Validating Risk Assessment Instruments Used in Community Corrections

January 1991

Christopher Baird

National Council on Crime and Delinquency
6409 Odana Road
Madison, Wisconsin

147395

Validating Risk Assessment Instruments Used in Community Corrections

January 1991

Christopher Baird

National Council on Crime and Delinquency
6409 Odana Road
Madison, Wisconsin

Validating Risk Assessment Instruments Used in Community Corrections

January 1991

Christopher Baird

National Council on Crime and Delinquency
6409 Odana Road
Madison, Wisconsin

TABLE OF CONTENTS

TABLE OF TABLES AND FIGURES

INTRODUCTION

A. Historical Perspective

While risk assessment was clearly not a new idea, for all practical purposes it was "discovered" in the 1970s and operationalized in the 1980s. Before 1980, risk assessment was limited to a few research papers or used somewhat idiosyncratically by a few correctional agencies. Even the current term "risk assessment" was not part of the correctional nomenclature, and risk scales took on a variety of titles, sometimes named for their developers (i.e., Burgess Scaling), other times using statistical or descriptive titles (the California Base Expectancy Tables; the Federal Salient Factor Scale). However, as probation and parole caseloads began to swell in the late 1970s, agencies sought methods for stretching their limited resources to continue to provide the most effective services possible. Obviously, as caseloads increased, exceeding 100 cases per officer in many agencies, corrections could no longer afford to see all offenders as often as desired; some method for establishing priorities was needed. The field turned quite naturally to risk assessment; it was an idea whose time had finally arrived.

Agencies adopting risk screening techniques had two options. Some -- to a large degree, those with in-house research capability -- developed their own instruments. Most, however, adopted instruments developed in other jurisdictions, sometimes incorporating minor modifications to reflect differences in policy or terminology. Remarkably, over the course of one short decade, the practice of probation and parole in the United States was altered significantly. Risk assessment went from a seldom-used technology in 1980 to the principle case management tool of probation and parole agencies by 1990.

The emergence of risk assessment as a method for sorting cases for supervision purposes was due, to a large extent, to the National Institute of Correction's (NIC) Model Probation/Parole Management Project and to other NIC technical assistance efforts. The model project not only spread the use of risk assessment instruments, but also led to considerable standardization in how these instruments were used by probation and parole agencies

throughout the nation. Comparisons of current practice indicate that reclassification schedules, contact standards, and use of needs assessments show only minor variance from jurisdiction to jurisdiction (NCCD, 1990).

In all instances of rapid change, solutions to existing problems create new problems, and the switch to risk assessment systems for setting supervision priorities in probation and parole proved no exception. As the use of risk assessment spread, the research community began to worry that instruments developed in one jurisdiction and transferred to another may not "work" for the adopting agency (Wright, Clear, Dickson; 1984). After all, populations, crime rates, and living situations vary significantly from region to region, state to state. What predicts risk in a rural midwestern state may have little connection to risk in New York or Los Angeles. Furthermore, a follow-up study of the NIC Model Project effort indicated that, in most agencies, original expectations regarding testing and validation of adopted risk assessment scales were lost in the crush of everyday operations (NIC, 1989).

Many agencies have also observed a gradual shift toward higher risk classifications. Since this often translates into the need for additional staff, funding agencies -- state legislatures and county boards -- began to directly question if these changes were legitimate and indirectly question the validity of risk assessment scales. Two obvious questions are raised by higher classifications: "Is the increase in average risk scores due to changes in offender characteristics?" and "Are risk scales developed ten to fifteen years ago still valid?" All of the issues raised by researchers, changes in offender profiles, and the passage of time have led to increased interest in scale validation.

The purpose of this monograph is to explore validation issues. It is designed as an operations or "how to" manual covering issues of sample size requirements; data needs; outcome or criterion variables; methods of analysis; what to look for and how to interpret results; and, finally, where to go for assistance. Results of recent validation studies are used throughout the monograph to illustrate specific points and to clarify the discussion of issues. It is also understood that not all probation and parole systems have the resources to support a

comprehensive validation effort. Hence, options are presented, ranging from a "bare bones" study with limited objectives to rather sophisticated reviews meant to influence policy decisions at various levels of the organization. The latter type of study goes beyond issues of scale validity and examines operations and impact. As such, these studies represent comprehensive evaluations of classification systems.

Before discussing the above topics, it is necessary to define terms and to establish appropriate expectations. Over the last decade, risk assessment has come to mean many different things to different people. Even among researchers, there is disagreement on how scale efficacy is measured. It is important that the discussion of risk validation begin with a common understanding of both the intent and potential of these instruments. The next section establishes these parameters, with an emphasis on practical rather than statistical or theoretical issues.

B. Issues in Validation

1. The Need for Validation

The 1980s saw record increases in the number of offenders under correctional supervision. While most of the public's attention has been focused on burgeoning prison populations, the fact is that the number of persons on probation and parole has risen at an even faster pace. Between 1980 and 1988, prison population grew 90%; during the same period the number of offenders on probation and parole more than doubled, growing at a 110% rate.

Table 1-A

CORRECTION POPULATIONS PERCENT CHANGE 1980 TO 1988

	1980	1988	% Change
Probation	1,118,097	2,356,483	111%
Jails	163,994	343,569	110%
Prison	329,821	627,588	90%
Parole	220,438	407,977	85%
Adult Arrests	6.1 million	8.5 million	39%
Reported Index Crimes	13.4 million	13.9 million	4%

Sources: Historical Corrections Statistics in the United States, 1850-1984, U.S. Department of Justice, Bureau of Justice Statistics.

Uniform Crime Reports: Crime in the United States, 1980 and 1988, U.S. Department of Justice, Federal Bureau of Investigation.

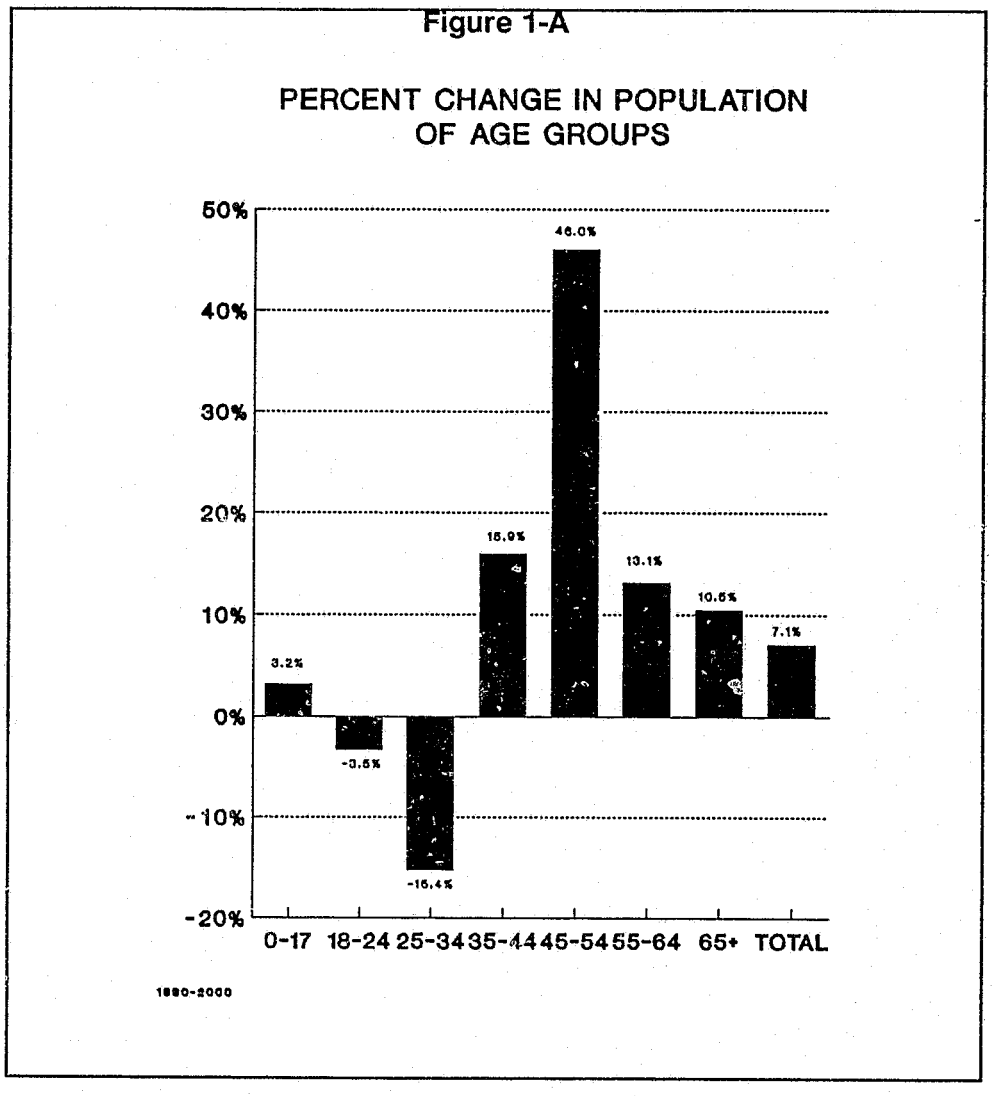
Census of Local Jails, 1988, U.S. Department of Justice, Bureau of Justice Statistics.

Prisoners in 1989, U.S. Department of Justice, Bureau of Justice Statistics.

Sourcebook of Criminal Justice Statistics, 1988, U.S. Department of Justice, Bureau of Justice Statistics.

The 1980s were also a time of substantial change in the demographic make-up of the United States as well as profound changes in sentencing policy. As the "baby boom" generation aged, the proportion of our population in the high crime-prone years (generally defined as those under 35) declined and, as Figure 1-A illustrates, this trend will continue through the 1990s.

Figure 1-A



Source: Paine Webber Research

Although this portends some changes in offender populations, they are almost inconsequential when compared to changes brought about by revisions in sentencing practices. In the 1980s, many states, as well as the federal government, instituted harsher penalties for all types of offenses, particularly for drug-related crimes. The war on drugs has resulted in massive increases in the number of drug offenders -- comprised largely of minority youth -- entering the

criminal justice system. In Florida, for example, 73% of all drug offenders are Black compared to 53% of all other prison admissions.

Table 1-B presents increases in the number of admissions to prison for drug offenses in eight selected states over the last few years. Similar patterns are noted in probation and parole. The incidence of substance abuse is currently so widespread among offender populations that several recent risk studies have demonstrated that drug or alcohol abuse no longer separates successes from failures. In essence, if nearly everyone in a population shares a characteristic, classification based on that characteristic is not possible.

Table 1-B

INCREASES IN DRUG OFFENSE PRISON ADMISSION FOR SELECTED STATES

States	Time Period	% Increase
Virginia	July 1986 - June 1989	136%
Michigan	July 1986 - June 1989	201%
Oklahoma	July 1986 - September 1989	174%
Florida	July 1986 - January 1989	168%
Tennessee	July 1986 - June 1989	128%
Illinois	July 1986 - June 1989	156%
Nevada	January 1986 - December 1988	107%
California	January 1982 - December 1987	635%

Mandatory arrest and sentencing practices resulted in other changes as well. Most notably, drunk driving and domestic violence cases on probation and parole caseloads have increased dramatically in recent years. For example, in a recent study of probation in Iowa, 47%

of admissions were "Operating While Intoxicated" (OWI or drunk driving) cases. These two populations were virtually non-existent in probation/parole when most risk scales currently in use were developed 15 years ago. Hence, little is known regarding their applicability to these offender groups.

Changes in population parameters represent one reason why scale revalidation is needed. The fact that most agencies are using risk assessment instruments from another jurisdiction is of equal importance. Despite the NIC recommendation that risk assessment and outcome data be collected routinely so that periodic revalidation could be completed, few agencies have designed and implemented information systems that support such research. As a result, few agencies have validated risk instruments that are used to make important decisions about offenders.

Validation studies need to address questions regarding applicability of risk scales to offender subgroups. Offender base rates -- that is, rates of success/failure on probation and parole -- vary significantly by ethnicity, gender, and offense groups. It is important to know if a single instrument is capable of effectively separating offenders based on risk for all of these subpopulations, or if different scales are required for various groups. Since women generally represent less than ten percent of an offender population, they have little influence in the statistical analyses used to develop risk assessment scales. Because most instruments used today were based primarily on male populations, their applicability to female offenders is a particularly significant issue.

In addition, decisions regarding high profile offenses, sex crimes, drug sales, and crimes of violence are important to both corrections officials and the general public. Information regarding recidivism rates and the ability of risk instruments to appropriately classify these offenders can help establish policy, enlighten the public, and defend agency practices when crises occur. Corrections, after all, is in the business of managing risk. While risk to the community cannot be completely controlled with anything less than total incapacitation, the public is right to insist that correctional decisions regarding supervision are based on the best

information available. Data on the risk presented by various offense groups are becoming increasingly important to policy development and practice in probation and parole.

Finally, validation studies are needed simply to increase staff confidence in the instrument used. Staff turnover, changing offender profiles, new administrators with new policies and procedures, and increases in overall workload often result in a loss of general agency knowledge regarding the origin and purpose of the classification system. In validation studies, staff concerns can be addressed and changes made that reflect current conditions and circumstances. Even if changes in scale design are relatively minor (or not required at all), data that demonstrate the effectiveness of the system will bolster staff confidence and diffuse the arguments of disbelievers.

In sum, due to changes in offender populations that occur over time (sometimes rather rapidly when public policy changes) and the need to examine the applicability of an instrument to subgroups of offenders, revalidation efforts should be completed periodically. While no rule of thumb can be applied to determine how frequently agencies should undertake such research, the degree and frequency of social and/or legislative changes determine when revalidation efforts should be undertaken.

2. What is Risk Assessment?

Although risk assessment is widely used in community corrections, the field is not entirely clear on what it actually represents. Even less is understood about properties of a good assessment system. Prior to embarking on a discussion of validation issues, it is, therefore, important to define terms and expectations and to identify issues often misunderstood by researchers working in the risk assessment arena. These misunderstandings can lead to a misuse of statistical procedures, as well as errors in the interpretation of results.

Several different terms have become associated with risk assessment in community corrections: chief among these are prediction and classification. These are often used interchangeably, yet really connote different results. Prediction, by definition, is more precise

than classification. According to Webster, prediction "declares in advance on the basis of observation, experience, or scientific reason." To predict accurately in any field is difficult; to predict human behavior accurately is especially complex as so many factors contribute to determining how individuals will act. Classification, on the other hand, is simply "a systematic arrangement in groups or categories according to established criteria." While accurate prediction would greatly benefit corrections and society, it has not proven feasible in criminal justice. We submit that goals of risk assessment are much more modest; it is simply meant to assign offenders to different categories based on observed rates of success or failure (however defined) on probation or parole.

The false expectation that probation or parole outcomes can be accurately predicted leads to the use of multivariate statistical techniques such as linear regression to evaluate the performance of risk instruments. Regression analysis is a rather powerful statistical technique which can be of some assistance in scale development. It is, however, not appropriate for evaluating the efficacy of risk assessment scales.

Regression produces a statistic (r^2 - the coefficient of determination) which represents the amount of variance explained in the outcome measure (recidivism) by factors in the equation (generally social and criminal history variables). For example, if a risk scale contains 10 factors, these are entered into an analysis that attempts to explain why some offenders fail and others succeed. If failure rates increase in exact increments relative to risk score increases, risk scores and outcomes are perfectly correlated, and an r^2 of 1 is attained. In effect, all variance in criminal behavior is explained. We would know, precisely, which offenders will always succeed, which will always fail, and the "relative" success or failure of all those offenders with less than perfect outcomes.

Risk scales, however, explain little of the variance in offender outcomes -- 8% to 15% is common. This fact leads some researchers to caution against the use of risk assessment, claiming these instruments are not valid because they fail to predict accurately who will succeed and who will fail. But if simple classification is the goal, the degree of variance in criminal activity

explained is of little consequence. What is important is the degree to which offenders in different risk groups perform differently. Valid risk instruments achieve significant differences in rates of recidivism among risk groups -- the greater the differences, the better the instrument. Data from South Carolina illustrate how a risk scale that explains less than 10% of the variance in criminal behavior can still provide valuable information to (in this case) a parole board (NCCD, 1985).

The overall failure rate (excluding minor violations and traffic offenses) for the South Carolina parole sample was 30.5%.

Failure rates for each risk group identified were:

Low risk	7.6%
Moderate risk	22.5%
High risk	44.7%

The rates of violent convictions reported were:

Low risk	0%
Moderate risk	4.6%
High risk	9.3%

Serious repeat and violent offense rates were:

Low risk	0%
Moderate risk	5.3%
High risk	19.3%

This scale obviously separates groups of offenders, based on probability of success on parole, very well despite the inability to explain much of the variance in criminal behavior. High-risk cases re-offended at nearly six times the rate of low-risk cases and they are far more likely

to commit serious offenses. The issue then is: What would have been the effect of using the risk assessment instrument in the parole decisionmaking process?

Because risk is only one factor considered in parole decisions, it is impossible to determine precisely how application of risk assessment would have affected release decisions. But, a hypothetical example can be established. If the half of the sample which scored the lowest on the scale had been paroled and the remaining half had served their full terms:

- Ninety-three more individuals would have been paroled, an increase of 12.6%.
- Total offenses among parolees would have decreased by 122.
- The number of violent and serious repeat offenders would have decreased from 112 to 44.

In other words, more people would be released on parole, yet community safety would be substantially enhanced.

To clarify the goal of risk assessment in community corrections, it may prove wise to drop the notion of prediction altogether. Although most researchers clearly understand the nuances of prediction terminology, it leads to false expectations among the less experienced. Recently, a risk instrument that effectively separated high-, moderate-, or low-risk youth in a midwestern state was judged "invalid" by an evaluator because it explained less than 10% of variance in outcomes. To discontinue use of the scale would have represented a serious setback, as the agency has incorporated risk assessment into a structured decision system that promised to enhance consistency and appropriateness of placements, and result in considerable savings as well. Fortunately, outside review was requested and the agency proceeded with implementation plans.

More is presented on issues of scale construction and validation in subsequent sections of this report. At this point, the discussion can be summarized with the following critical point:

- **The purpose of risk assessment instruments is to separate groups of offenders to the maximum extent possible based on rates of success/failure. Therefore, the value of risk instruments should be based on their ability to separate offender groups rather than their ability to explain variance in criminal behavior.**

3. Properties of Good Risk Assessment Systems

There are four properties present in all good decision systems including risk assessment. These are:

- Validity
- Reliability
- Equity
- Utility

Validation studies, of course, directly examine the issue of scale validity. In its broadest context, validity means that a system accomplishes its objectives. A risk assessment instrument is valid if it separates groups of offenders based on rates of failure. When developing risk instruments, the goal is to achieve the maximum difference in failure rates possible. Experience demonstrates that, in most instances, a four to one ratio or at least a difference of 30% in failure rates between the highest and lowest risk groups can generally be achieved.

The remaining properties -- reliability, equity, and utility -- should also be examined in a comprehensive evaluation of risk assessment. Reliability is present if the risk assessment score given an individual is the same regardless of who completes the scale. The best way to attain reliability is to use objective factors to rate risk, provide thorough definitions of items and item values, and adequately train staff in use of the instrument. Reliability and validity are inextricably linked, since errors in ratings obviously produce invalid results.

Equity goes beyond validity and reliability to require that use of given factors in risk assessment must be fair -- it does not discriminate against subgroups in a society -- and

justifiable -- its use is consistent with broader social values (Clear, Baird, 1986). The issue of equity in scale construction and validation centers around areas of gender, ethnicity, and age. Various offender populations have substantially different rates of recidivism. This fact alone makes the identifying characteristic (e.g., male, under 25) "predictive." Because of bias in society and the criminal justice system, there are no absolutely "clean" criteria for equitable differentiation of offenders other than the current offense. Systems should, however, expunge factors that directly discriminate (such as ethnicity) and then monitor operations to determine the effects of risk assessment systems on various offender populations. For this reason, validation studies should examine how the system works for males, females, the major ethnic groups in the population, and for various age breakdowns.

Utility is a basically pragmatic criterion. Risk assessment scales should be simple, efficient, and the relationships between risk factors and outcomes evident to staff (face validity). The most valid of scales will not benefit operations if not accepted and used appropriately by staff in the supervision process. Experience indicates that complex systems will be resisted by staff who, generally, already feel inundated with paperwork and case processing requirements of the legal system. Simplicity will also enhance rater reliability, as errors in scoring will be minimized. To enhance utility, validation efforts should seek to simplify scales whenever possible, provided such changes do not reduce the scale's ability to effectively separate risk groups.

C. Needs Scale Validation

Most agencies have been primarily interested in validating risk instruments, and few have expressed interest in the validity of needs assessment tools. This probably reflects increasing reliance on risk management concepts, diminishing resources for service delivery, and the fact that the majority of supervision level decisions are based on risk rather than need, even in agencies which utilize both assessments.

Conceptually, needs validation is not nearly as well defined as the validation of risk instruments. The weights associated with need assessment items are based on supervision time requirements rather than outcomes. Hence, the data base required and the type of analysis undertaken to "validate" a needs scale differ substantially from risk scale validation efforts. However, because these instruments remain an important component of the classification process, a short synopsis of a recent validation effort is presented in the appendix.

D. Glossary of Terms

The following sections discuss study parameters and data analysis. Although these chapters are not overly technical, some explanation of terms used throughout may prove helpful to the reader, particularly those with limited statistical knowledge. Simple, operational definitions of these terms are presented below:

Criterion Variable:	This is simply the outcome measure used in the study. Some studies select a single outcome such as revocation. Others test risk factors against several outcomes: arrests, convictions, revocations. In statistical analyses, the criterion variable is sometimes referred to as the dependent variable.
Independent Variables:	Factors or items used to define risk are called independent variables. These are generally social or criminal history measures which have a potential relationship to the criterion variable.
Item Values:	Each independent variable is categorized or scaled based on its relationship to outcomes. For example, cases with no prior probation experience may, in the aggregate, be less likely to re-offend than those serving their second or third probation. If no significant differences in outcomes are evident between those serving a second or third probation term, the item values for prior probation will be simply "None" and "One or More."

Item Weights:	Weights assigned to each item value reflect its relationship to the criterion variable(s), relative to all others on the scale. A variable that separates outcome groups to a small degree may receive weights of 0 and 1; a factor which separates offender groups to a greater degree receives a higher item weight.
Follow-up Period:	This term refers to the length of time cases are studied to determine outcomes. Standard research practice requires a uniform follow-up period for all cases in the study. Since terms of probation and parole vary significantly, time on supervision is generally not acceptable as a follow-up period.
Base Rates:	The rate at which an observed event occurs within a population is termed the base rate. In validation studies, base rates typically refer to the rate of revocation or re-offending. Base rates vary among subpopulations and it is, therefore, important to analyze these groups independently. Generally speaking, high base rates produce better risk studies.
Multiple Regression Analysis:	A commonly used statistical method in scale construction. It is (most often) a linear technique that attempts to identify the best combination of factors to explain the variance in the dependent or criterion variable.

STUDY PARAMETERS

This section of the report presents parameters for validation studies, ranging from minimal requirements to what is ideal. We begin with identification of the research questions that all analyses should address:

1. How well does the risk instrument currently used separate risk groups based on rates of success/failure?

2. Can the scale's ability to separate risk groups be increased through: a) different value aggregations within risk factors, b) different weights for risk factor values, c) the addition of new variables to the risk scale, d) the deletion of factors currently used, or e) different cut-off scores for risk groups?

3. How does the scale perform for population subgroups including various ethnic groups, female offenders, and special offender groups?

Risk validation studies need not be overly complex if the study's goals are clearly understood. There are, however, a few important research issues confronting these studies including:

- selection of outcome measures (or criterion variables)
- sampling methods and sample size
- follow-up period required
- data base requirements

Each of these issues is addressed below. Statistical technique, as well as interpretation and presentation of results, are presented in the next section.

A. Criterion Variables

The first step in any validation effort is the selection of the criterion or outcome variables. Common measures include arrests, convictions (sometimes broken down into felonies and misdemeanors or assaultive/non-assaultive offenses) and revocations (often delineated by

reason for revocation -- new offense or technical violation of probation or parole). A few studies have also attempted to include violations that did not result in revocation. In Wisconsin, for example, a scaled outcome variable was created, ranking behaviors observed from best (no violations) to worst (new felony convictions and rules violations reported) as follows:

0 = No arrests, rules violations, or convictions
1 = Rules violations only; no revocation
2 = Absconding recorded; no revocation
3 = Arrests and/or convictions recorded; no revocation
4 = Revocation due to rules violations
5 = Revocation due to arrest (in lieu of conviction)
6 = Revocation following new conviction
7 = Revocation with both new conviction(s) and rules violation(s) (including absconding(s))

Few studies have data bases available to support this type of outcome scaling. While the above scale includes all types of supervision behavior, analyses demonstrate that it is highly correlated with simpler outcome measures and, in essence, adds little to a validation effort.

Each type of outcome measure has associated strengths and weaknesses. Arrests and convictions generally represent actual law-violating behavior and the correlation between the two measures is often so high that either can be used as the principle measure of recidivism without affecting the statistical analysis. However, some caution is required. Arrests are only allegations which, in certain circumstances or jurisdictions, may have a limited relationship to actual behavior. Parolees (particularly high-profile offenders) in some areas are routinely arrested, questioned, and released when crimes are reported with no further action taken by the criminal justice system. Convictions can also pose problems as the time required to obtain a conviction can be substantial. Hence, the actual behavior may have occurred within the study period, but the conviction did not. When new convictions are used as the outcome variable, such cases can be misrepresented as "successes." Further complicating the issue is the fact that crime-reporting systems are far from reliable. They depend on local sheriffs, police departments, and court personnel to properly record and enter data. When studies have compared data from a variety

of sources -- state crime reporting systems, the National Crime Information Center, and state correctional systems -- serious inconsistencies and incomplete records have been encountered.

Revocation as an outcome measure can also prove problematic. Revocation is often more representative of the system's response to violations than the frequency or severity of offender misbehavior. Prior studies have clearly demonstrated that substantial variance exists among probation/parole officers and areas of a state in the use of revocation (Clear, Baird, Harris; 1986).

Despite the problems of data reliability, justice system delays, and inconsistent use of revocation, the data available on criminals have generally proven adequate for the task of validation. In most cases, it can be assumed that errors and omissions are random across all risk groups. In other instances, where problems with certain measures are discovered, these measures should be avoided. For example, high rates of arrest with no subsequent action were noted in a recent study of parolees in Tennessee (NCCD, 1990). Since the problem was especially evident in certain offender groups, random distribution of arrest "errors" could not be assumed. Therefore, arrests were dropped as an outcome criterion.

As noted earlier, the correlation between arrests and convictions is often very high -- .8 or above. Thus, either measure can be used without altering the results of the analysis. Use of arrests rather than convictions is generally related to the length of the follow-up period or the base rate of each criterion. When studies are hampered by short follow-up periods -- particularly those of a year or less -- arrests probably represent the best alternative, as many new convictions would not be captured within study time frames (due to delays in court processing). The higher base rate associated with arrests may also prove valuable in establishing statistical relationships, again particularly when the follow-up period is limited.

In the ideal situation, different data sources can be tapped, compared, and merged to present the best overall picture of offender behavior. Arrests, convictions, and revocations can all be used to evaluate risk scale performance. Arrests and convictions should be reported by level of frequency and revocations delineated by type: technical violations and those due to a

new offense. Prior to beginning the analysis of risk, base rates should be established for the entire sample as well as population subgroups to help identify issues that require further study. The following two tables from recent studies are presented to illustrate how base rates can be presented and the degree to which they may differ among subpopulations.

Table 2-A

MICHIGAN YOUTH STUDY
NEW OFFENSES

Technical Violations Recorded	
None	43.2%
One	23.4%
Two	11.8%
Three	8.9%
Four or More	12.3%
Arrests	
None	67.9%
One	21.5%
Two	7.7%
Three or More	2.9%
Felony Arrests	
None	75.2%
One	18.3%
Two or More	6.5%
Adjudications	
None	71.3%
One	21.7%
Two or More	7.1%
Assaults Recorded	
None	88.7%
One or More	11.3%
Out-of-Home Placements	
None	45.8%
One	28.9%
Two	12.6%
Three or More	12.7%
DSS/Private Child Care Institutional Placements	
None	74.4%
One	18.1%
Two or More	7.5%

Table 2-A also illustrates that outcome criteria can and should be different for juvenile offender populations. When constructing risk instruments for juveniles, arrests may, in fact, represent a better indicator of behavior than actual adjudications because the juvenile justice system generally has more latitude in dealing with offenses. Youths arrested for similar behaviors may, therefore, be treated very differently by the system, depending on chronicity, prior attempts by the system to deal with delinquent behavior, etc.

Table 2-B

WISCONSIN COMMUNITY CORRECTIONS STUDY

	Revocation Rate	Arrest Rate
Age:		
25 or under	19.6%	28.4%
26 through 39	17.7%	23.6%
40 or older	12.9%	15.9%
Sex:		
Male	19.5%	26.8%
Female	8.6%	17.3%
Race:		
White	15.7%	24.3%
Black	23.6%	28.1%
Supervision Status:		
Probation	15.4%	23.6%
Parole	28.6%	29.8%

Unfortunately, studies are sometimes constrained by data availability or lack of funds required to conduct record checks on sample cases. In such instances, outcomes may be limited to those recorded in a correctional data base -- principally revocations and returns to prison or probation. Simply comparing revocation rates to initial risk scores can provide an indication of scale efficacy. This requires only that total risk scores and probation/parole outcomes are known. It does not allow for item analysis and frequently is hampered by the

existence of a variable follow-up period (probation/parole terms vary and data on behavior after termination may not be available). Hence, such comparisons should not be confused with a formal validation study, but could well indicate just how crucial it is to revalidate the current system.

B. Follow-up Period

When conducting a validation study, it is important that a standard follow-up period be used for all cases. Some offenders may be on probation or parole for the entire period while others are discharged and spend only part of the follow-up under supervision. The variance in degree of control exerted on cases should be acknowledged, but unless some cases spend minimal time on supervision and/or supervision is particularly intrusive, the affect of the different length of probation or parole terms is probably negligible. In any event, it is more than offset by the value of standard follow-up periods.

In selecting the length of the follow-up period, two issues should be considered. First, the time frame analyzed should be long enough to capture the vast majority of cases that will have new violations reported -- arrests, convictions, revocations. Most research indicates that 18 months is adequate but that 24 to 36 months or longer is ideal. However, the length of the follow-up period should be chosen in context with the need to use cases recently admitted to probation or parole. When changes occur in legislation, policy, or social conditions, offender profiles can change substantially. Hence, studies strive to use the most recent admission or prison release cohort possible and still allow for an adequate follow-up period. For example, study cases admitted to probation during the last six months of 1988 provide a reasonably contemporary sample, but still permit analysis of a 24-month "at risk" period for a study beginning in January 1991. However, if major legislative initiatives were enacted in January 1989 that resulted in a significant shift in probation profiles, shortening the follow-up period to 18 months may produce results more reflective of current conditions.

C. Sampling Parameters

In selecting cases for a validation study, two basic rules apply: (1) The selection should be random (although oversampling -- systematic stratification -- of some offender groups may be desirable) and (2) large samples are superior to smaller samples. Beyond these two generalizations many factors should be considered.

Readers willing to take the time to look into sampling issues will find a confusing array of recommendations. Many statisticians, for example, suggest 100 cases (50 for construction, 50 for validation) for each risk scale factor -- generally about 1,000 cases (Clear, 1988). Alexander and Austin (1991) recommend various sample sizes dependent upon the level of confidence required in the estimates attained in the study. (Slightly fewer than 400 cases will produce an error rate of 5%.) In any case, sample characteristics should be compared to those of the general population to determine if the sample is truly representative. Table 2-C illustrates the type of comparisons done in a recent risk study completed for the Tennessee Parole Board.

Table 2-C

COMPARISON OF POPULATION AND SAMPLE CHARACTERISTICS (TENNESSEE)

Characteristic:	Releases	% Releases	Sample Cases	% Sample
Sex:				
Male	2312	91%	778	92%
Female	230	9%	69	8%
Total	2542	100%	847	100%
Race:				
White	1260	50%	463	56%
Black	1070	42%	333	38%
Other	212	8%	51	6%
Total	2542	100%	847	100%
Type of Supervision Release:				
Parole	1313	52%	437	52%
Safety Valve	984	39%	340	40%
Expired Sentence	245	9%	70	8%
Total	2542	100%	847	100%
Released From:				
Prison	1305	51%	494	58%
Jail	1237	49%	353	42%
Total	2542	100%	847	100%
Instant Offense:				
Sex Offense	148	6%	74	9%
Drug Offense	257	10%	144	17%
All Other	2137	84%	629	74%
Total	2542	100%	847	100%

In many studies, sample size is limited by data availability. Manual file searches can be expensive and time consuming, but studies can rarely avoid manual data collection because few automated data bases provide the level of detail needed to support a comprehensive evaluation of risk assessment. Furthermore, since some proportion of cases in the study will not have spent the entire follow-up period on supervision, it is necessary to go beyond the agency's data system and obtain rap sheets from state and/or federal crime information centers. Because of the time and expense involved, the size of the study sample may have to be curtailed.

If the study objective is simply to ascertain the validity of the existing scale, 350 to 400 cases will provide an adequate sample (as it produces about a 95% confidence interval). If construction of a new improved instrument is a potential goal, the sample should be doubled. Standard scale construction methodology requires dividing the sample into two halves: the first is used to construct the risk scale; the remaining half is used for validation purposes. The use of construction and validation samples allows a scale to be developed on one population and tested on another. The validation sample better indicates how the scale will perform when implemented.

Much larger samples are needed to accommodate questions of validity for subpopulations -- females, ethnic groups, specific offender groups, urban or rural populations, etc. The greater the number of breakdowns contemplated, the larger the sample required. To obtain large enough samples, oversampling of some groups may be essential. For example, females generally comprise only 5% to 15% of an offender population. Hence, to obtain a large enough sample to produce conclusive results, it may be necessary to obtain data on all females admitted during the sample period or even to extend the sample time frame for women.[1] In this instance, even though a 100% sample is used, because the cohort is small, the statistics produced may not prove stable enough to project future results.

Ideally, each subsample analyzed would be comprised of three to four hundred cases. This, however, is not always possible, particularly in smaller jurisdictions. It suffices to say that results from small samples must be interpreted with caution. Obviously, large samples produce greater confidence in study results.

The following table presents an example of oversampling. In this example, the sample size was increased by nearly 800 cases in order to produce results for groups of particular interest. Obviously, the cost of the study will increase proportionately.

---

1 When specific subgroups are oversampled, their numbers should be randomly reduced to representative levels for the general analysis, so that the scale construction is not unduly influenced by any particular group.

Examples of Oversampling

	Random Selection (10% sample)	Over-Sampling	% Sampled	Total Samples
White Males	556	---	---	556
Black Males	610	---	---	610
White Females	31	309	100%	309
Black Females	37	351	100%	351
Sex Offenders	107	294	30%	294
TOTAL	1341			2120

D. Data Base Requirements

Following selection of the outcome criteria, the length of the follow-up period to be analyzed, and sampling parameters for the study, overall data needs can be identified. Again, we will attempt to describe the range of possibilities from minimum requirements to data needed to support a comprehensive evaluation of the classification process. Included in this discussion are methods of data collection and strengths and weaknesses of each approach.

A "bare bones" approach to validation requires limited information. At the most basic level, risk item scores, a few offender characteristics -- sex, age, race -- and an acceptable measure of outcome are all that are required. This low cost option will produce important information and permit reweighting of items, deletion of factors not related to outcomes, and changes in cut-off scores. It will not, however, allow re-aggregation of item values or

consideration of additional risk factors. Nevertheless, given the budget constraints faced by many correctional agencies, and the minimal revisions in risk scales that have resulted from more sophisticated studies, this simpler approach to validation may produce the highest benefit to cost ratio.

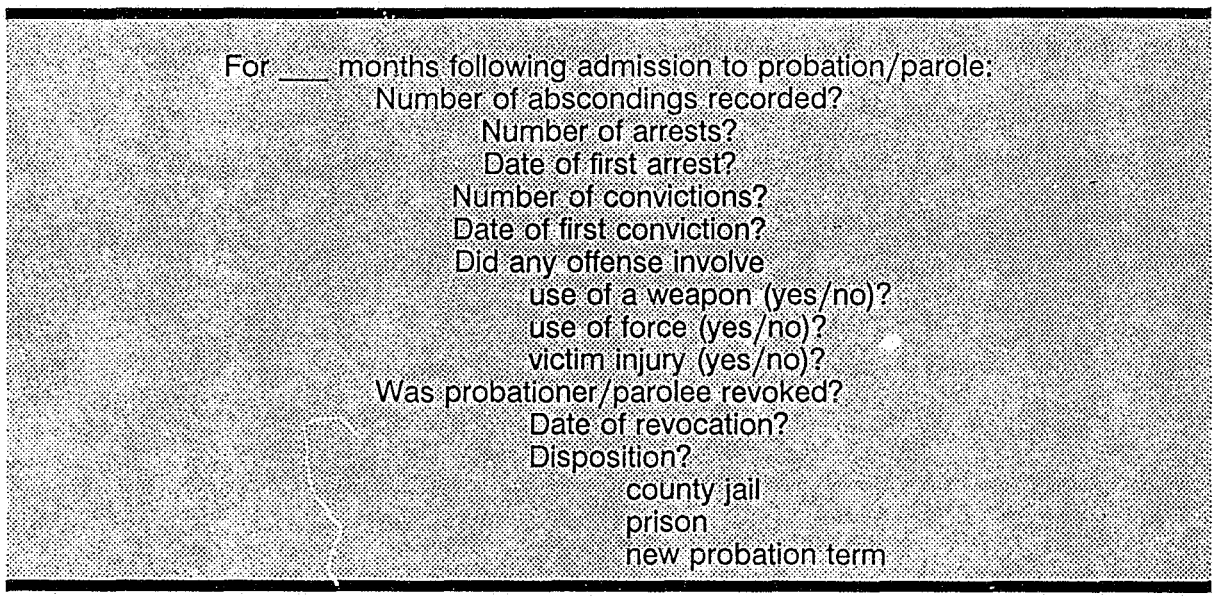
The Kansas Department of Corrections, for example, recently completed a validation study relying totally on data from the Department's information system. The outcome variables used were:

- revocation due to a technical violation
- revocation due to a new offense
- return on a new conviction

The study was completed with no manual data collection or NCIC record checks and, as a result, was done for about $5,000.

More elaborate studies sometimes cost $50,000 or more, depending on (1) the relative ease with which data can be collected and (2) the agency's ability to assist researchers with data collection tasks. A large part of any research effort is purely clerical: assembling, recording, computerizing, and editing data. Distributing these responsibilities among agency staff can reduce the time required to conduct record searches and avoid the cost of paying researchers to perform this function. Furthermore, it may prove far more manageable to have each line worker conduct a limited number of NCIC checks than to submit a request for data on a large number of offenders or have data collectors tie up a few terminals for an extended period of time. In addition, workers are likely to have access to case files allowing them to augment incomplete information obtained from state or national crime information files. Data on offense disposition is frequently missing from automated crime information systems.

When agency staff are surveyed to collect data on sample cases, they should be asked to conduct record checks and review files to answer the following questions:



When staff surveys are used for data collection, a small randomly selected portion of responses should be verified by the research team to ensure that the data obtained are accurate. Ideally, this would be done in a pilot phase to determine if surveys will solicit reliable data. Verification may involve as few as 50 to 100 cases.

For criminal and social history information, many studies rely generally on risk and need forms completed at the time of admission to supervision. Use of classification data presents one drawback. Data values have already been aggregated (e.g., Two or More Felony Offenses = 4; Age at First Conviction of 19 or Less = 4). Without the source data that resulted in these aggregations, researchers cannot test whether different aggregations produce better results. Hence, raw data -- actual number of prior felonies, actual age at first conviction, etc. -- provide a superior data base for risk scale validation.

When agencies have a fairly sophisticated computerized data base available, much of this information is readily accessible. Producing a "tape dump" of all cases admitted during the sample period will allow the researcher to select a sample and obtain available data on each

case. Automated files often contain "face sheet" type of data -- details on offense and offense history as well as demographic information. Identifying elements where substantial data are missing will help determine what needs to be collected manually.

In assembling a data base, the following steps are recommended:

Step 1: Establish an advisory committee of staff and administrators to set goals and objectives for the study, assist researchers in identifying what additional offender characteristics (other than existing risk factors) should be analyzed, select the criterion variable for the study, and recommend data collection methods.

Step 2: Determine what data are available from automated sources and how best to obtain this information. Establish sampling parameters based on goals and objectives of the study. Select sample.

Step 3: Construct data collection forms, instructions, definitions, and protocols.

Step 4: If manual data collection is required, identify the most expeditious and cost-efficient method for collecting reliable information.

Step 5: Establish a short pilot study of the data collection process to determine if problems exist with instructions, definitions, or data availability. Conduct inter-rater reliability checks during the pilot phase to determine if the data forms are completed consistently. Normally this can be accomplished by having all raters independently complete forms on the same cases and comparing the results. The number of cases required for a reliability check ranges from five to ten depending on the number of raters involved. The higher the number of raters, the fewer cases required to produce an adequate number of observations. We recommend the following table as a guide to minimum requirements for a reliability check:

Number of Raters	Minimum Cases	Total Observations
2	10	20
3	10	30
4	8	32
5	8	40
6	6	36
7 or more	5	35 or more

Step 6: Verify information collected on at least 20% of the forms completed during the pilot phase.

Step 7: Finalize forms, instructions, and procedures based on the above studies. Begin data collection process.

Step 8: Establish routine edit procedures for discovering errors and omissions. Monitor return rates to ensure a representative sample. Implement procedures to correct obvious errors and collect missing data.

When manual data collection is required, researchers should anticipate some reduction in the sample size due to missing files or non-compliance problems. When surveys are used, recent return rates in NCCD studies have ranged from about 80% to 97%. If substantial numbers of data collection forms are not returned, it is essential that sample characteristics be compared with general population profiles to ensure that the sample is representative.

In some instances, return rates from a particular geographical area may be far below those of areas of the jurisdiction. If the problem cannot be rectified, the following steps should be taken:

a) Cases submitted by the area with the low return rate should be compared to all cases from that area on critical factors including (at least) age, sex, race, and offense to ensure that cases were not excluded on a systematic basis.

b) A weighting technique should be employed to increase the area subsample to a proportional share of the overall sample.

Putting It All Together

The chart on the following page summarizes the discussion of study parameters, listing minimum requirements for a "bare bones" validation of a risk instrument currently in use compared to what is needed to produce a comprehensive evaluation of an agency's classification process:

Study Parameter	Minimal Requirements	Optimal Recommendations
Criterion variable	Revocations by type	All violations recorded and arrests, convictions (by type), and disposition.
Follow-up period	12 months (18 if convictions are used as criterion)	24 to 36 months
Sample size	350 - 400 cases	5,000 cases
Comments:	Allows only an overall assessment of the instrument's ability to classify cases. Will not support many breakdowns (racial, gender, etc.) or construction of a revised scale.	A large sample permits analysis of all relevant subpopulations: women, ethnic groups, age groups, etc. to identify ethnic, gender, or age bias.
Sample selection methods	Random	Random and systematic stratification. Small groups of interest should be oversampled.
Data base	Risk factors and criterion variable, sex, age, race, offense. Demographic factors and offense can be used to determine if sample is representative	Risk and need data in raw form, outcome variables, demographics, offense data, juvenile records, criminal histories including offenses and dispositions, outcome variables. Time study data.
Cost	$5,000	$25,000 - $100,000 depending on data availability.

ANALYSIS

In the past, most efforts at scale construction employed multivariate statistical techniques -- principally multiple regression analysis. In an earlier section, we noted the problems that are encountered when regression is used to evaluate scale performance. Experience now indicates that multivariate analyses, although helpful, should not be relied upon entirely to construct new instruments. The validation study demonstrates that changes are required. A decade ago, in a study for the National Institute of Corrections, researchers found that no one statistical method produced significantly better results than others in constructing risk instruments (Gottfredson & Gottfredson, 1979). The study assessed the efficiency and accuracy of varying mathematical models of risk prediction, comparing complex procedures with simpler methods. After an exhaustive analysis, the authors concluded that none of the approaches (combining bivariate analyses -- Burgess Scaling, Multiple Regression, or Predictive Attribute analyses) offered advantages over the others.

However, recent NCCD efforts in risk prediction (California, South Carolina, Oregon, Alaska, Illinois) have demonstrated that combining the results of simple bivariate analyses (guided by results of multivariate analyses) to create a scale produce the best results. This is probably due to correctional data base problems, principally that of missing, incomplete, or inaccurate data. The fact that combining highly correlated variables benefits predictive accuracy may also reflect on inconsistencies in our criminal justice system that disrupt "normal" patterns of events. For example, arrests for similar offenses may result in different conviction patterns due to a number of criteria. This is especially true in jurisdictions where judges exercise considerable discretion, and dispositions will be based on such factors as family support and availability, offender needs, program availability, and the different use of available sanctions by individual judges.

The redundancy incorporated in such scales is adjusted for in establishing cut-off scores for each decision point. In addition, systematic addition and deletion of highly correlated items

from the scale and testing each combination against the construction sample illustrates which items add to the scale's ability to identify groups of high-, moderate-, and low-risk offenders. In essence, this technique mirrors the goals of discriminant function analysis without the complications of two or more functions and high tolerance levels for item inclusion.

Maximum separation of groups based on actual rates of re-offending is the goal of classification systems. As noted earlier, this should not be confused with the ability to explain the variance in criminal behavior among sample cases. A scale may explain very little of the variance in outcomes, yet effectively separate groups of offenders and significantly improve correctional decisionmaking.

A. Approach to Validation

A stepwise procedure for conducting a validation study is presented below. Examples of output are presented throughout to help clarify this discussion.

1. **THE EXISTING INSTRUMENT IS TESTED AGAINST THE ENTIRE SAMPLE TO: (A) DETERMINE THE DEGREE TO WHICH OFFENDER GROUPS ARE SEPARATED RELATIVE TO OUTCOMES, (B) DETERMINE IF EXISTING CUT-OFF SCORES ARE THE OPTIMAL THRESHOLDS FOR DEFINING RISK GROUPS, AND (C) ILLUSTRATE HOW SAMPLE CASES ARE DISTRIBUTED AMONG THE EXISTING RISK CLASSIFICATIONS. (THESE FINDINGS SHOULD BE COMPARED TO EARLIER STUDIES, IF ANY WERE CONDUCTED.) TABLES 3-A AND 3-B FROM THE 1989 VALIDATION OF THE WISCONSIN INSTRUMENT ARE USED AS EXAMPLES OF THE ANALYSIS DESCRIBED ABOVE.**

   THESE EXAMPLES SHOW THAT (1) THE SCALE IS STILL EFFECTIVE IN SEPARATING OFFENDER GROUPS BASED ON RATES OF REVOCATION; (2) THE REVOCATION RATE HAS INCREASED OVER THE LAST DECADE IN WISCONSIN; AND (3) OVER TWO OF EVERY FIVE CASES ADMITTED TO PROBATION OR PAROLE WERE RATED HIGH RISK IN 1987.

Table 3-A

RISK DISTRIBUTION USING CURRENT RISK SCALE SCORES[2] (WISCONSIN)

Risk Classification	(Total Score)	N of Cases	Percentage of 1989 Study
Low Risk	(0 - 7)	1363	25.4%
Moderate Risk	(8 - 14)	1643	30.6%
High Risk	(15 - 37)	2365	44.0%
TOTAL		5371	100.0%

Table 3-B

COMPARISON OF OFFENDER REVOCATION RATES BY RISK SCORE FOR THE 1979 AND 1989 STUDIES (WISCONSIN)

Offender Risk Score Range	Offender Revocation Rate: 1979 Study	1989 Study
0 - 3	0.9%	2.0%
4 - 7	2.4%	5.5%
8 - 9	5.6%	7.2%
10 - 11	9.8%	12.8%
12 - 14	12.5%	15.4%
15 - 19	15.6%	22.1%
20 - 24	25.9%	29.4%
25 - 29	37.5%	37.0%
30+	42.5%	44.3%
Summary		
Low Risk (0 - 7)	2.0%	4.4%
Moderate Risk (8 - 14)	9.2%	11.6%
High Risk (15 - 37)	26.0%	31.5%
TOTAL	11.3%	18.5%

[2] These scores omit the 15 points assigned for an assaultive offense committed within the last five years.

2. NEXT, THE EXISTING SCALE SHOULD BE ANALYZED USING RELEVANT SUBPOPULATIONS TO DISCERN ITS ABILITY TO EFFECTIVELY SEPARATE RISK GROUPS WITH EACH SUBGROUPING. AGAIN, DATA FROM THE WISCONSIN STUDY IS USED TO ILLUSTRATE RESULTS OF THIS STEP AND TO INDICATE HOW OUTCOME RATES (IN THIS INSTANCE, REVOCATION) VARY AMONG OFFENDER GROUPS.

Table 3-C

REVOCATION RATES BY CURRENT RISK SCORE INTERVALS FOR POPULATION SUBGROUPINGS (WISCONSIN)

	N of Cases	Current Risk Level[3] Low (0 - 7)	Moderate (8 - 14)	High (15 - 37)
Gender:				
Male	4459	5.3%	12.6%	33.1%
Female	871	2.0%	7.1%	21.4%
Race:				
White	3551	4.1%	10.6%	27.2%
Black	1426	4.9%	14.5%	34.0%
Status:				
Probation	4096	4.6%	11.6%	29.3%
Parole	1275	2.7%	11.8%	34.4%
Age:				
25 or Under	2677	5.9%	10.9%	34.2%
26 - 39	2124	3.6%	13.7%	28.8%
40 and Over	528	3.2%	7.8%	28.5%

THESE DATA ILLUSTRATE THAT WHILE BASE RATES VARY SUBSTANTIALLY AMONG OFFENDER SUBPOPULATIONS, THE INSTRUMENT CURRENTLY USED SEPARATES HIGH, MODERATE, AND LOW RISK GROUPS VERY WELL FOR ALL GROUPS ANALYZED.

[3] Risk scores do not include 15 points for assaultive offenses.

3. IN THE THIRD STEP, ITEM ANALYSIS BEGINS. FOR INFORMATIONAL PURPOSES, THE RELATIVE INFLUENCE OF INDIVIDUAL RISK FACTORS IS DETERMINED THROUGH MEAN PERCENTAGE OF THE AVERAGE TOTAL SCORE EACH FACTOR REPRESENTS AND ITEM CORRELATIONS WITH TOTAL SCORE. TABLE 3-D ILLUSTRATES THIS STEP.

Table 3-D

CURRENT RISK SCALE ITEM ANALYSIS FOR THE 1989 STUDY (WISCONSIN)

Risk Factors	Mean Score	Mean Score as % of Subtotal	Correlation w/Subtotal
Address Changes	1.74	11.8%	.37
Time Employed	1.10	7.5%	.40
Alcohol Use	1.70	11.5%	.44
Drug Use	0.60	4.1%	.45
Attitude	1.88	12.8%	.48
Age at First Conviction	2.39	16.2%	.63
Prior Probations/ Paroles	1.91	13.0%	.78
Prior Revocations	0.95	6.4%	.71
Prior Felonies	1.11	7.5%	.75
Convictions for Burglary, Theft, etc.	1.35	9.2%	.56
SUBTOTAL	14.74	100.0%	---
Assaultive Offense History	6.56	30.7%*	.11
RISK TOTAL	21.30	---	---

* Represents percentage of Risk Total.

4. INDIVIDUAL ITEMS ARE THEN TESTED AGAINST PRINCIPLE OUTCOME OR CRITERION VARIABLES. THIS ANALYSIS WILL INDICATE (A) HOW THE CURRENT DISCRIMINATORY POWER OF EACH ITEM RELATES TO ITEM WEIGHTS, AND (B) WHAT CHANGES, INCLUDING ITEM DELETION, RE-AGGREGATION OF ITEM VALUES, OR REVISIONS IN ITEM WEIGHTS MAY IMPROVE SCALE PERFORMANCE. DATA FROM AN ACTUAL STUDY ARE PRESENTED IN TABLE 3-E.

Table 3-E

ITEM ANALYSIS

Risk Item	Current Weight	Percent of Cases	Percent Revoked	Percent Arrested
Number of Address Changes				
None	0	32.3%	14.1%	24%
One	2	28.8%	17.9%	31%
Two +	3	38.9%	27.4%	36%
Percent of Time Employed				
60% +	0	39.2%	13.8%	25%
40-59%	1	14.3%	16.8%	32%
39% -	2	46.5%	26.5%	36%
Alcohol Problems				
None	0	43.4%	14.0%	24%
Occasional	2	26.9%	20.2%	36%
Frequent	4	29.7%	30.2%	38%
Other Drug Problems				
None	0	59.7%	14.2%	26%
Occasional	1	21.0%	24.4%	39%
Frequent	2	19.3%	34.4%	39%
Attitude				
Motivated, receptive	0	50.8%	13.5%	25%
Dependent, unwilling	3	31.3%	23.2%	37%
Negative, rationalizes	5	18.0%	33.7%	37%
Age at First Conviction				
24 or older	0	30.5%	8.7%	14%
20-23	2	20.3%	17.0%	30%
19 or younger	4	49.2%	28.7%	42%
Prior Probations/Paroles				
None	0	53.7%	10.4%	20%
One +	4	46.3%	31.1%	43%
Prior Revocations				
None	0	77.6%	14.2%	25%
One +	4	22.4%	40.0%	51%
Prior Felony Convictions				
None	0	67.4%	14.0%	23%
One	2	12.4%	20.3%	39%
Two +	4	19.5%	39.7%	49%
Convictions for:				
Neither a or b	0	46.1%	12.2%	21%
a) Burglary, Theft, Auto Theft, Robbery	2	40.8%	27.2%	41%
b) Worthless Checks, Forgery	3	6.8%	17.4%	31%
c) Both a and b	5	6.4%	35.1%	44%

THE ABOVE ANALYSIS INDICATES THAT SCALE ITEMS HAVE RETAINED THEIR DISCRIMINATORY POWER BUT THAT CHANGE IN ITEM WEIGHTS MAY FURTHER IMPROVE THE SCALE'S PERFORMANCE. DIFFERENCES NOTED BETWEEN ARREST AND REVOCATION RATES ARE ALSO INTERESTING (NOTE THE FIGURES FOR SUBSTANCE ABUSE ITEMS AND ATTITUDE). IN GENERAL, CASES WITH HIGHER RATINGS ARE MORE LIKELY TO BE REVOKED DESPITE SIMILAR ARREST FIGURES. (THESE CASES MAY HAVE BEEN ARRESTED FOR MORE SERIOUS CRIMES OR EXPERIENCED MORE FREQUENT ARRESTS. IF NOT, THESE DATA MAY INDICATE VARIANCE IN THE USE OF REVOCATION -- LESS TOLERANCE OF BEHAVIORS WITHIN SPECIFIC OFFENDER PROFILES -- AND NEED TO BE ADDRESSED BY MANAGEMENT.)

5. IF NO ADDITIONAL DATA ARE AVAILABLE THAT COULD POTENTIALLY BE USED TO REVISE THE RISK INSTRUMENT, REVISIONS ARE MADE BASED ON ABOVE ANALYSES AND THE "NEW" SCALE TESTED AGAINST CRITERION VARIABLES. IMPROVEMENT, IF ANY, IN THE SCALE'S ABILITY TO DISCRIMINATE BETWEEN RISK GROUPS IS IDENTIFIED AS ARE CHANGES IN CUT-OFF SCORES AND DISTRIBUTION OF OFFENDERS. THE LATTER STATISTIC IS EXTREMELY IMPORTANT IF THE AGENCY USES A WORKLOAD-BASED STAFF ALLOCATION SYSTEM. UNLESS ADDITIONAL DATA CAN BE ADDED TO THE ANALYSIS, PROCEED TO STEP 14.

BASED ON THE RESULTS PRESENTED IN TABLE 3-E, THE FOLLOWING RE-WEIGHTING OF ITEMS WAS DONE TO BETTER REFLECT CURRENT RELATIONSHIPS BETWEEN EACH FACTOR AND PROBATION/PAROLE OUTCOMES:

	Previous Weight	Revised Weight
Address Changes		
None	0	0
One	2	1
Two+	3	2
Percent Time Employed		
60%+	0	0
40% - 59%	1	1
39%-	2	2
Alcohol Problems		
None	0	0
Occasional	2	1
Frequent	4	2
Other Drug Problems		
None	0	0
Occasional	1	2
Frequent	2	3
Attitude		
Motivated	0	0
Dependent	3	2
Negative	5	3
Convictions for		
Neither a or b	0	0
a) Burglary, Theft, Auto Theft, Robbery	2	2
b) Worthless Checks, Forgery	3	1
c) Both a and b	5	3

IF ADDITIONAL DATA ELEMENTS ARE AVAILABLE, THESE ITEMS ARE TESTED AGAINST OUTCOMES. AT THIS POINT, THE VALIDATION EFFORT IS TRANSFORMED INTO A NEW CONSTRUCTION/VALIDATION STUDY TO DETERMINE IF A BETTER INSTRUMENT CAN BE DEVELOPED.

6. THE SAMPLE IS DIVIDED INTO TWO EQUAL GROUPS: THE FIRST TO BE USED TO CONSTRUCT A SCALE, THE SECOND USED FOR VALIDATION PURPOSES. THE USE OF CONSTRUCTION AND VALIDATION SAMPLES ALLOWS A SCALE TO BE DEVELOPED ON ONE POPULATION AND TESTED ON ANOTHER. IN GENERAL, SCALES BEST "FIT" THE POPULATION USED FOR DEVELOPMENT. VALIDATING THE SCALE ON A SEPARATE POPULATION BETTER INDICATES HOW A RISK ASSESSMENT INSTRUMENT WILL PERFORM WHEN ACTUALLY IMPLEMENTED. THE AMOUNT OF PREDICTIVE POWER LOST FROM CONSTRUCTION TO VALIDATION SAMPLES IS TERMED "SHRINKAGE." SOME SHRINKAGE IS NORMAL AND FULLY EXPECTED; EXCESSIVE SHRINKAGE INVALIDATES THE SCALE. NO RULE ON ALLOWABLE SHRINKAGE IS APPLICABLE TO ALL SITUATIONS; EACH ANALYSIS MUST BE VIEWED IN THE CONTEXT OF THE BASE RATE AND OUTCOME DEFINITIONS.

7. SIMPLE CORRELATIONS ARE DEVELOPED BETWEEN EACH BACKGROUND FACTOR COLLECTED AND MEASURES OF OUTCOME. ITEMS WITH SIGNIFICANT CORRELATIONS (.05 LEVEL) WITH ANY OF THE OUTCOME MEASURES ARE SELECTED FOR FURTHER ANALYSIS.

8. MULTIPLE LINEAR REGRESSION ANALYSIS IS CONDUCTED TO HELP GUIDE SELECTION OF THE BEST COMBINATION OF PREDICTIVE ITEMS. THIS ANALYSIS PROVIDES SOME INSIGHTS AS TO WHICH ITEMS SHOULD RECEIVE PRIMARY CONSIDERATION FOR INCLUSION (BASED ON LOW COLINEARITY). HOWEVER, ADDITIONAL VARIABLES ARE INCLUDED IN SUBSEQUENT STEPS.

9. CROSSTABULATIONS (WITH A NUMBER OF ASSOCIATED STATISTICS SUCH AS CHI SQUARES AND CORRELATIONS) ARE COMPLETED TO FURTHER DETERMINE RELATIONSHIPS BETWEEN OUTCOMES AND ALL POTENTIAL SCALE ITEMS. THESE ANALYSES HELP TO DETERMINE 1) HOW VALUES OF EACH INDEPENDENT FACTOR COULD BEST BE COMBINED TO MAXIMIZE THE VARIABLE'S RELATIONSHIP TO THE VARIOUS OUTCOME MEASURES, AND 2) HOW OUTCOME VALUES SHOULD BE COMBINED (E.G., THREE OR MORE ARRESTS AS A SINGLE OUTCOME VALUE).

10. VARIABLES ARE RE-CODED BASED ON THE ABOVE ANALYSIS, AND THE CROSSTABULATIONS, CHI SQUARES, AND CORRELATIONS ARE REPEATED. ITEM WEIGHTS ARE SELECTED BASED ON THE ABILITY OF EACH FACTOR TO SEPARATE OFFENDER GROUPS WITH DIFFERENT RATES OF SUCCESS/FAILURE REPORTED DURING THE FOLLOW-UP PERIOD.

11. ITEMS ARE SELECTED FOR SCALE INCLUSION BASED ON THE RESULTS OF ALL THE ANALYSES CONDUCTED ABOVE.

12. THE NEWLY DEVELOPED SCALE IS CROSSTABULATED WITH OUTCOMES TO DETERMINE OVERALL DISCRIMINATORY CAPABILITIES AND OPTIMAL CUT-OFF POINTS FOR EACH IDENTIFIED LEVEL OF RISK. ITEMS ARE ADDED AND DELETED FROM THE SCALE AND THESE CROSSTABULATIONS REPEATED TO TEST VARIOUS COMBINATIONS OF FACTORS.

13. THE BEST COMBINATION OF FACTORS IS SELECTED AND THE SCALE IS FINALIZED.

14. THE SCALE IS TESTED AGAINST THE VALIDATION SAMPLE TO DETERMINE THE DEGREE OF SHRINKAGE. IF SHRINKAGE IS WITHIN ACCEPTABLE LIMITS, IT IS COMPARED TO THE EXISTING INSTRUMENT TO DETERMINE THE ADDED LEVEL OF DISCRIMINATION ATTAINED (IF ANY). TABLES 3-F AND 3-G PRESENT THESE COMPARISONS FOR A 1991 STUDY OF TENNESSEE PAROLEES AND A 1990 STUDY OF IOWA PROBATIONERS. THESE EXAMPLES ARE USED BECAUSE SUBSTANTIAL IMPROVEMENTS TO EXISTING SYSTEMS ARE EVIDENT.

Table 3-F

COMPARISON OF CONVICTION RATES
CURRENT AND PROPOSED SCALES
(TENNESSEE)

Current Scale	Cases	% Distribution	New Conviction %
Minimum	286	(46.4%)	35.7%
Medium	175	(28.4%)	40.0%
Maximum	155	(25.2%)	47.2%
TOTAL	**616**	**(100.0%)**	**Base 38.5%**
Proposed Scale			
Low	164	(23.2%)	15.2%
Low Moderate	158	(22.4%)	28.5%
Moderate	162	(22.9%)	46.3%
High	170	(24.1%)	54.7%
Very High	52	(7.4%)	65.4%
TOTAL	**706**	**(100.0%)**	**Base 38.5%**

Table 3-G

COMPARISON OF CONVICTION RATES*
CURRENT AND PROPOSED SCALES
(IOWA)

Risk Classification	Conviction Rates		% Distribution	
	Proposed	Current	Proposed	Current
Administrative	7.7%	9.3%	13.1%	5.8%
Minimum	21.9%	19.1%	36.9%	16.3%
Normal	36.9%	28.1%	36.2%	45.0%
Intensive	52.5%	41.0%	13.8%	32.8%
TOTAL	**Base 29.7%**		**100.0%**	

*New conviction of any type within 24 months of the initial classification.

15. THE SCALE IS THEN TESTED AGAINST ALL RELEVANT SUBSAMPLES: BLACKS, WOMEN, PAROLEES (IF APPROPRIATE), AND REGIONAL BREAKDOWNS IN THE SAMPLE TO DETERMINE IF THE SCALE DEMONSTRATES ANY RACIAL OR GENDER BIAS. RESULTS OF THESE ANALYSES FROM A VALIDATION STUDY CONDUCTED ARE PRESENTED BELOW:

Table 3-H

OUTCOMES BY RISK LEVEL AND RACE (WISCONSIN)

Risk Levels (Total Score)	N	% of Cases	Revocation Rate	Conviction Rate	Arrest Rate
Low (0 - 5)					
Whites	948	27%	3.6%	8.2%	11.2%
Blacks	241	17%	5.4%	9.1%	14.1%
Medium (6 - 12)					
Whites	1301	37%	10.2%	16.6%	22.3%
Blacks	415	29%	13.3%	13.3%	21.4%
High (13 - 30)					
Whites	1302	37%	30.0%	27.3%	35.8%
Blacks	770	54%	34.9%	21.4%	36.0%

Figure 3-A

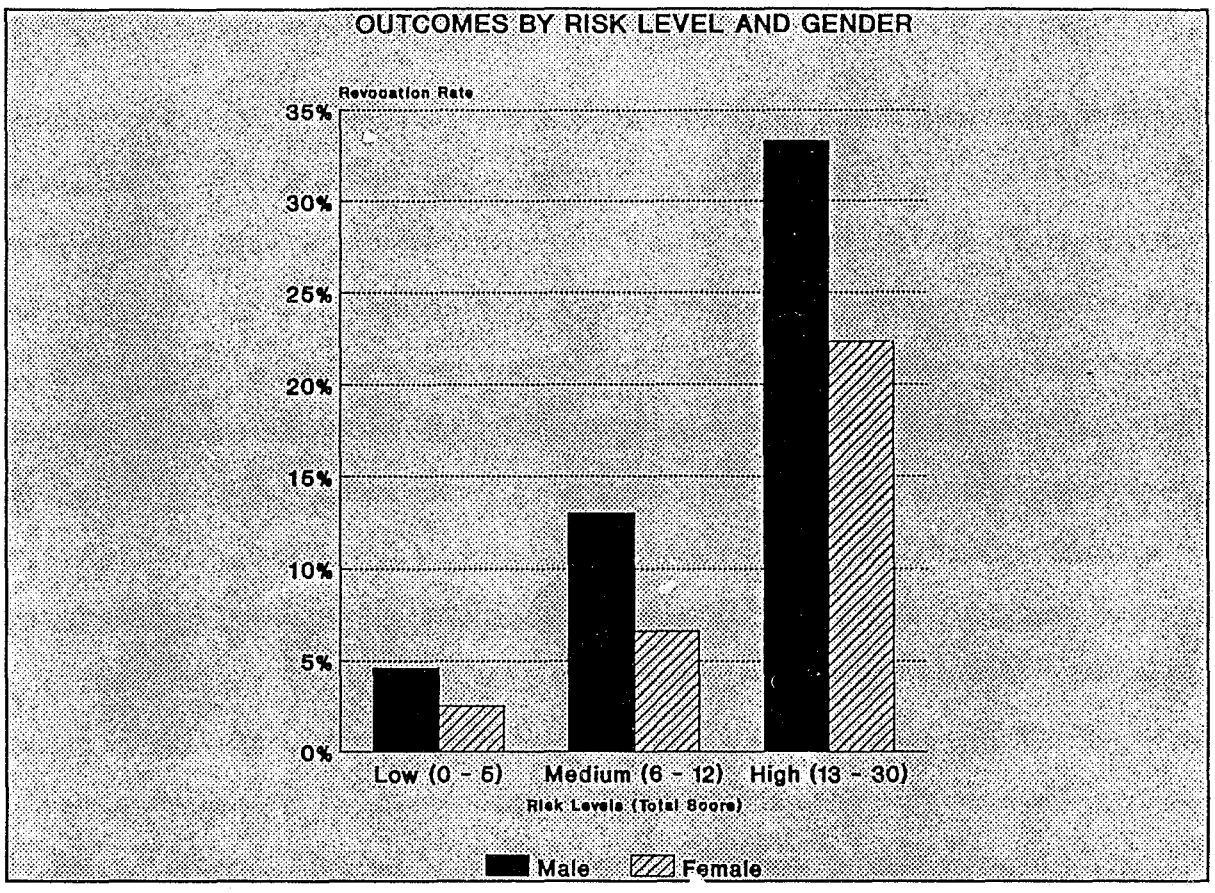
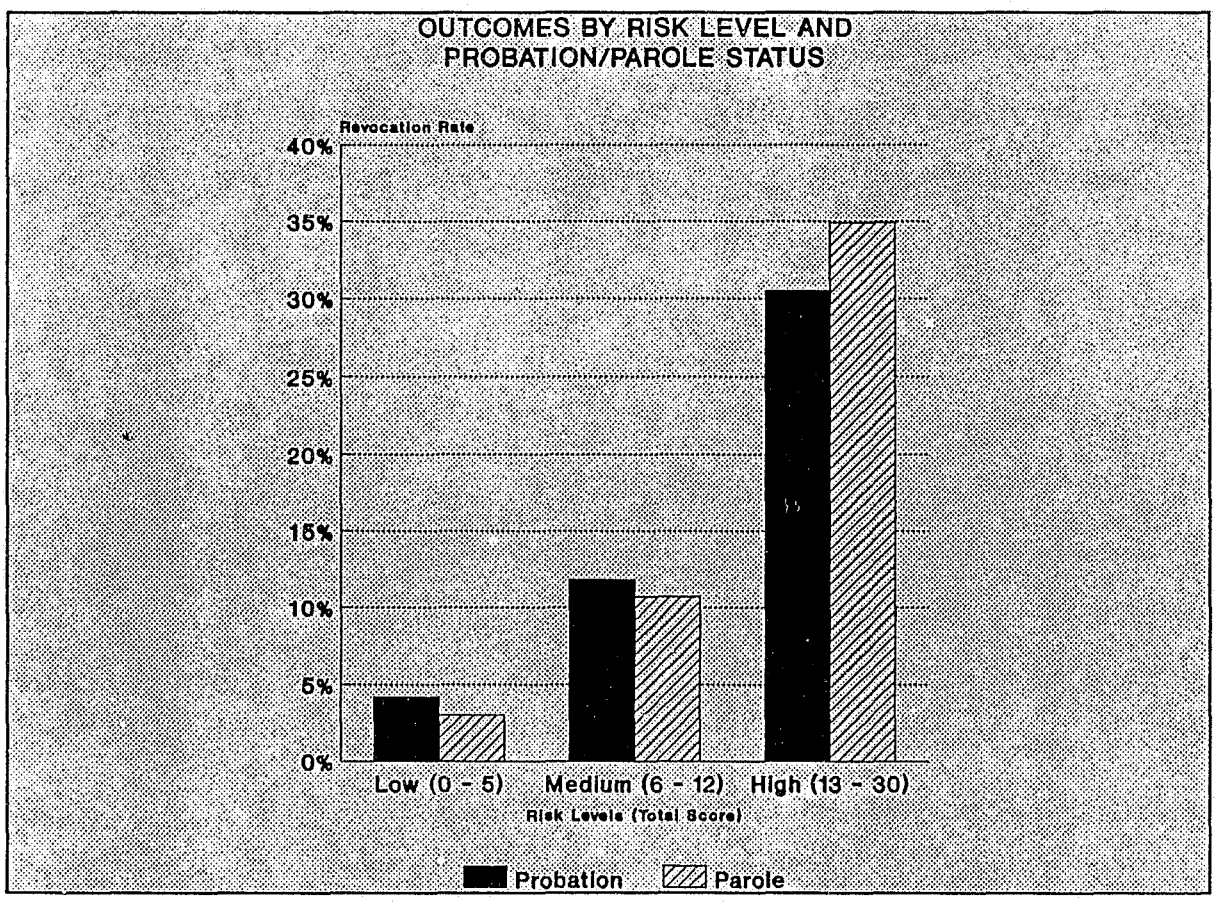


Figure 3-B



B. The Frequency of Validation Efforts

In an attempt to determine how frequently revalidation studies should be conducted, NCCD recently developed risk profiles of cases at two-year intervals from 1980 through 1988, using the automated data base of a midwestern state. Two thousand admissions from each year were randomly selected and comparisons of total risk scores and the scores for individual items were developed. The goals were to ascertain how the offender population had changed over time and how rapidly these changes occurred. As noted earlier, the 1980s were a period of rather dramatic change in both social and sentencing policy. If offender profiles remained relatively stable over such a volatile period in corrections, it would provide some evidence that frequent validation efforts are unnecessary. Major changes in profiles, however, would support the need for more frequent risk studies.

The results of this analysis are presented in the following two tables (3-I and 3-J). The first presentation shows a gradual decline in the lowest risk groups and an increase in the number of cases scoring 25 or more. (The 1988 data, however, may signal a reversal in the latter trend as percentages at the two highest levels declined slightly from 1986 levels.) These data also demonstrated that shifts in risk can be rather abrupt. From 1984 to 1986, the proportion of cases with scores of 25 or higher increased from 11.9% to 17.9%.

Table 3-I

RISK SCORES BY YEAR

Risk Score Ranges	1980	1982	1984	1986	1988
0 - 3	9.4%	11.5%	11.6%	9.5%	8.1%
4 - 7	21.3%	20.5%	18.3%	18.2%	17.3%
8 - 10	16.1%	17.0%	15.4%	14.4%	16.4%
11 - 14	15.9%	16.9%	15.2%	17.3%	15.5%
15 - 19	15.2%	12.4%	14.5%	12.4%	15.6%
20 - 24	11.5%	10.3%	13.1%	10.3%	11.4%
25 - 29	6.9%	7.7%	7.6%	10.5%	9.6%
30 - 37	3.7%	3.6%	4.3%	7.4%	6.1%

Most notable is the increase in cases with serious alcohol and drug abuse problems. The proportion of cases with frequent abuse of drugs reported more than doubled from 1980 to 1988; clients who frequently abuse alcohol increased by 74% over the same period.

The second data presentation reviews how cases scored on individual risk items by year. This analysis identifies what factors are responsible for changes in overall risk profiles.

Table 3-J

RISK SCALE ITEMS BY YEAR

Risk Score Ranges	1980	1982	1984	1986	1988
Address Changes					
None	41.6%	43.0%	41.7%	36.5%	35.5%
One	28.6%	29.1%	27.8%	28.5%	29.4%
Two +	28.3%	27.9%	30.5%	34.9%	35.1%
Percent Time Employed					
60% +	45.0%	42.3%	42.6%	42.0%	42.9%
40% - 59%	18.8%	15.1%	12.3%	14.4%	17.5%
0% - 39%	36.2%	42.6%	43.1%	43.6%	39.5%
Alcohol Problems					
None	49.3%	48.9%	47.8%	42.4%	38.1%
Occasional	32.5%	32.3%	32.3%	27.7%	30.0%
Frequent	18.2%	18.8%	19.8%	29.9%	31.8%
Other Drug Problems					
None	64.8%	67.8%	68.2%	62.4%	56.1%
Occasional	25.3%	21.5%	20.4%	19.7%	22.3%
Frequent	9.9%	10.6%	11.4%	17.8%	21.6%
Attitude					
Motivated, Receptive	56.1%	60.3%	56.8%	50.3%	51.6%
Dependent, Unwilling	29.4%	26.5%	28.5%	30.6%	30.1%
Negative, Rationalizes	14.5%	13.2%	14.6%	19.1%	18.3%
Age at First Conviction					
24 or older	23.2%	30.6%	32.6%	30.5%	33.3%
20 - 23	23.6%	21.6%	18.9%	20.5%	20.6%
19 or younger	53.3%	47.9%	48.5%	49.0%	46.1%
Prior Probations/Paroles					
None	57.6%	59.1%	55.0%	55.1%	55.5%
One +	42.4%	40.9%	45.0%	44.9%	44.5%
Prior Revocations					
None	83.8%	84.1%	80.7%	77.4%	80.4%
One +	16.2%	15.9%	19.3%	22.6%	19.6%
Prior Felony Convictions					
None	72.1%	71.9%	69.0%	69.1%	70.3%
One	13.1%	12.7%	13.8%	12.5%	12.1%
Two +	14.8%	15.5%	17.2%	18.4%	17.5%
Convictions for:					
Neither A nor B	48.9%	47.9%	48.0%	49.1%	46.6%
a) Burglary, Theft, Robbery	35.5%	38.0%	36.6%	38.2%	37.5%
b) Worthless checks/forgery	7.4%	6.9%	7.1%	5.8%	7.3%
Both A and B	8.2%	7.2%	8.3%	6.9%	8.6%
Assaultive Offense within Five Years					
No	72.9%	71.6%	65.0%	57.8%	59.3%
Yes	27.1%	28.4%	35.0%	42.2%	40.0%

Employment problems appeared to peak in the mid-eighties and the percentage of cases employed less than 40% of the prior 12 months decreased from 1984 to 1988. At the same time, another stability factor -- number of address changes -- showed that the offender population was involved in significantly more changes of residence in 1988 than in 1980. This may be attributed to increases in drug and alcohol abuse or homelessness.

On criminal history items, it appears that the 1988 population became involved in offenses at a later age than 1980 admissions, but had a more involved history as adults, as evidenced by a greater proportion of cases with prior probations/paroles and prior revocations. The percentage with prior felony convictions, however, has changed very little over the decade.

Though not a risk item, prior assaultive behavior can have significant impact on levels of supervision assigned in jurisdictions that, by policy, assign all assaultive cases to the maximum level of supervision. The proportion of cases with assaultive offense histories rose from 27% in 1980 to about 41% in 1988. It is important that this is evaluated to determine if it is the result of a change in the definition of assaultive behavior or a real change in the offender population.

In sum, these data indicate that:

1. A significantly higher proportion of the offender population has a history of serious drug and/or alcohol abuse in 1988 than in 1980.

2. The proportion of cases with prior periods of supervision and prior revocations increased slightly from 1980 to 1988; the percentage of cases with prior felony convictions remained relatively stable.

3. Serious employment problems peaked in the mid-eighties and, more recently, began to decline. The proportion of admissions employed less than 40% of the time declined from 1986 to 1988.

4. The percentage of cases with prior assaultive offenses increased substantially from 1980 to 1988. The growth rate seems somewhat inconsistent with a lack of change in the proportion of cases with prior felony convictions. Therefore, it may be attributable to a change in definition as the criminal justice system has become more sensitive to domestic violence, drunk driving, and other social problems.

This study, and others conducted throughout the country (Illinois, Colorado, Iowa, Kansas, etc.), generally demonstrate that risk scales continue to discriminate well over time;

changes in offender populations, even when rather substantial, have not severely affected these scales' ability to accurately classify offenders into high-, moderate-, and low-risk groups. Still, at least minor revisions to scales nearly always result from revalidation studies which increase the discriminatory power of these scales and, in some instances, major improvements are possible.

USING RESULTS OF VALIDATION STUDIES

Revalidation efforts frequently identify some changes that can be made in scales to improve their ability to accurately classify cases. More often than not, however, the degree of improvement is relatively minor, leaving jurisdictions with the dilemma of whether small enhancements to risk classification are worth the time and expense involved in revising current operations. Fortunately, the decision rarely hinges on this factor alone. The implications of scale revisions when viewed in the context of total agency operations may eclipse the impact of a small improvement in discriminatory power. Even relatively minor changes in scale design, for example, can result in substantial shifts in the proportion of cases classified at each risk level. Such shifts can be beneficial or disruptive depending on their direction, current agency staffing levels, and anticipated resource issues. Furthermore, when risk scales are used by parole boards to assist with release decisions, significant issues regarding current practices often emerge. All of the above issues fall under the rubric of policy issues and deserve careful consideration before changes in risk assessment procedures are introduced.

A. Distribution of Cases

Revalidation efforts may suggest significant changes in the distribution of cases among risk levels, or even identify additional classification levels based on better assessment of proclivities for continued criminal behavior. For example, rather dramatic shifts were produced in a recent Iowa study (see Table 3-G). The proportion of cases sentenced for operating a motor vehicle while intoxicated (OWIs) on probation caseloads had skyrocketed in recent years significantly altering the Iowa risk profile. A validation study conducted in 1990 demonstrated that the number of cases classified as "administrative" or "minimum" could be increased significantly while, at the same time, the rate of recidivism for these groups would be lower than for cases classified administrative and minimum by the existing scale. Fewer cases were

classified to the higher supervision categories, but had a much higher rate of new convictions. In essence, classification decisions could be improved substantially.

Adoption of the new scale in Iowa could significantly reduce staffing requirements for probation. Unless supervision standards are revised, fewer staff will be required. However, because the new system will be far more selective in placing offenders in maximum supervision and those placed in the highest level have historically recidivated at very high levels, such data could be used to justify increased supervision requirements for this group. Such a policy could increase community protection and offset the reduction in staffing caused by greater use of lower supervision levels.

The Iowa results are atypical in that this degree of improvement in classification accuracy coupled with substantial changes in distribution of cases is rare. However, less dramatic results can still produce a need to review supervision standards and/or staffing requirements. Consider, for example, policy implication from a 1987 study of Illinois probationers:

> Because recommended changes in the risk scale are fairly minimal (one item is eliminated and one is added), the cost of implementing our recommendations should also be minimal. The added item, Age at Admission, is easy to score and requires no special instructions. The remaining changes actually simplify the scoring of existing items and should require little staff training. Therefore, the basic expense involved in changing the risk forms is in printing costs.
>
> Of far greater concern is the impact of change on the distribution of workload. Although our analysis identified four distinct risk levels, we recommend that Illinois continue to use three levels of supervision. The three-level supervision system could be maintained by collapsing the two middle risk groups into a single moderate risk category. It is also recommended that Illinois continue the practice of placing all probationers convicted of an assaultive offense in maximum supervision. The combined effects of these recommendations are:

1. The percentage of cases assigned to each supervision level at initial classification would change as follows[4]:

	Maximum	Medium	Minimum
Current System	34.2%	33.4%	32.4%
Proposed System	25.4%	52.1%	22.5%

In effect, the proposed system places fewer individuals at both the minimum and maximum level. The new distribution more resembles a bell shaped curve and better separates cases based on failure rates.

2. The average workload represented by 100 cases would change very little. Using the current workload values of 3 hours per month for maximum cases, 1.5 hours for medium supervision, and .5 hours for minimum cases results in the following workload totals:

	Current System	Proposed System
Maximum Cases	102.6 hours	76.2 hours
Medium Cases	50.1 hours	78.2 hours
Minimum Cases	16.2 hours	11.3 hours
Workload Total	168.9 hours	165.7 hours

Using the time frames generated by recent time study results shows even less difference in 1.9 hours less time required per 100 cases.

The above figures are based on changes to initial risk scale only. Obviously, the reclassification scale should be revised in a corresponding manner. This results in a change of weights assigned to 10 of the 12 scale items. Using the same cut-off points recommended for the initial risk scale produced the following results at reclassification:

---

4 The following figures assume that the sample used in this study is representative of Illinois cases.

NOTE: This discussion is taken directly from the 1987 Validation Study conducted by NCCD for the Administrative Office of the Illinois Courts.

Table 34

Reclassification Scores by County Groups

	Total Sample	Cook County	Large Counties	Small Counties
0 - 4	46.5%	59.0%	34.9%	43.3%
5 - 10	42.2%	35.9%	50.9%	40.9%
11 +	11.3%	5.2%	14.2%	15.8%

Assuming an 18-month average length of stay on probation, the combined effects of initial and reclassification scores on sample cases result in the following breakdown:

Maximum Supervision	16%
Medium Supervision	46%
Minimum Supervision	38%

The above figures are based on sample cases and, therefore, do not necessarily represent the breakdown of supervision levels that would be attained statewide if the revisions were implemented. The actual proportion of cases in each county group may be substantially different than the proportion of cases from each group in the study sample. Furthermore, the continued use of the needs scale, with existing cut-off points, would place about 3% of the minimum risk cases in medium supervision and 1% of moderate risk cases in maximum supervision.

Lowering cut-off scores on the needs assessment instrument to 10 for medium supervision and 20 for maximum supervision would give needs a larger role in classification. It may also better represent the point at which time requirements increase significantly (Wisconsin, 1978). This revision would move 5.4% of minimum risk cases into medium supervision and 6.8% of minimum and moderate risk cases into maximum supervision.

Conclusion

The existing Illinois Risk Assessment Scale is a valid indicator of risk. It also meets acceptable standards of utility and equity. Reliability problems are, however, evident. Although these could not be specifically identified within the parameters of the study, it appears that cases in Cook County are scored differently on several of the more subjective items than are cases from other Illinois counties.

The revisions recommended would strengthen the system somewhat. While from an outside perspective, it seems these changes could be rather easily implemented, the AOIC will need to weigh the benefits of change against the costs involved (NCCD, 1987).

B. Equity Issues

Studies that analyze the relationship between risk assessment, race, and gender frequently encounter issues of different base rates for different offender populations. Most recent studies have demonstrated that risk instruments effectively separate offenders by risk level within each group. The same studies, however, have found that the terms "high, moderate, and low risk" translate into different rates of recidivism when comparisons are made between groups. Consider, for example, revocation rates reported for Blacks, Whites, Males, Females, Probationers, and Parolees from a 1989 revalidation conducted for Wisconsin.

REVOCATION RATE BY CASE TYPE (WISCONSIN)

	N	High Risk	Moderate Risk	Low Risk
Blacks	1426	34.9%	13.3%	5.4%
Whites	3551	30.0%	10.2%	3.6%
Males	4459	33.3%	13.0%	4.6%
Females	871	22.3%	6.6%	2.5%
Probation	4096	30.5%	11.8%	4.2%
Parole	1275	34.9%	10.7%	3.0%

The above data indicate that while failure rates vary to a degree, no crossover exists between rates of revocation and risk levels. Hence, the agency should have no difficulty justifying existing classification and supervision policies. (Arrest and conviction rates showed even less variance than revocation rates among subgroups). The situation would be different, however, if moderate risk females recidivated at or below the rate of recidivism recorded for low risk males. (In fact, combining the low and moderate risk groups results in a revocation rate of 4.7% -- or about that recorded for low-risk males.) In this case, the following issue emerges: would the agency be correct in lowering supervision requirements for females scoring at the moderate risk range (should they be supervised like low risk males)? Implementation of such

a policy would mean that some males and females with identical risk scores would be treated differently by the correctional system.

Risk scores have meaning only in that they reflect group probabilities of success/failure. Therefore, the above question could be juxtaposed to read, "Is it correct to hold groups with different rates of failure to the same supervision requirements?" In our opinion, the answer is "no." If experience demonstrates that "moderate-risk" females behave like low-risk males, equal treatment should be required. In essence, in classification, equity issues should relate to base rates, not risk scores.

Equity considerations must, of course, be viewed in the context of overall agency operations. It would create chaos to attempt to implement different cut-off points or supervision standards for many different offender groups. Therefore, changes must be made judiciously -- only in instances where serious breaches in equitable treatment of offenders are evident.

Obviously, several issues other than increased effectiveness of the classification process must be considered when implementing changes in risk assessment. The integration of agency policy, budgeting, resource deployment, information systems, and classification dictate the need to view changes recommended in the context of total agency operations.

C. Where to Go for Assistance

Generally, sources of assistance are local university staff and criminal justice research groups. The National Institute of Corrections maintains lists of consultants with expertise in areas of classification and research and the National Information Center can provide copies of concluded studies for review. The National Council on Crime and Delinquency, Rutgers University, and the University of Cincinnati have all conducted several risk construction/validation studies in recent years.

APPENDIX

VALIDATION OF THE WISCONSIN NEEDS ASSESSMENT SCALE

Two separate data sources are needed to validate a needs scale. These include: a self-reported client-based time study, in which probation or parole officers record all time devoted to sample cases over a defined time frame; and the most recent needs assessment scale completed for each case in the time study.

As explained earlier, the weights assigned to need scale items theoretically represent the amount of time required to supervise a case with that client problem. Thus, the higher the weight, the more time consuming the problem. The remainder of this section of the report uses a recent study of needs assessment conducted for the Wisconsin Division of Community Corrections to illustrate how such studies are completed.

A. THE 1979 NEEDS STUDY

Earlier tests of the Wisconsin scale have demonstrated a relatively strong relationship between the total needs scores and time devoted to cases, even within a specific supervision level where contact requirements were identical. Summary results of a 1979 time study are presented in Table A1.

Table A1

RELATIONSHIP OF NEEDS SCORES TO SUPERVISION TIME IN MINUTES
1979 TIME STUDY

Needs Assessment Score	Low Supervision	Medium Supervision	Maximum Supervision	Average Time
	(Average Minutes Per Client Per Month)			
9 or Less	40.0	61.9	92.4	47.7
10 - 14	45.3	90.6	105.3	79.7
15 - 19	NA	69.7	116.9	86.8
20 - 24	NA	95.4	184.3	142.0
25 - 29	NA	104.3	180.9	160.2
30 or More	NA	107.2	196.7	185.3

B. THE 1989 NEEDS VALIDATION STUDY

A time study conducted in the Spring of 1989 provided the data base for the new validation study. Since needs or problems seldom exist in isolation, it is impossible to determine precisely how a particular offender problem or need influences supervision time requirements. However, with a large data set, time can be related to case needs and reasonable estimates established. The goal is to establish a hierarchy, assigning the highest weight to the most time-consuming needs.

Our analysis proceeded with the following steps:

1. Correlations were developed between all need scale items and time devoted to cases.
2. Time per case per month was crosstabulated with total need score within each level of supervision.
3. Time per case per month was crosstabulated with each need item.
4. Combinations of need items (e.g., alcohol abuse, employment, and companions) were combined and crosstabulated with time.

Table A2 presents average time devoted at various need score levels within maximum and medium supervision levels. At medium supervision, cases with very low need scores (0 - 7) had an average of 50 minutes per month recorded on case-based activities. Time gradually increased to 88 minutes per month devoted to cases scoring 25 or above. At maximum supervision, average time ranged from 95 minutes per month for cases scoring 0 to 7, to 173 minutes per month for cases with scores of 35 or more.[5]

Table A2

TIME SPENT ON ALL CASES BY NEEDS SCORES AND SUPERVISION LEVEL 1989 STUDY

Need Score Ranges	Medium Supervision	% Increase	Maximum Supervision	% Increase
0 - 7	50 min./mo.		95 min./mo.	
8 - 14	58 min./mo.	16%	101 min./mo.	6%
15 - 24	65 min./mo.	12%	111 min./mo.	11%
25 and above	88 min./mo.*	35%		
25 - 29	NA	NA	122 min./mo.	10%
30 - 34	NA	NA	138 min./mo.	13%
35 and above	NA	NA	173 min./mo.	25%

* Because of the small number of Medium cases scoring over 25, breakdowns into the ranges of 25-29, 30-34, and 35 plus were not advisable.

In Table A3, relationships between individual scale items and average time are presented. In every instance but one (mental ability), average time increases in conjunction with increases in item scores. The reversal evident with mental ability is probably either an artifact of the low number of cases with serious problems or represents the fact that seriously disabled persons are put in sheltered living and work situations where only limited agent involvement is required. Overall, the relationship between item scores and average time demonstrate that the scale performs as designed: higher ratings translate into more time devoted to cases.

5 In minimum supervision, there was too little variance in need scores to present meaningful results.

Table A3

TIME BY NEED RATINGS

Need Items/Score	N of Cases	Mean Time Recorded (in minutes per month)	Correlation Coefficient
Academic/Vocational Skills			
-1	902	108 min	
0	1519	113 min	
2	877	128 min	
4	333	149 min	.0951
Employment			
-1	305	79 min	
0	1283	100 min	
3	1572	128 min	
6	471	166 min	.1902
Financial Management			
-1	62	75 min	
0	791	86 min	
3	1837	121 min	
5	941	145 min	.1669
Marital/Family Relationships			
-1	49	61 min	
0	1180	89 min	
3	1592	120 min	
5	810	162 min	.2112
Companions			
-1	30	76 min	
0	1546	92 min	
2	1652	132 min	
4	403	173 min	.2157
Emotional Stability			
-2	15	90 min	
0	1426	90 min	
4	1751	130 min	
7	439	168 min	.2067
Alcohol Usage			
0	1722	102 min	
3	1225	127 min	
6	684	145 min	.1332
Other Drug Usage			
0	2147	99 min	
3	916	133 min	
5	568	169 min	.2037
Mental Ability			
0	3309	117 min	
3	277	142 min	
6	45	124 min	.0449
Health			
0	2044	103 min	
1	1237	133 min	
2	350	157 min	.1476
Sexual Behavior			
0	3079	116 min	
3	263	129 min	
5	289	142 min	.0623
Agent Impression on Needs			
-1	77	47 min	
0	650	64 min	
3	1467	116 min	
5	1437	151 min	.2581

C. REVISING THE NEEDS SCALE

The correlations and crosstabulations presented in Table A3 do suggest that changes in item weights and cut-off scores are appropriate at this time. In addition, some categories of severity seldom apply; hence, they could be collapsed or eliminated to simplify the scale.

Although we hypothesized that specific combinations of needs may result in major increases in time devoted to cases (independent of total need scores), extensive testing failed to indicate that the current additive format of needs assessment should be replaced.

In summary, our analyses of the Wisconsin needs assessment instrument indicates that the following changes should be made:

1. The strength categories, represented by negative weights, should be eliminated. The data suggest they are seldom used and the fact that negative numbers are assigned to them probably cause problems with addition, decreasing accuracy and reliability.

2. Our analysis also indicates that changes to the weights assigned to need scale items are in order. Each item is listed below with its current weights and the recommended change.

Need Item	Current Values	Proposed Values
Academics/Vocational Skills	-1,0,2,4	0,2,4
Employment	-1,0,3,6	0,3,6
Financial Management	-1,0,3,5	0,2,4
Marital/Family Relationships	-1,0,3,5	0,3,5
Companions	-1,0,2,4	0,3,6
Emotional Stability	-2,0,4,7	0,3,6
Alcohol Usage	0,3,6	0,2,4
Other Drug Usage	0,3,5	0,3,6
Mental Ability	0,3,6	0,2,4
Health	0,1,2	0,2,5
Sexual Behavior	0,3,5	0,2,4
Agent's Impression	-1,0,3,5	0,3,5

The proposed changes reflect time by item categories as reported in Table A3. With the current scale, the maximum attainable score is 60. With the recommended changes, the maximum possible score is 59.

3. Cut-off points should be revised to reflect the changes in item values. Current and proposed cut-off scores are presented below:

	Current Cut-Offs	Proposed Cut-Offs
Low Needs	0 - 14	0 - 14
Moderate Needs	15 - 29	15 - 24
High Needs	30 - 60	25 - 59

Time requirements jumped significantly when need scores reached 20 points in the 1979 study. In 1989, the largest increase was noted at 25 points. Time devoted to cases at every level of need dropped from 1979 to 1989, probably reflecting a shift in emphasis toward risk management in probation and parole. Nevertheless, the new need values should produce a total need score that corresponds more closely to supervisory time. A simulation of the recommended revisions indicates that at 25 points, time reported per case increases substantially.

Revised Needs Scale

Select the appropriate answer and enter the associated weight in the score column. Higher numbers indicate more severe problems. Total all scores. If client is to be referred to a community resource or to clinical services, check appropriate referral box.

ACADEMIC/VOCATIONAL SKILLS

0	Adequate skills; able to handle every day requirements	2	Low skill level causing minor adjustment problems	4	Minimal skill level causing serious adjustment problems

EMPLOYMENT

0	Secure employment; no difficulties reported, or homemaker, student or retired	3	Unsatisfactory employment, or unemployed but has adequate job skills	6	Unemployed and virtually unemployable, needs training

FINANCIAL MANAGEMENT

0	No current difficulties	2	Situational or minor difficulties	4	Severe difficulties; may include garnishment, bad checks or bankruptcy

MARITAL/FAMILY RELATIONSHIPS

0	Relatively stable relationships	3	Some disorganization or stress but potential for improvement	5	Major disorganization or stress

COMPANIONS

0	No adverse relationships	3	Associations with occasional negative results	6	Associations almost completely negative

EMOTIONAL STABILITY

0	No symptoms of emotional instability; appropriate emotional responses	3	Symptoms limit but do not prohibit adequate functioning; e.g., excessive anxiety	6	Symptoms prohibit adequate functioning; e.g., lashes out or retreats into self

ALCOHOL USAGE

0	No interference with functioning	2	Occasional use some disruption of functioning	4	Frequent abuse; serious disruption; needs treatment

OTHER DRUG ABUSE

0	No interference with functioning	3	Occasional substance abuse, some disruption of functioning	6	Frequent substance abuse; serious disruption; needs treatment

MENTAL ABILITY

0	Able to function independently	2	Some need for assistance; potential for adequate adjustment; mild retardation	4	Deficiencies severely limit independent functioning; moderate retardation

HEALTH

0	Sound physical health; seldom ill	2	Handicap or illness interferes with functioning on a recurring basis	5	Serious handicap or chronic illness; needs frequent medical care

SEXUAL BEHAVIOR

0	No apparent dysfunction	2	Real or perceived situational or minor problems	4	Real or perceived chronic or severe problems

AGENT'S IMPRESSION OF CLIENT'S NEEDS

0	Low	3	Medium	5	Maximum

D. IMPLICATIONS FOR CLASSIFICATION

The changes recommended will result in minor shifts in the number of clients assigned to each supervision level, but will better reflect actual time requirements. Figures A1 and A2 present the number of cases in the study sample that would be assigned to each supervision level (excluding the impact of the assaultive offense item and overrides) first under the current classification and then with the proposed changes.

Figure A1

CURRENT RISK AND NEED SCALES CLASSIFICATION MATRIX

Risk/Need	Low Need	Moderate Need	High Need
Low Risk	15.4%	9.1%	0.9%
Moderate Risk	10.4%	16.3%	3.9%
High Risk	7.6%	20.3%	16.1%

RESULTING SUPERVISION LEVELS AT INITIAL CLASSIFICATION:

Maximum	48.8%
Medium	35.8%
Minimum	15.4%
TOTAL	100.0%

It should be noted that cases will continue to be reclassified to lower supervision levels at six-month intervals. The above figures only represent how new admissions break out at initial classification.

Figure A2

REVISED RISK AND NEED SCALE CLASSIFICATION MATRIX

	Low Need	Moderate Need	High Need
Low Risk	14.7%	7.4%	1.5%
Moderate Risk	12.2%	13.8%	8.5%
High Risk	7.5%	13.7%	20.6%

RESULTING SUPERVISION LEVELS AT INITIAL CLASSIFICATION:

Maximum	51.8%
Medium	33.4%
Minimum	14.7%
TOTAL	100.0%

The slight increase in cases assigned to maximum supervision is the result of giving greater emphasis to client needs (a lower cut-off point for maximum) and greater weight to need areas which require more agent time. The 3% shift could well be off-set by a lower rate of override and the fact that fewer cases may be placed in maximum solely on the basis of past assaultive behavior.

BIBLIOGRAPHY

Baird, S. Christopher, Todd R. Clear, and Patricia M. Harris (1986). *The Behavior Control Tools of Probation Officers.* Final Report to the National Institute of Corrections.

Baird, S. Christopher, Richard Heinz, and Brian J. Bemus (1979). *The Wisconsin Classification and Workload Deployment Project, Final Report.* Madison: Bureau of Community Corrections.

Clear, Todd R., and S. Christopher Baird (1987). "In-Out Decisionmaking: A Conceptual Framework," *Perspectives.* (Fall) 11:4, p. 10.

Clear, Todd R., and Vincent O'Leary (1982). *Controlling the Offender in the Community.* Lexington, MA: Lexington.

Gottfredson, Stephen D. (1987). "Prediction: An Overview of Selected Methodological Issues. Classification and Prediction." In D. Gottfredson and M. Tonry, eds. *Prediction and Classification.* Chicago: University of Chicago.

Gottfredson, Stephen D., and Don M. Gottfredson (1986). "The Accuracy of Prediction." In Christy A. Visher, *Criminal Careers and "Career Criminals".* Washington, DC: National Academy of Sciences.

National Council on Crime and Delinquency (1988). *Development of Risk Assessment Indices for Alaska Family Services.* Report to the Alaska Division of Family and Youth Services.

_____ (1989). *Revalidation of the Colorado Risk Assessment System.* Report to the Colorado Office of the State Court Administrator.

_____ (1988). *Revalidation of the Illinois Risk Assessment System.* Report to the Administrative Office of the Illinois Courts.

_____ (1990). *Iowa Community Corrections Risk/Needs Assessment Study - Preliminary Report.* Report to the Iowa Department of Corrections.

_____ (1990). *Development of the Michigan Delinquency Risk Assessment Instruments.* Report to the Michigan Office of Children and Youth Services.

_____ (1987). *Oregon Risk Assessment Project - Final Report.* Report to the Oregon Criminal Justice Council.

_____ (1985). *Risk Assessment in Parole Decision Making.* Report to the South Carolina Board of Parole and Community Corrections.

_____ (1990). *Tennessee Board of Paroles Risk Assessment Report.* Report to the Tennessee Board of Paroles.

_____ (1990). *Revalidation of the Wisconsin Probation/Parole Classification System.* Report to the Wisconsin Bureau of Community Corrections.

National Institute of Corrections (1981). *Model Probation and Parole Management Project.* Washington, DC: National Institute of Corrections.

_____ (1989). Classification and Case Management for Probation and Parole: A Practitioner's Guide. Washington, DC; National Institute of Corrections.

O'Leary, Vincent, and Todd R. Clear (1984). *Community Corrections in the 1990s.* Washington, DC: National Institute of Corrections.

Petersilia, Joan (1986). *Prison versus Probation in California: Implications for Crime and Offender Recidivism*. R-3323-NIJ. Santa Monica, CA: The RAND Corporation.

Wright, Kevin W., Todd R. Clear, and Paul Dickson (1984). "A Critique of the Universal Applicability of Risk Assessment Instruments," Criminology 22:1, February, pp. 113-133.