example, only one DOL analysis (Property with the CGR Scale) remains in the Stage 2 analysis . Rearrest recidivism is difficult to predict with any

Table 7.

SUMMARY INFORMATION CONCERNING PREDICTION ACCURACY FOR STAGE TWO ANALYSES, BY SCALE

(k = the number of Crime type x Data set analyses involved)

A. All Stage Two Analyses

SCALE	MEAN %CORR	MEAN RIOC	MEAN CORIOC
RAND (k=15)	73.3%	.061	.277
INSLAW (k=15)	73.9%	.133	.333
SFS81 (k=15)	68.6%	-.149	.214
CGR (k=16)	67.7%	.098	.236

B. Significant RIOC Values Only: Stage Two Analyses

SCALE	MEAN %CORR	MEAN RIOC	MEAN CORIOC
RAND (k=10)	68.3%	.177	.331
INSLAW (k=9)	65.1%	.161	.315
SFS81 (k=9)	62.6%	.194	.357
CGR (k=10)	61.8%	.146	.299

Table 7 (Continued)

SUMMARY INFORMATION CONCERNING PREDICTION ACCURACY
FOR STAGE TWO ANALYSES, BY SCALE

C. Crime types of Significant Stage Two Analyses

SCALE	PRESTON	YCOT	P&P	DOL
RAND (k=10)	Violent Robbery Drugs	Violent Robbery	Violent Robbery Property Drugs Total	
INSLAW (k=9)	Violent Robbery Drugs	Violent Robbery	 Robbery Property Drugs Total	
SFS81 (k=9)	Violent Robbery	Violent Robbery	Violent Robbery Property Drugs Total	
CGR (k=10)	Violent Robbery Drugs	Violent Robbery Property	 Property Drugs Total	 Property

significant accuracy among the young adult arrestees in the DOL sample, largely because the number of recidivists is very small (the base rate is low). Only the CGR Scale--which was originally developed on arrestee samples--ever exceeds random accuracy on the DOL arrestee data. The P&P data set, on the other hand, contained samples of probationers and prisoners, had intermediate BR's, and consistently produced the largest number of Stage 2 analyses. With the Preston and YCOT data sets, which had the highest BR's, two or three crime types remained for each scale.

Finally, Table 7(C) suggests that there is no systematic relationship between the prediction scales and their performance on specific crime types. Instead, all scales perform well on the same crime types. While such a relationship theoretically might not have been expected, since the scales were constructed for different purposes and on varying criminal justice populations, the simple empirical correspondence between the BR and SR in each analysis seems to be the more salient determinant of prediction success.

An alternative approach to assessing the predictive ability of the scales is to compare their performance within each unique data set/crime type. Table 8 provides comparative information for those analyses with statistically significant RIOC statistics.

As can be seen in Table 8, the INSLAW scale had the highest CORIOC value in four of the ten available comparisons. These analyses involve a variety of crime types in the

Table 8.
PREDICTION ACCURACY DATA FOR SCALES WITHIN EACH DATA SET/CRIME TYPE FOR WHICH RIOC IS SIGNIFICANT (Ordered by Base Rate)

ANALYSIS[a]	N	BR	MinTER*	%CORR	RIOC[b]	CORIOC	FPR	FNR
YCOT P RAND / R INSLAW / O SFS81 / P CGR	830	69.0	-17.9	54.1	.116*	.251@	27.4	65.3
PRES D RAND	1596	47.2	-15.7	55.7	.139***	.278	45.4	43.9
R INSLAW	1596	47.2	-28.8	55.2	.175***	.311@	43.5	45.2
E U SFS81								
S G CGR	1056	50.8	-14.9	55.1	.153***	.290	41.7	46.5
P T RAND	1022	45.6	-30.0	60.4	.434***	.512@	30.8	41.3
& O INSLAW	1022	45.6	-26.4	58.1	.295***	.414	40.3	42.3
P T SFS81	1022	45.6	-10.2	63.2	.309***	.442	37.6	36.4
A CGR	979	44.7	-24.8	56.8	.165**	.306	46.2	42.5
L								
P V RAND	1596	45.0	-13.5	56.5	.135***	.276	47.6	41.6
R I INSLAW	1596	45.0	-26.6	57.4	.209***	.346@	43.5	42.4
E O SFS81	1596	45.0	-15.2	56.1	.125***	.265	48.1	42.1
S L CGR	1056	43.9	-8.0	59.6	.195***	.341	45.1	37.8
P R RAND	1596	35.8	-4.3	62.3	.174***	.329	53.0	30.7
R O INSLAW	1596	35.8	-17.4	63.9	.210***	.366@	50.7	32.8
E B SFS81	1596	35.8	-6.0	60.6	.126***	.276	56.1	32.4
S CGR	1056	35.0	+0.9	60.3	.136***	.286	56.5	30.3
YCOT V RAND	1079	35.1	-25.5	63.6	.126*	.283	56.7	34.3
C I INSLAW	1079	35.1	-15.6	62.9	.149***	.306@	55.2	32.8
O O SFS81	1079	35.1	-8.6	59.9	.084*	.224	59.4	33.2
T L CGR	830	34.3	+16.8	53.1	.103*	.234	62.3	30.8
Y R RAND	1079	27.2	-17.6	69.8	.101*	.266	65.4	26.5
C O INSLAW	1079	27.2	-7.7	66.5	.090**	.245	66.2	25.7
O B SFS81	1079	27.2	-0.7	64.5	.096***	.249	65.7	24.7
T CGR	830	26.9	+24.2	53.8	.166**	.299@	68.9	22.4
P P RAND	1022	24.9	-9.3	72.4	.222***	.401@	58.5	21.8
& R INSLAW	1022	24.9	-5.7	68.2	.097***	.257	67.9	23.1
P O SFS81	1022	24.9	+10.5	65.1	.244***	.399	64.1	18.8
P CGR	979	23.9	-4.0	69.7	.131***	.302	66.2	21.4
DOL P RAND / R INSLAW / O SFS81 / P CGR	746	18.9	-12.6	78.8	.160**	.355@	68.1	18.0

Table 8 (Continued)
PREDICTION ACCURACY DATA FOR SCALES WITHIN EACH DATA SET/CRIME TYPE
FOR WHICH RIOC IS SIGNIFICANT (Ordered by Base Rate)

ANALYSIS[a]	N	BR	MinTER*	%CORR	RIOC[b]	CORIOC	FPR	FNR
P D RAND	1022	9.6	+6.0	78.9	.070*	.235	86.8	8.9
& R INSLAW	1022	9.6	+9.6	77.1	.141**	.330	84.7	8.2
P U SFS81	1022	9.6	+25.8	64.2	.194**	.353[@]	87.0	7.7
G CGR	979	9.3	+10.6	76.7	.135**	.322	85.6	8.0
P R RAND	1022	8.2	+7.4	80.9	.154***	.353	84.9	7.0
& O INSLAW	1022	8.2	+11.0	76.9	.087*	.259	88.8	7.5
P B SFS81	1022	8.2	+27.2	65.2	.281***	.428[@]	87.6	5.9
CGR								
P V RAND	1022	5.3	+10.3	82.6	.211***	.417	88.7	4.2
& I INSLAW								
p O SFS81	1022	5.3	+30.1	65.0	.283**	.429[@]	92.0	3.8
L CGR								

---

* p <= .05
** p <= .01
*** p <= .001

[a]The CGR scale utilizes an educational achievement variable. Data to support this variable were sometimes missing for individuals in the various data sets.

When this occurred, cases missing this datum were excluded from the analysis. This accounts for the reduction in the CGR sample size and the Base Rate differences when compared to other scales.

[b]The assumption of a systematic normal distribution for sample estimates of the RIOC statistic do not apply when a cell frequency in the 2 x 2 table of predicted and actual outcomes do not exceed five. The consequence of low cell frequencies is substantial biases in estimates of standard error for the RIOC statistic. (Copas and Loeber, 1989) Therefore when cell frequencies of five or less occur in this analysis no significance levels are reported.

[@]Largest CORIOC value within a Data set/Crime type

Preston and YCOT data sets, but all have relatively high base rates. The second best performing scale, SFS81, consistently produced the highest CORIOC values when base rates were relatively low.

It is interesting to note that the scale with the highest CORIOC value in the Table 8 comparisons also tends to be the scale with the highest values of MinTER (MinTER = |MinTER*|). High values of MinTER identify a scale with considerable disparity between the BR and SR. Relatively large MinTER values produce higher floors on minimum error and reduce the range of potential improvement in accuracy reflected in the denominator of the RIOC statistic. In the analyses reported here, the magnitude of RIOC is substantially reflected in the resulting CORIOC values. Thus, it is easier to do better, both in relative terms and in the CORIOC statistic, as SR departs from BR and thereby narrows the range of possible correct predictions.[14] Again, absent improvement in measurment precision, it appears that statistical characteristics of the scale (SR) and sample (BR) are more important for predictive accuracy than are theoretical considerations like congruence between the type of offender population used to develop a scale and the population to which a scale is applied.

Identifying the highest CORIOC values among the scales in dataset/crimetype grouping provides a series of ten comparisons (excluding the two analyses with no

[14]Each percentage point improvement in correct predictions will have a larger proportional effect on accuracy as the range of possible values in the denominator declines.

comparisons). The best scale in each set of comparisons (the scale with the highest CORIOC value) is indicated by an ampersand in the CORIOC column of Table 8. The percentage of correct predictions in these 10 analyses ranged from 54% to 72%. The average percent of correct predictions for the best scales is 62% (63% when the two analyses for which there were no comparisons are added). The relatively low RIOC values (ranging from .30 to .51) indicate that the potential for improvement is substantially larger than is achieved by all of the prediction scales. Finally, the differences between the scale ranking (from best to worst) based on mean CORIOC's (across crime types and data sets) in Table 7(B) [SFS81 → RAND → INSLAW → CGR] is different from the ranking observed within specific data set/crime type samples [INSLAW → SFS81 → RAND → CGR]. The change in rank for INSLAW is due to the greater variability in CORIOC values for that scale, which includes the lowest CORIOC values in Table 8. Finally, the scale for assessing parole risk among Federal parole applicants (SFS81) seems to be the most robust of those assessed (ranking first or second in the two ranking procedures).

Focused Analyses.

One way to conceptualize an appropriate test of the transportability of existing scales is to examine their performance across a variety of populations. The results of this type of test, reported above, indicates that none of the scales in our sample performed in a consistently effective manner across the data sets. This broad approach to testing transportability arguably may not be the "fairest" test. For example, it might be argued that a fairer test would examine the performance of scales using populations and decision

contexts that are similar to those used in original scale construction. It is unfair, from this perspective, to assess the performance of a scale designed to predict reappearance at trial (like CGR) on a sample of violent offenders under correctional custody (like P&P Violent). We, therefore, created a conceptually-based test by focusing on the analysis that most closely matches the data set and crime type used to construct each scale. As a final test we examine the best empirical performance of each scale across all the analyses remaining at Stage 2.

To identify the conceptually best fitting population, we considered all analyses available at Stage 1. While the matching process is imperfect, because we were limited to the scales and data set/crime types available, we believe the pairings identified are sufficiently congruent to support meaningful assessments. From the eighty possible pairings, we identified the following as the best conceptual fits:

> The RAND scale was constructed to identify high-rate offenders whose incarceration ought to be extended and was based on samples of jail and prison inmates from California, Michigan and Texas. The conceptually closest data set was P&P which was based on samples of California prisoners and probationers. The Robbery crime type was used because robbery offenders tend most closely to resemble prison inmates.
>
> The INSLAW scale was constructed to identify career criminals for prosecutors using samples of Federal prisoners and probationers. No samples of Federal prisoners or probationers were available. The data set most closely approximating the construction sample is P&P for the category of Total crime type.
>
> The SFS81 scale predicts the risk of reoffending for Federal prisoners who are eligible for parole. Again, the closest pairing was the Total crime type in the P&P data set.

The CGR scale was based on samples of persons in pretrial detention in upstate New York and was used to predict the probability of reappearance at trial and the risk of rearrest for defendants eligible for release on recognizance. The best conceptual fit for this scale was the Total crime type in the DOL data set. These individuals, who came from Miami, Albuquerque and New York City, were referred to jobs training programs after being arrested.

The analysis results associated with these pairings is reported in Table 9. The first column reports data about each scale's performance in the data on which it was originally constructed/validated (hereafter referred to as its "construction data" or "construction samples.") Comparing these data to the results for the best conceptual fit in our data (column 2) presents a somewhat different picture from that obtained in the overall Stage 2 analyses. The overall prediction accuracy of the RAND scale actually improves by almost eight percentage points from the construction samples (73.0%) to the P&P Robbery sample (80.8%). However, the RIOC statistic deteriorates by more than half (dropping to .154), indicating that the ability to minimize prediction errors is drastically lower in our analysis where the Robbery sample has a lower BR and the scale produces a lower SR. The conjoint effect of the two assessment criteria, seen by comparing the CORIOC values, is that when transported to the P&P Robbery sample the RAND scale is not robust as is evident in the deterioration of performance from its original construction samples.

The percent correct for the INSLAW scale deteriorated 12.2 percentage points from the construction sample (70.3%) to the P&P Total sample (58.1%) and the RIOC value

Table 9.

COMPARISONS OF BASELINE PREDICTION ACCURACY DATA FROM EACH SCALE WITH DATA FROM THE CONCEPTUALLY CLOSEST DATA SET/CRIME TYPE FOR EACH SCALE AND WITH DATA FROM THE BEST EMPIRICAL FIT WITH EACH SCALE

	Construction/ Validation Data	Conceptual Best Fit[1] Data	Empirical Best Fit Data
RAND		P&P Robbery	P&P Total
BR	28.0	8.2	45.6
SR	28.9	15.6	15.6
MinTER*	+.009	+.074	-.300
%CORR	73.0	80.8	60.4
RIOC	.345***	.154***	.434***
CORIOC	.502	.353	.512
INSLAW		P&P Total	P&P Total
BR	41.6	45.6	45.6
SR	11.7	19.2	19.2
MinTER*	-.299	-.264	-.264
%CORR	70.3	58.1	58.1
RIOC	.743***	.295***	.295***
CORIOC	.703	.414	.414
SFS81		P&P Total	P&P Total[3]
BR	31.5	45.6	45.6
SR	34.2	35.4	35.4
MinTER*	-.027	-.102	-.102
%CORR	65.8	63.2	63.2
RIOC	.239***	.309***	.309***
CORIOC	.397	.442	.442
CGR		DOL Total	Preston Violent
BR	22.5	46.9	43.9
SR	9.0	6.3	35.9
MinTER*	-.135	-.406	-.080
%CORR	72.4	55.6	59.6
RIOC	.181***	.439***	.195***
CORIOC	.362	N/A[2]	.341

---

[1]Closest conceptual match of Scale and Data set
[2]Did not meet the criteria for inclusion in the Stage 2 analysis
[3]Best overall empirical performance of any Scale

again dropped below half of the value in the original construction sample (from .743 to .295). The indicator of joint effect, CORIOC, was the highest of all four scales in their construction samples (.703). This declined to .414 in the P&P Total sample indicating the INSLAW scale was moderately robust in this application. It is interesting to note from the values of MinTER* in both the construction sample (-.229) and the P&P Total sample (-.264), the INSLAW scale is tilted toward making a substantial proportion of false negative errors.

With the SFS81 scale, the total percent correct deteriorated the least of all the conceptually matched pairs (2.6 percentage points), although the original percent correct was somewhat lower than the other scales (65.8%). The RIOC value was higher in the P&P Total sample (.309) than in the original construction sample (.239) as was the CORIOC value (.442 compared with .397). Thus, the SFS81 scale was robust when transported to the P&P Total sample. Based on a MinTER* value near zero in the construction data, the originally designed prediction model did not structurally favor false negative or false positive errors. A slight structural bias favoring false negative errors did arise in the P&P Total Sample (MinTER* = -.102).

The optimum conceptual match for the CGR scale was excluded from the Stage 2 Analysis because the discrepancy between the BR and SR was so large that the structural minimum level of prediction errors would have involved more than 40% of the cases (MinTER = -.406). When compared with the DOL Total pairing there is a 17 percentage

point reduction in correct predictions. The RIOC values, however, indicate a gain (from .181 to .439) in relative prediction accuracy when the scale is applied to the DOL Total sample (CORIOC value = .244). On balance, the conjoint effect is dominated by the decline in total correct predictions.

In comparisons to the conceptually best matched data the SFS81 scale is the most robust, even though its total prediction accuracy is considerably below the RAND scale. The RAND scale had the highest overall prediction accuracy (over 80%), but the lowest relative improvement in accuracy (RIOC = .154). The INSLAW scale was the least robust, resulting in the largest reductions in both total correct predictions and relative improvements in accuracy. The conceptual match for the CGR scale did not produce a viable analysis. Since the structural constraint on minimum error was high, the model resulted in a low percentage of correct predictions on the conceptual best fit data.

For the INSLAW and SFS81 scales, the empirical best fits (last column of Table 9) were identical to the conceptually based pairings. In both cases this involved the sample reflecting the P&P Total data. For the RAND scale the best empirical fit also was with P&P Total but the conceptual best fit was with P&P Robbery. However, these two samples are conceptually quite similar. Finally, Preston Violent was the best fit for the CGR scale. The results from examining the empirical best fit for each scale do not modify our previous conclusions about the relatively better performance of the SFS81 and RAND scales when transported to new data.

Finally, it is interesting to note that the RAND scale, both in the construction samples and in the P&P Robbery sample, has a value of MinTER* is that is very close to zero indicating that the scale is not structurally fitted to make false negative prediction errors. This is the only scale among those analyzed whose design does not require that it make a certain number of false negative errors. While it may not have been intentional, it is a matter of empirical fact that the structure of errors in the INSLAW, SFS81 and CGR scales systematically favor releasing offenders who reoffend, a result that might be expected when greater priority is given to minimizing false positive errors in the design of a scale.

CONCLUSIONS

First, we conclude that the two stage procedure employing the MinTER and CORIOC statistics is a useful approach for making comparisons among scales and across data sets. The MinTER screening criterion eliminated analyses that are structurally constrained to have excessive prediction errors. The CORIOC statistic, and its constituent parts (the total percent correct and the RIOC statistic), effectively discriminate among scales on the basis of their predictive performance and their relative fit. In our analysis, the best performing scale for samples that have high base rates was INSLAW. The SFS81 scale was most effective with samples that have low base rates, and it was the best performing scale both in the overall analysis of all study data and when the conceptual "best fit" analysis was compared with scale performance in its original construction data. It also was the second best performing in the individual Stage 2 analyses.

Notably, the SFS81 scale employs Burgess-type scores instead of empirically estimated weights confirming yet again Wilkins' (1969) observations about this matter. The former are less vulnerable to overfitting the construction data, which increases their potential transportability to new data. The SFS81 scale - - designed to inform parole release decisions about federal prisoners - - also was developed on selected subsamples of offenders who are likely to be more homogeneous. The resulting scale instrument is thus more likely to be sensitive to offender attributes that are meaningfully related to recidivism and influenced less by extraneous variations that are likely to be found in highly heterogeneous samples of offenders.

On the other hand, none of the scales performed very well or very consistently. Even after the analyses in which excessive amounts of structural error were excluded, the magnitude of observed error averages 35%, of which about half is structurally induced by the relationship between the BR and SR. Furthermore, most scales were dominated by false negative errors, or recidivists among offenders who are classified as non-recidivists. In today's political climate, it seems unlikely that policy makers would choose scales systematically producing errors of commission - - ie., releasing offenders more likely to recidivate. In addition, prediction errors increased in most of the applications to new data. On the basis of our findings, we must support the conventional wisdom that transporting existing scales is not a viable option. At a minimum, transporting scales across populations or decision contexts would require re-norming the scale on the target population.

Interestingly, we found that without considerable improvement in measurement ability, theoretical considerations about what factors affect recidivism are of less importance in successful analyses than the statistical characteristics of the sample and the scale (ie., BR and SR). Finally, we believe that the statistics we devised for this analysis (MinTER* and CORIOC) are well suited to examining the performance of prediction scales generally and especially during scale development when optimal prediction variables and scale cutpoints are being selected.

REFERENCES

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.

Copas, J. and Loeber, R.
1989 The statistical properties of the index: Relative improvement over chance (RIOC). British Journal of Mathematical and Statistical Psychology.

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.

APPENDIX A

SCALE SCORING RULES

RAND Scale

Scoring: +1 for each attribute of offender.

Classification Rule: offenders with scores greater than 3 are classified as high-rate.

Scale Criteria:

Prior convictions for the same charge (robbery or burglary)
Incarcerated more than 50 percent of 2 years prior to present commitment to prison or jail
Convicted before age 16
Served time in state juvenile facility
Drug use during 2 years preceding present commitment to prison or jail
Drug use as a juvenile
Employed less than 50 percent of the 2 years prior to present
0 commitment to prison or jail

(Source: Greenwood and Abrahamses, 1982:Table A-4)

INSLAW Scale

Scoring: points assigned for each attribute of offender.

Classification Rule: offenders with scores greater than 46 are classified as "career criminals."

Scale Criteria:

Variable	Points
Heavy use of Alcohol	+ 5
Heroin Use	+10
Age at time of instant arrest	
Less than 22	+21
23 - 27	+14
28 - 32	+ 7
33 - 37	0
38 - 42	- 7
43+	-14
Length of criminal career (time from 1st arrest to target arrest)	
0 - 5 years	+ 0
6 - 10 years	+ 1
11 - 15 years	+ 2
16 - 20 years	+ 3
21+ years	+ 4
Arrests during last five years	
Crimes of violence	+ 4/arrest
Crimes against property	+ 3/arrest
Sale of drugs	+ 4/arrest
Other offenses	+ 2/arrest
Longest time served, single term	
1 - 5 months	+ 4
6 - 12 months	+ 9
13 - 24 months	+18
25 - 36 months	+27
37 - 48 months	+36
49+ months	+45
Number of probation sentences	+1.5/sentence
Instant offense was a crime of violence[a]	+ 7
Instant was a crime labeled "other[b]	-18

[a]Violent crimes include homicide, assault, robbery, sexual assault and kidnaping.

[b]"Other" crimes include military violations, probation, parole, weapons and all others except arson, burglary, larceny, auto theft, fraud, forgery, drug sales or possession, and violent crimes.

(Source: , Rhodes, et al, 1982:Table V.1)

SFS81 Scale

Scoring: points assigned for each attribute of offender.

Classification Rule: offenders with scores of less than 4 are classified as poor parole risks.

Scale Criteria:

Variable	Points
Prior convictions/adjudications (adult or juvenile)	
None	+3
1	+2
2 or 3	+1
4 or more	0
Prior commitment(s) of more than 30 days (adult or juvenile)	
None	+2
1 or 2	+1
3 or more	0
Age at current offense/prior commitments	
Age at commencement of current offense:	
26 years or more	+2
20 - 25 years	+1
19 years or less	0
Except	
If five or more prior commitments of more than 30 days (adult or juvenile),	
Place an X here _____, and	
Score this item	0
Recent commitment free period (3 years)	
No prior commitment of more than 30 days (adult or juvenile) or released to the community from last such commitment at least 3 years prior to the current offense	+1
Otherwise	0
Violations of probation/parole/confinement conditions	
or escaped at time of instant offense	+1
No history of heroin/opiate dependence	+1

(Source: U.S. Parole Commission, 1985:45)

CGR Scale (Composite model)

Scoring: add logistic regression weights for each attribute of offender.

Classification Rule: offenders with scores greater than 1.43 are classified as high risk for pretrial release.

Scale Criteria:

Variable	Weight
Number of prior violent felony arrests in the last 5 years	+.3680/arrest
Number of prior non-felony arrests in the last 5 years	+.1205/arrest
Length of time at current employment (in months)	-.0082/month
Years of education	-.0766/year

(Source: Center for Governmental Research, 1982-83:158)

APPENDIX B
RIOC ANALYSIS TABLE

PRESTON		N	BR	SR	MinTER*	%CORR	RIOC	CORIOC	FPR	FNR
VIOLENT	RAND	1596	45.0	31.5	-13.5	56.5	.135***	.276	47.6	41.6
	INSLAW	1596	45.0	18.4	-26.6	57.4	.209***	.346	43.5	42.4
	SFS81	1596	45.0	29.8	-15.2	56.1	.125***	.265	48.1	42.1
	CGR	1056	43.9	35.9	- 8.0	59.6	.195***	.341	45.1	37.8
ROBBERY	RAND	1596	35.8	31.5	- 4.3	62.3	.174***	.329	53.0	30.7
	INSLAW	1596	35.8	18.4	-17.4	63.9	.210***	.366	50.7	32.8
	SFS81	1596	35.8	29.8	- 6.0	60.5	.126***	.276	56.1	32.4
	CGR	1056	35.0	35.9	+ 0.9	60.3	.136***	.286	56.5	30.3
PROPERTY	RAND[c]	1596	74.5	31.5		45.1	.258***		18.9	71.5
	INSLAW[c]	1596	74.5	18.4		36.5	.213*		20.1	73.3
	SFS81[c]	1596	74.5	29.8		43.2	.201**		20.4	72.3
	CGR[c]	1056	73.8	35.9		46.4	.165**		21.9	71.3
DRUGS	RAND	1596	47.2	31.5	-15.7	55.6	.139***	.278	45.4	43.9
	INSLAW	1596	47.2	18.4	-28.8	55.1	.175***	.311	43.5	45.2
	SFS81	1596	47.2	29.8	-17.4	51.6	.016	.091	51.9	46.9
	CGR	1056	50.8	35.9	-14.9	55.2	.153***	.291	41.7	46.5
TOTAL	RANDc	1596	93.2	31.5		36.2	.529***		3.2	91.6
	INSLAW	1596	93.2	18.4		24.4	.689***		2.0	92.2
	SFS81[c]	1596	93.2	29.8		33.0	.131		5.9	92.9
	CGR[c]	1056	93.7	35.9		41.0	.709***		1.8	91.1

YCOT		N	BR	SR	MinTER*	%CORR	RIOC	CORIOC	FPR	FNR
VIOLENT	RAND	1079	35.1	9.6	-25.5	63.5	.126*	.283	56.7	34.3
	INSLAW	1079	35.1	19.5	-15.6	62.8	.149***	.306	55.2	32.8
	SFS81	1079	35.1	26.5	- 8.6	59.9	.084*	.224	59.4	33.2
	CGR	830	34.3	51.1	+16.8	53.1	.103*	.234	62.3	30.8
ROBBERY	RAND	1079	27.2	9.6	-17.6	69.8	.101*	.266	65.4	26.5
	INSLAW	1079	27.2	19.5	- 7.7	66.4	.090**	.244	66.2	25.7
	SFS81	1079	27.2	26.5	- 0.7	64.5	.096***	.249	65.7	24.7
	CGR	830	26.9	51.1	+24.2	53.8	.166**	.299	68.9	22.4
PROPERTY	RAND[c]	1079	68.6	9.6		37.1	.357**		20.2	67.4
	INSLAW[c]	1079	68.6	19.5		39.4	.060		29.5	68.1
	SFS81[c]	1079	68.6	26.5		47.0	.343***		20.6	64.7
	CGR	830	69.0	51.1	-17.9	54.1	.116*	.251	27.4	65.3
DRUGS	RAND	1079	35.3	9.6	-25.7	63.2	.108	.261	57.7	34.6
	INSLAW	1079	35.3	19.5	-15.8	58.4	-.043[b]	.000	66.2	35.7
	SFS81	1079	35.3	26.5	- 8.8	57.1	.005	.053	64.3	35.2
	CGR	830	35.5	51.1	+15.6	51.5	.051	.162	62.7	33.7
TOTAL	RAND[c]	1079	92.3	9.6		16.6	.500		3.8	91.9
	INSLAW[c]	1079	92.3	19.5		25.3	.381*		4.8	91.6
	SFS81[c]	1079	92.3	26.5		32.2	.500**		3.8	90.9
	CGR[c]	830	69.0	51.1		53.9	.379***		4.7	89.4

P&P		N	BR	SR	MinTER*	%CORR	RIOC	CORIOC	FPR	FNR
VIOLENT	RAND	1022	5.3	15.6	+10.3	82.6	.211***	.417	88.7	4.2
	INSLAW	1022	5.3	19.2	+13.9	78.3	.083	.255	92.9	4.8
	SFS81	1022	5.3	35.4	+30.1	65.0	.283**	.429	92.0	3.8
	CGR	979	5.3	19.9	+14.6	76.8	.035	.000	94.9	5.4
ROBBERY	RAND	1022	8.2	15.6	+ 7.4	80.8	.154***	.353	84.9	7.0
	INSLAW	1022	8.2	19.2	+11.0	76.9	.087*	.259	88.8	7.5
	SFS81	1022	8.2	35.4	+27.2	65.2	.281***	.428	87.6	5.9
	CGR	979	8.3	19.9	+11.6	74.9	-.070	.000	92.3	8.4
PROPERTY	RAND	1022	24.9	15.6	- 9.3	72.4	.222***	.401	58.5	21.8
	INSLAW	1022	24.9	19.2	- 5.7	68.3	.097**	.257	67.9	23.1
	SFS81	1022	24.9	35.4	+10.5	65.2	.244***	.399	64.1	18.8
	CGR	979	23.9	19.9	- 4.0	69.7	.131***	.302	66.2	21.4
DRUGS	RAND	1022	9.6	15.6	+ 6.0	78.9	.070*	.235	86.8	8.9
	INSLAW	1022	9.6	19.2	+ 9.6	77.1	.141**	.330	84.7	8.2
	SFS81	1022	9.6	35.4	+25.8	64.2	.194**	.353	87.0	7.7
	CGR	979	9.3	19.9	+10.6	76.6	.135**	.321	85.6	8.0
TOTAL	RAND	1022	45.6	15.6	-30.0	60.3	.434***	.512	30.8	41.3
	INSLAW	1022	45.6	19.2	-26.4	58.1	.295***	.414	40.3	42.3
	SFS81	1022	45.6	35.4	-10.2	63.2	.309***	.442	37.6	36.4
	CGR	979	44.7	19.9	-24.8	56.8	.165**	.306	46.2	42.5
DOL		N	BR	SR	MinTER*	%CORR	RIOC	CORIOC	FPR	FNR
VIOLENT	RAND	746	6.6	1.5	- 5.1	91.9	-1.00[b]	.000	100.0	6.7
	INSLAW	746	6.6	0.4	- 6.2	92.9	-1.00[b]	.000	100.0	6.6
	SFS81	746	6.6	0.1	- 6.5	93.3	-1.00[b]	.000	100.0	6.6
	CGR	746	6.6	6.3	- 0.3	88.4	.044	.197	89.4	6.3
ROBBERY	RAND	746	11.5	1.5	-10.0	87.3	-.211[b]	.000	90.9	11.6
	INSLAW	746	11.5	0.5	-11.1	88.6	.623[b]	.743	33.3	11.3
	SFS81	746	11.5	0.1	-11.4	88.4	-1.00[b]	.000	100.0	11.5
	CGR	746	11.5	6.3	- 5.2	84.3	.062	.229	83.0	11.2
PROPERTY	RAND	746	18.9	1.5	-17.4	80.7	.215[b]	.417	63.6	18.6
	INSLAW	746	18.9	0.4	-18.5	81.2	.589[b]	.692	33.3	18.7
	SFS81	746	18.9	0.1	-18.8	71.0	-1.00[b]	.000	100.0	18.9
	CGR	746	18.9	6.3	-12.6	78.8	.160**	.355	68.1	18.0
DRUGS	RAND	746	6.2	1.5	- 4.7	92.6	.031[b]	.169	90.9	6.1
	NSLAW	746	6.2	0.4	- 5.8	93.7	.290[b]	.521	66.7	6.1
	SFS81	746	6.2	0.1	- 6.1	93.7	-1.00[b]	.000	100.0	6.2
	CGR	746	6.2	6.3	+ 0.1	88.8	.049	.209	89.4	5.9
TOTAL	RAND[c]	746	46.9	1.5		53.2	.144		45.5	46.8
	INSLAW[c]	746	46.9	0.4		53.5	1.00		0.0	46.7
	SFS81[c]	746	46.9	0.1		52.9	-1.00[b]		100.0	47.0
	CGR[c]	746	46.9	6.3		55.6	.439***		29.8	45.4

* p <= .05
** p <= .01
*** p <= .001

[a]The CGR scale utilizes an educational achievement variable. Data to support this variable were sometimes missing for individuals in the various datasets. When this occurred, cases missing this datum were excluded from the analysis. This accounts for the reduction in the CGR sample size and the Base Rate differences when compared to other scales.

[b]The test for statistical significance of the RIOC is found in Farrington and Loeber (1989). The assumption of a systematic normal distribution for sample estimates of the RIOC statistic do not apply when a cell frequencie in the 2 x 2 table of predicted and actual outcomes do not exceed five. The consequence of low cell frequencies is substantial biases in estimates of standard error for the RIOC statistic. (Copas and Loeber, 1989) Therefore, when cell frequencies of five or less occur in this analysis no signifi ance levels are reported.

[c]Cases are excluded from analysis because the minimum TER that is possible (MinTER=/MinTER') exceeds 33%. For excluded cases, MinTER* and CORIOC are not reported in this table.