Centennial 100
NCCD

February 2009

SPECIAL REPORT

National Council on Crime and Delinquency

A Question of Evidence: A Critique of Risk Assessment Models Used in the Justice System

Christopher Baird

Prologue

In 1972, I conducted my first "parole prediction" studies for the Illinois Department of Corrections. The purpose of these studies was simple—to identify groups of offenders with significantly different rates of recidivism. As others, including Burgess (1928) and Babst, Gottfredson, and Ballard (1968) had discovered, making clear distinctions between low, moderate, and high risk offenders is possible.

Three years later, while working in the Wisconsin Division of Corrections, I again explored the principal aim of risk classification but added the concepts of differential supervision and treatment, based not only on risk assessment but also on each offender's treatment needs. The question was, does responding differently to offenders with different risk/needs profiles produce better outcomes? The Wisconsin Case Classification/ Staff Deployment Project was successful enough by 1980 to form the basis for the National Institute of Corrections' Model Probation and Parole Management Program. This program spread the use of actuarial risk assessment from coast to coast.

When I joined the National Council on Crime and Delinquency (NCCD), we turned our attention to the child welfare arena to determine if a similar classification process could be applied to decision making in child protection and foster care. We were successful, and despite a near-universal reluctance to apply "numbers" to social work, actuarial risk assessment is now promoted as best practice in child welfare and used across the country.

In 2002, NCCD re-engaged in development and evaluation of assessment systems for juvenile justice, taking part in a joint project with the National Council

of Juvenile and Family Court Judges (sponsored by the Office of Juvenile Justice and Delinquency Prevention) to create the Graduated Sanctions Center.

At the time NCCD undertook the graduated sanctions project, it seemed that significant progress had been made in risk assessment systems in the justice field. New risk assessment systems, referred to as generations three and four, were being touted throughout the country. Risk factors were now delineated into two groups, static and dynamic. "Criminogenic needs" and protective factors were added to risk assessment instruments. These instruments, previously limited to between eight and 12 factors, now contained anywhere from 25 to more than 100 factors. Separate risk instruments were used for sex offenders and for violence prediction. The field was experiencing a virtual explosion of reliance on evidence-based practice (EBP), and the new assessments were playing a critical role.

These developments seemed admirable, but the more NCCD investigated the evidence provided to support newer approaches to risk assessment, the more we questioned the results. The field had rightly embraced EBP, but at least on the assessment side of the equation, the promise of innovation had trumped actual performance. It appeared as though the justice field had abandoned clarity, parsimony, and to a significant degree, the validity and reliability of existing systems. Today there is a confusing array of overlapping terms, objectives, and reliability and validity measures in place.

This report explores the problems with the present state of risk assessment in the justice field as we at NCCD see them. The critique offered here is the result of many conversations with others in the justice community as well as a review of predictive research conducted in other fields. We recognize that much of what is presented is contrary to current understanding and acceptance, but we hope that it clarifies what evidence is required for the designation of best practice.

Our critique deals first with the primary purpose of risk assessment, and includes a short review of the recent literature on validity and reliability. A subsequent section addresses the terminology introduced with these approaches and discusses why, from NCCD's perspective, these terms convey false expectations. Throughout this report, we present data from studies of risk assessment systems currently in wide use throughout the country. The point is not to criticize any particular approach (as many other examples could be used to demonstrate a problem), but to use actual data to clarify issues for policy makers and practitioners who are attempting to select and implement reliable and valid assessment systems.

A Question of Evidence

Over the last few years, an emphasis on EBP has become a predominant theme in the justice field. EBP has, in fact, become the litmus test for both assessment and treatment approaches. This focus on EBP is needed and has the potential to significantly improve practice. In some instances, substantial progress is already evident.

Several related factors initially gave rise to growing interest in EBP. In juvenile justice, for example, the field had seen funding for treatment and service programs decimated by "get tough" public policy mandates that placed more juveniles in adult-type facilities (or in many cases, actual adult facilities). In many jurisdictions, all emphasis on treatment virtually disappeared. Delinquents were often warehoused away from public scrutiny, provided little in the way of education or treatment, and returned to communities in far worse shape than they had been before they were incarcerated. Needless to say, this trend was an abysmal failure. In the backlash that naturally ensued, the search for better approaches was spearheaded by advocates eager for solutions. Consequently, the door was wide open to proponents of new methods for dealing with offenders. The only requirement for entry was some evidence that an approach was effective. A number of risk assessment models quickly gained widespread acceptance, and their accompanying terminology has become firmly entrenched in the lexicon of the justice community.

While some good has certainly been accomplished, there is an unsettling perspective expressed more frequently

in recent years that something has also been lost (Gottfredson & Moriarty, 2006). The sheer volume of the nation's correctional population suggests it is time to step back and take a fresh look at the evidence behind current approaches before the field is locked into less than optimal methods for the assessment, supervision, and treatment of offenders.

Model Validity

We begin with a discussion of the most important issue in risk assessment: validity. The intent of actuarial risk assessment is to identify subgroups within an offender population who have significantly different rates of recidivism. A number of successful actuarial models have been developed over the last six decades. Although the models are frequently depicted as a means to predict which offenders will reoffend, actuarial risk assessment is more appropriately described in terms of classification. These systems simply apply group statistics to individual decisions to help agencies identify where they should focus their resources. In essence, these tools establish base expectancy rates for offenders who have different profiles.

The early actuarial risk assessment instruments were simple, usually consisting of fewer than a dozen factors. More recently, risk assessment systems like the LSI-R, COMPAS, PACT, LS/CMI, and the YASI have focused on differences between static and dynamic risk factors (often defined as criminogenic needs) and have added the goal of risk reduction to the models (Schwalbe, 2008).* As a result, these systems contain many more factors, usually arranged within domains. The LSI-R, for example, is comprised of 54 risk factors and eight domains (Andrews & Bonta, 1995). The full YASI contains 117 factors (Orbis Partners, 2008).

The use of a large number of risk factors and/or domains represents a significant departure from both the format and content of earlier instruments, and also raises a number of questions.

First, what is meant by the term "risk reduction" in the context of risk assessment? Because any actual reduction in risk can only be attained through intervention, we assume it merely refers to the inclusion of risk measures that can change over time and therefore, reduce the scored risk level.

This assumption raises a critical issue, as it appears that many of these factors are not statistically related to any measure of recidivism. Austin and colleagues (Austin, Coleman, Peyton, & Johnson, 2003), for example, found that relatively few of the LSI-R factors had significant correlations with outcomes. Other studies, even those done by proponents of these models, have frequently determined that a substantial number of risk factors demonstrate little or no relationship to recidivism (Flores, Travis, & Latessa, 2004).

If a factor on a risk assessment instrument has no demonstrated relationship to recidivism, how can positive changes in this factor reduce risk?

Out of these questions emerges a conundrum: If a factor on a risk assessment instrument has no demonstrated relationship to recidivism, how can positive changes in this factor reduce risk? Logically, one must conclude that if a factor is not related to recidivism in the first place, changes in that factor should have little or no impact on outcomes. (This report's section on criminogenic needs will further this discussion.)

Of all these issues, the most important is this: Can the inclusion of factors without significant statistical relationships to recidivism actually *reduce*, rather than improve, a model's ability to accurately classify cases? Simple logic suggests that this is indeed not just possible but likely. Non-validated factors, when included in a risk scoring system, introduce substantial "noise" and dilute the relationship between legitimate risk factors and recidivism. It is for precisely this reason that Gottfredson and Moriarty (2006) conclude:

* LSI-R: Level of Service Inventory–Revised; COMPAS: Correctional Offender Management Profiling for Alternative Sanctions; PACT: Positive Achievement Change Tool; LS/CMI: Level of Service/Case Management Inventory; YASI: Youth Assessment and Screening Instrument.

> We would argue that predicting who will or will not behave criminally is risk assessment, whereas using predictive methods to attempt a reduction in criminality through assignment to differential treatments is needs assessment.
>
> Often, then, risk assessments are being used to predict success and failure in treatment programs. We do not argue that treatment success or failure is unrelated to recidivism; we argue that using risk assessment tools to predict treatment success and/or failure is a misapplication of the tool and that properly constructed needs assessment devices, assessed against a proper criterion variable (e.g., treatment outcome), would prove to have greater validity for that purpose and hence greater value to those concerned with offender treatment.
>
> We find the focus on static and dynamic variables in this context to be of little help. The focus should be diverted back to public safety risk and treatment risk, two separate yet often commingled concepts. The best predictor of future behavior is past behavior, and it can be argued that past behavior is static. However, even if we entertain the notion of change and introduce variables that measure change in an offender, the discussion is not relevant to the fundamental problems of current application: confusion over appropriate criterion measures, with consequent misapplication of otherwise perfectly acceptable decision-making aids. We find the discussion in the literature about static and dynamic variables to be neither helpful nor germane to the clarity of purpose that the field appears desperately to need. If a variable can be measured reliably and if it is predictive, then of course it should be used—absent legal or ethical challenge. (p. 193)

We agree entirely: risk and needs should not be combined as a composite measure. We find arguments to the contrary unpersuasive (for example, Schwalbe, 2008).

An example from another field can help clarify this point. Suppose health researchers find that using six factors, they can divide cancer patients into three risk levels—high, moderate, and low—based on recurrence rates observed in a large cohort of cancer patients. In the same study, they discover six additional factors that, although not related to the chance of recurrence, are related to the quality of life following initial medical intervention. The research team decides to put all 12 factors on the same scale and weight them equally. Their highest risk patient, who scores on all six factors related to recurrence, does not score on any of the quality-of-life factors. Their lowest risk patient, who scores on none of the factors related to recurrence but on every quality-of-life factor, will receive the same score. Consequently, two drastically different cases—or combinations of risk and needs—score exactly the same, placing them in the same intervention category even though their individual situations are quite different. This is precisely what occurs in the justice community when factors not related to recidivism are included in a risk index. Frankly, this is poorly conceived practice.

Despite the inclusion of factors without significant relationships to recidivism, these risk models contain enough valid risk factors to attain, in many instances, a modest relationship with various measures of recidivism (see, for example, Flores et al., 2004). Most researchers never ask the next logical question: Would classification results improve if these non-related factors were left out of the instrument? A study of the LSI-R in Pennsylvania (Austin et al., 2003) explored this issue, and produced a dramatic improvement in accuracy using only eight of the 54 LSI-R factors. Results of this analysis are presented in Table 1.

Note that the more concise scale not only produced better separation among risk categories, it also dramatically altered the proportion of cases at each risk level, placing more cases in the moderate and low risk categories. This has substantial implications for both release decision making and allocation of resources, including staff supervision and reentry programs and services. In this instance, because the instrument is used by the parole board, the potential impact on individual offenders is especially profound.

Since few researchers have actually scrutinized these systems as closely as did Austin and colleagues, direct comparisons like the Pennsylvania example are rare. Most studies seem constrained by the model being evaluated. While they present data describing how

the models in question perform, there is rarely any attempt to improve a model's performance. The models themselves appear to be sacrosanct, which may stem from concerns with copyright infringement. To the extent this is true, it represents a grave concern. A principal tenet governing the use of risk assessment is that identification of high risk cases is critical to effective case management. High risk cases are to be targeted for additional services and supervision. A model that does not optimally classify cases based on risk will undoubtedly lead to less effective service delivery and a less-than-optimal use of resources.

Flores, Travis, and Latessa (2004), in a study of the Y-LSI, found that "relatively few of the 42 items contribute to accuracy in risk classification" (p. 1). Such a finding obviously supports the need to separate risk and needs into distinct indices. However, rather than suggesting modification of the existing model based on the results of their study, the authors stated, "If the purpose of classification is limited to risk classification, other instruments should be considered…If, on the other hand, the agency wishes to assess needs, and use need assessment information to develop and deliver effective interventions, our data suggest the Y-LSI is a useful tool" (Flores et al., 2004, p. 2). NCCD recommends a different solution: agencies need to obtain measures of both risk and needs. Simply separating risk and need factors into different indices will produce better measures of both.

Table 1

Outcome Comparisons by Risk Level: Pennsylvania Parolees

	Full LSI-R		Eight Factors From LSI-R	
Risk Level	N	Rate of Recidivism	N	Rate of Recidivism
Low	86 (9%)	43%	146 (15%)	34%
Moderate	398 (40%)	51%	614 (65%)	53%
High	522 (52%)	58%	186 (20%)	69%

Source: Austin, Coleman, Peyton, & Johnson, 2003.

The approach to validation used by Flores and colleagues stands in direct contrast to a study conducted for the Nevada Department of Probation and Parole (Wagner, Quigley, Ehrlich, & Baird, 1998). Nearly 20 years earlier, Nevada had adopted the Wisconsin risk assessment model. Wagner and colleagues found that the Wisconsin instrument accurately classified cases in Nevada, but improved classification results were attained when jurisdiction-specific revisions to the scale were introduced, including deleting a risk factor and changing factor weights and cut-off points used to identify high, moderate, and low risk offenders. A revised instrument was recommended.

There is substantial evidence available to suggest that relatively brief risk indices outperform longer, more complex models. For example, the short criminal history pre-screen used in the YASI (or PACT) system appears to have a significantly stronger relationship to recidivism than the full system (Wagner, 2008). Furthermore, when we compare results obtained from jurisdictions using simple risk indices to those published for generation three and four instruments, the simple scales typically produce better results. Table 2 compares results of a validation of the 42-factor LS/CMI with the 11-factor risk assessment instrument used in Nevada (Onifade, Davidson, Campbell, Turke, Malinowski, & Turner, 2008; Wagner et al., 1998). Neither instrument was developed specifically for the agency where it was tested, so these data show a true comparison of validation efforts.

Although we do not know how each instrument would have performed when administered to the other population, the data serve to illustrate what many studies have demonstrated: concise (eight- to 12-item) risk assessment scales provide accurate estimates of risk, and

almost always exceed levels of discrimination produced by more complex systems (see Wagner et al., 1998; DMJM & Huskey and Associates, 2007; Wagner & DeComo, 1995). Furthermore, because simple jurisdiction-specific revisions to the Nevada assessment were permitted, the degree of discrimination attained between risk levels improved, resulting in better decisions regarding supervision requirements. If nothing else, the field needs to determine if the added time and effort required by a risk inventory that includes between 50 and more than 100 factors is worth the investment.

risk levels, this is sometimes not discussed at all (see, for example, Flores, Lowenkamp, Smith, & Latessa, 2006) despite the fact that this is the clearest measure of how risk affects actual practice.

There are a number of reasons why a simple analysis of recidivism rates by risk level should be the standard for evaluating risk assessment systems. The first involves clarity: it is a measure readily understood by all who use the system and conveys more useful information than a correlation coefficient of .25 or an AUC of .70. Other reasons include the fact that many of the traditional measures of "predictive accuracy" are based on the

Table 2

Comparison of Validation Results for LS/CMI and Nevada Department of Probation and Parole Risk Assessment Instruments

	LS/CMI Validation		Nevada Validation		Nevada Revised	
Risk Level	N	Rate of Recidivism	N	Rate of Recidivism	N	Rate of Recidivism
Low	82 (25%)	11%	286 (23%)	9%	229 (18%)	8%
Moderate	167 (51%)	26%	433 (34%)	24%	470 (47%)	22%
High	79 (24%)	39%	549 (43%)	45%	355 (35%)	51%

Sources: Onifade et al., 2008; Wagner et al., 1998.

Measures of Validity

We believe that the corrections field also needs to rethink how the efficacy of risk classification is measured. Typically, statistical measures of association between risk scores and outcomes are reported. In some instances, this is limited to correlations between risk scores and recidivism, but more often, studies include additional measures of association (area under the curve, or AUC, for example) and/or measures of specificity and sensitivity (receiver operating characteristic, or ROC). The latter measure reflects the rates of produced false positives and false negatives. While some studies present outcome rates by assigned

assumption that decision options are dichotomous, e.g., whether offenders recidivate or not (Silver & Banks, 1998). This assumption simply does not reflect practice, where risk assessment typically guides decisions involving a continuum of options. Nearly all risk classification systems assign cases to at least three different risk levels, and some employ six or more levels of risk.

From a practice perspective, we must ask: What prediction is being made for cases at mid-range risk levels? Is moderate risk a prediction of success and failure? NCCD contends that it is neither. It is simply recognition that cases in these categories reoffend at

higher levels than some offenders and lower levels than others. This knowledge helps agencies determine the level of interventions—and resources—needed to appropriately supervise these individuals.

As Silver and Banks (1998) noted, "The primary utility of a risk classification model is in providing a continuum of risk estimates…which can be used to guide a range of decision making responses" (p. 8). Silver and Banks developed a summary statistic, the dispersion index for risk (DIFR), which assesses how a cohort is partitioned into different risk groups and the extent to which group outcomes vary from the base rate for the entire cohort. While the DIFR has not been widely utilized, it represents a clear illustration of what should be the primary issues addressed in system efficiency: proportionality and differences in outcome rates among several risk groups.

Similar measures are found in medical research. Altman and Royston (2000) have proposed a simple index of separation (the PSEP) because it is both "interpretable and pragmatic" (p. 460). The PSEP simply measures the distance between failure rates of the lowest and highest risk groups. The measure's chief weakness is that it does not include proportionality. Still, it illustrates the fact that the "distance" in outcome rates between risk levels is the critical measure of model validity.

Reliability

Nearly all of the literature on popular risk models refers to their demonstrated validity and reliability. In actuality, there is little information available that supports model reliability, and much of what is available either addresses the wrong issue (internal consistency) or provides inadequate tests of inter-rater reliability.

Inter-rater reliability is particularly critical when models include 25 or more items, many of which are scored using subjective judgment. When there is little or no consistency among staff members completing risk instruments, the validity of the system cannot be assumed. Reliability studies have demonstrated that the "static" factors related to criminal history are the most consistently rated risk factors (Austin et al., 2003; Baird, Heinz, & Bemus, 1979). Items requiring greater subjective judgment, such as marital/family factors, use of leisure time, and peer relationships, have significantly lower reported rates of reliability. The more of these factors included in a scale, the greater the potential for classification error. Some models, such as the YASI and the LSI-R, are predominantly composed of such measures. At least one test of the reliability of the LSI-R found serious deficiencies, particularly with factors described as criminogenic needs (Austin et al., 2003). The study also found that additional staff training improved reliability results. However, it should be noted that measuring reliability soon after the completion of training sessions may inflate the rate above what will be observed in practice. Regression to the mean is a frequent occurrence; that is, as time passes, staff tend to revert to old (pre-training) work habits.

The best measure of inter-rater reliability is a simple statistic that is both pragmatic and interpretable: specifically, percent agreement across a number of independent raters. (Frequently, Cohen's kappa is applied to percent agreement to account for "chance agreement.") Before the Wisconsin needs assessment instrument was introduced statewide, percent agreement was determined for 45–50 raters who independently assessed nine different cases (Baird et al., 1979, p. 17). Given the importance of reliability in instruments that require considerable subjective judgment, rigorous tests of reliability are critical before these instruments are widely disseminated.

Some model developers bypass any discussion of inter-rater reliability entirely and focus instead on the level of "internal consistency" across all items on a risk instrument (see, for example, Brennan & Ehret, 2007). However, NCCD contends that examining internal consistency of items on a risk assessment scale is actually counterproductive. The focus on internal consistency comes from the field of psychology, which typically tests "constructs" rather than observable outcomes. Cronbach's alpha is often the preferred measure of internal consistency and is used in many risk studies in the field of corrections. However, Cronbach's alpha is not a proper measure of the reliability (or validity) of risk assessment instruments. Cronbach's

alpha measures the extent to which item responses obtained at the same time correlate with each other (Garson, 2003). This is important when measuring a construct, such as depression. Since there is no objective test for depression, it is impossible to correlate an item such as depression with an observable criterion to validate it. The next best approach is to hypothesize that all items on a depression scale should have some degree of covariance. Cronbach's alpha is ideal for this purpose.

Despite the fact that many of these generation three and four models have been widely used for years, there is little evidence to suggest they have any impact on outcomes.

Recidivism, on the other hand, is not a construct, but an observable outcome. The relationship between risk factors and recidivism does not have to be estimated; it can be measured. For risk assessment, it is best when all risk items are totally independent of each other but each has a relatively strong relationship to the outcome measure utilized. These principles are in direct conflict with maximizing Cronbach's alpha. Cronbach's alpha, in fact, is used more to measure scale validity than reliability when the dependent variable is a psychological construct. Logic suggests that if all factors are hypothesized to be related to the construct in question and all are strongly related to each other, it follows that they indeed measure that construct.

Our perspective is that measures of internal consistency, while important in psychology, are not relevant to risk assessment in corrections. Their use only serves to divert attention from the important issue of inter-rater reliability and to create confusion among practitioners.

Measuring Impact

Despite the fact that many of these generation three and four models have been widely used for years, there is little evidence to suggest they have any impact on outcomes. This is particularly puzzling given that these new instruments focus strongly on criminogenic needs and "risk reduction." The premise is that identifying criminogenic needs and focusing interventions in these areas will reduce recidivism for offenders. Testing this hypothesis would seem to be a critical step before extensive replication of these models occurs. It is widely assumed that use of these instruments addresses the correct needs and that the measures adequately identify interventions best suited to prevent subsequent offending. However, there is little empirical evidence that this is true. Further, even if criminogenic needs were adequately assessed, the fact that these models may produce less-than-optimal assessments of risk (as discussed earlier) indicates that they may still target the wrong cases for intervention.

While it is always difficult to establish experimental designs in agencies implementing new practices, simply comparing outcomes before and after implementation of a risk assessment model should be possible in many jurisdictions. Such evaluation designs have their limitations, but they can provide data to help determine the degree of effectiveness of a particular model in reducing recidivism.

Terminology

As noted earlier, we are concerned with some of the terminology that has emerged over the last two decades. This concern stems not from the terms themselves, but from their use within the context of risk assessment. Two such terms are "criminogenic needs" and "protective factors." The use of these terms has become so widespread that it is nearly impossible to discuss risk assessment models without referencing both. Indeed, despite our ambivalence regarding these terms, we have felt compelled to include them in our own published literature, as the field expects to see them addressed. They are nevertheless problematic.

The term "criminogenic" was coined in the 1980s to convey a relationship between dynamic risk factors (or criminogenic needs) and offending behavior. Combining "criminal" with "genesis"—the Greek word

for the point at which something comes into being—implies that particular needs can create or generate criminal behavior. Typical lists of criminogenic needs generally encompass six to eight needs categories or domains, including the following:

- Parenting/Family Relationships
- Education/Employment
- Substance Abuse
- Leisure/Recreation
- Peer Relationships
- Emotional Stability/Mental Health
- Criminal Orientation
- Residential Stability

Most professionals would agree that any one of these factors could contribute to criminal behavior in individual cases. However, the mere existence of a need does not always mean it is criminogenic. Further, nothing in these risk models systematically identifies which needs truly are criminogenic for an individual offender. For example, association with the wrong peer group could lead one youth into delinquent behavior, while for another youth, association with delinquent peers may simply be an artifact of his/her delinquency. In both cases, many risk assessment models label the need as criminogenic, implying a claim about causality that generally far exceeds what can legitimately be concluded from the assessment data.

The practice of labeling all needs as criminogenic appears to be a misguided effort to merge risk assessment—which uses group data to inform certain fundamental case decisions—with case planning, which must be based on the individual circumstances of each offender. Labeling a need as criminogenic when it has little or nothing to do with criminal behavior is counterproductive, leading to ineffective interventions and unnecessary expense.

Over the last 60 years or so, assessment systems have been developed that attempt to identify the underlying reasons for criminal behavior, but these systems are much more complex than the simple listing of needs factors. Typically, the developers attempted to design offender typologies and use these profiles to identify how specific needs/problems/developmental characteristics may generate offending behavior in individual cases. The first of these systems, the I-Level, showed promise, but training in the use of this model was prohibitively expensive; certification as an I-Level counselor required approximately six weeks of training. Later attempts, the Quay system (a derivative of the I-Level) and Lerner and Arling's CMC and SJS systems, proved more manageable in terms of cost (Harris, 1988; Lerner, Arling, & Baird, 1986). Evaluations of the CMC have shown it can have an extraordinary impact on recidivism (Leninger, 1998; Eisenberg & Markly, 1987; McManus, Stagg, & McDuffie, 1988). NCCD's position is that most generation three and four instruments simply do not include the level of analysis required to accurately identify needs as criminogenic in individual cases.

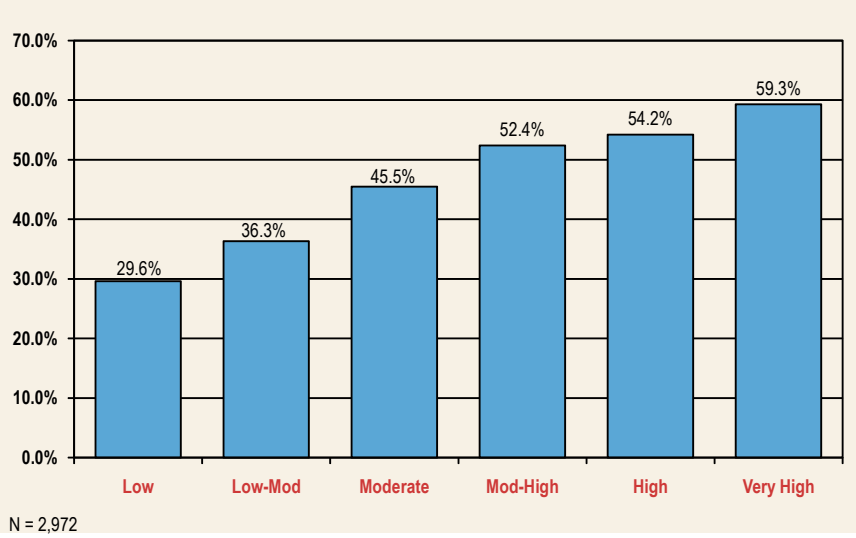
The role of protective factors in risk assessment is particularly mystifying. For the most part, each protective factor seems to represent the absence or opposite of a risk factor (e.g., substance abuse is a risk factor; ergo, the fact that a youth does not abuse substances is a protective factor). To the extent that this is true, protective factors offer nothing in terms of increasing our ability to accurately classify cases. They are simply measures of the same condition or behavior from a different perspective.

A recent evaluation of the YASI illustrates this point (Orbis Partners, 2007). Two graphs are reproduced here. Figure 1 delineates negative outcomes by dynamic risk level. Figure 2 presents negative outcomes by dynamic protective level. They are nearly mirror images of each other, with the dynamic risk level showing a slightly stronger relationship to recidivism, as would be expected. Hence, if the objective is to accurately classify cases into groups with significantly different outcomes, there is simply no point in using both measures.

The mere absence of a risk factor does not translate into a protective factor for an individual. If substance abuse has nothing to do with an offender's criminal behavior, its absence is not likely to protect the individual from subsequent criminal activity. This is not meant to imply

Figure 1

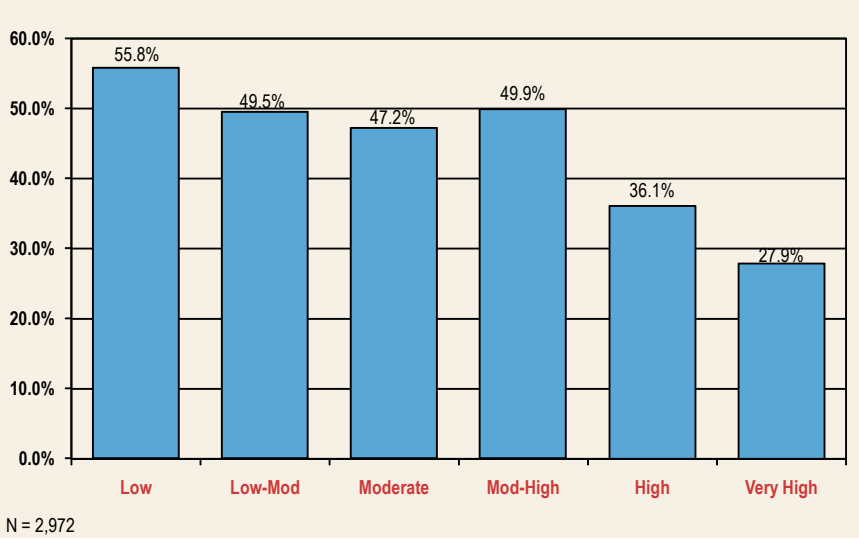
Negative Outcomes by Dynamic Risk Level in New York State Juvenile Probation



Source: Orbis Partners, 2007.

Figure 2

Negative Outcomes by Dynamic Protective Level in New York State Juvenile Probation



Source: Orbis Partners, 2007.

that protective factors are not important to both case planning and case management. However, the manner in which protective factors are assessed and represented in many risk assessment models undermines their usefulness in helping workers complete these functions.

Summary

The focus on evidence-based practice in corrections is an extraordinarily positive development. That said, the need remains to carefully review some of the more recent assessment models used in juvenile and adult correctional systems, because by expanding the objectives of these models, their clarity, validity, and reliability have all been compromised. Further, expectations introduced by the terminology included in these models exceed what can legitimately be accomplished. While some authors (see Schwalbe, 2008) make a case for continuing down this road, others (Gottfredson & Moriarty, 2006; Andresen, 2008; Austin et al., 2003) express reservations similar to those discussed here.

NCCD has conducted dozens of separate risk assessment studies over the last three decades, many in juvenile and adult corrections and many others in child welfare. We recognize that creation of valid, reliable, and robust risk assessment instruments is both a science and an art. We have witnessed the degree to which poorly designed systems can negatively affect practice. We contend that the justice field needs to step back and carefully review both the logic and the level of evidence supporting many current assessment practices. NCCD recommends that the following points, at a minimum, should guide this review:

1. It is obviously important to identify offenders at high risk of recidivism and to devote more services and resources to these cases. With this in mind, the justice field must recognize that combining factors that have little or no relationship to recidivism with validated risk factors cannot improve but can seriously reduce the relationship between risk scores and outcomes.

2. The standard for measuring the efficacy of a risk assessment model should be the level of discrimination attained between risk levels. Correlation coefficients, analyses of false positives and false negatives, and other measures of association may all be helpful in scale construction, but they fail to convey how well a risk model can inform actual case decisions.

3. Simplicity, clarity, and parsimony are important. Assessment scales containing 25–125 variables introduce significant noise and create potential problems with reliability. Since these instruments also include many items that require subjective judgment, well-designed tests of inter-rater reliability are essential. To date, such analyses have been less than adequate.

4. The internal consistency of risk factors or domains that make up a risk assessment model is an inappropriate measure of the model's reliability. Such tests generally provide an estimate of scale validity when measuring a construct. Recidivism is not a construct, but rather a measurable outcome.

5. More caution should be exercised in identifying factors as "criminogenic" or "protective." These are important concepts, but ones that require a significantly deeper level of assessment than many risk models currently provide. As such, they can raise false expectations and lead to inappropriate case plans and services.

6. Since little is known about the effectiveness of many of the newer risk assessment models due to the reliability and validity issues raised here, there are good reasons for skepticism about their actual impact on recidivism and treatment outcomes.

NOTE: NCCD supports standard case assessment and management models for both juvenile and adult offenders (JAIS™ and CAIS™). Assessments of risk and needs within these systems are completely independent indices. Further, if an agency has a risk instrument that has been validated on its population, it replaces the risk assessment embedded in the model. NCCD also revalidates the risk assessment for each agency periodically and makes all appropriate revisions.

References

Altman, D., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19*, 453–473.

Andresen, C. (2008). *Infomercial criminology vs. research: A review and analysis of the research on the Level of Service Inventory–Revised (LSI-R)*. Unpublished manuscript.

Andrews, D., & Bonta, J. (1995). *LSI-R: The Level of Service Inventory–Revised*. Toronto, ONT: Multi-Health Systems, Inc.

Austin, J., Coleman, D., Peyton, J., & Johnson, K.D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, D.C.: Institute on Crime, Justice, and Corrections at The George Washington University.

Babst, D.V., Gottfredson, D.M., & Ballard, K.B. (1968). Comparison of multiple regression and configural analysis techniques for developing base expectancy tables. *Journal of Research in Crime and Delinquency, 5*(1), 72–80.

Baird, C. (1991). *Validating risk assessment instruments used in community corrections*. Madison, WI: National Council on Crime and Delinquency.

Baird, C., Heinz, R., & Bemus, B. (1979). *The Wisconsin Case Classification/Staff Deployment Project: Two-year follow-up report*. Madison, WI: Wisconsin Division of Corrections.

Brennan, T., & Ehret, B. (2007). *COMPAS reclassification scale validation study*. Williamsburg, MI: Northpointe Institute for Public Management.

Burgess, E.W. (1928). Factors determining success or failure on parole. In A.A. Bruce, E.W. Burgess, J. Landesco, & A.J. Harno (Eds.), *The workings of the indeterminate sentence law and the parole system in Illinois*, (pp. 221–234). Springfield, IL: Illinois State Board of Parole.

DMJM & Huskey and Associates. (2007). *State of Illinois Department of Juvenile Justice Comprehensive Master Plan: A Vision for the Future*. Chicago: Huskey and Associates.

Eisenberg, M., & Markly, G. (1987). Something works in community supervision. *Federal Probation, 51*(4).

Flores, A., Lowenkamp, C., Smith, P., & Latessa, E. (2006). Validating the Level of Service Inventory–Revised on a sample of federal probationers. *Federal Probation, 70*(2).

Flores, A.W., Travis, L.F., & Latessa, E.J. (2004). *Case classification for juvenile corrections: An assessment of the Youth Level of Service/Case Management Inventory (YLS/CMI), executive summary* (98-JB-VX-0108). Washington, D.C.: U.S. Department of Justice.

Garson, G.D. (2003). *Scales and measures*. Raleigh: North Carolina State University.

Gottfredson, S.D., & Gottfredson, D.M. (1979). *Screening for risk: A comparison of methods*. Washington, D.C.: U.S. Government Printing Office.

Gottfredson, S.D., & Moriarty, L.J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency, 52*(1), 178–200.

Harris, P.M. (1988). The Interpersonal Maturity Level Classification System: I-Level. *Criminal Justice and Behavior, 15*(1), 58–77.

Hoge, R., & Andrews, D. (1996, August). *The Youth Level of Service/Case Management Inventory: Description and evaluation*. Paper presented at the Annual Conference of the American Psychological Association, Toronto, Canada.

Leninger, K. (1998). *Effectiveness of Client Management Classification*. Tampa, FL: Florida Department of Corrections Research and Analysis.

Lerner, K., Arling, G., & Baird, C. (1986). Client Management Classification strategies for case supervision. *Crime and Delinquency, 32*(3), 254–271.

Lowenkamp, C. & Latessa, E. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections*. Washington, D.C.: National Institute of Corrections, Department of Justice.

McManus, R., Stagg, D., & McDuffie, C.R. (1988). CMC as an effective supervision tool: The South Carolina perspective. *Perspectives*, Summer 1988.

National Institute of Corrections. (1981). *The Model Probation and Parole Management Program*. Washington, D.C.: Author.

Onifade, E., Davidson, W., Campbell, C., Turke, G., Malinowski, J., & Turner, K. (2008, April). Predicting recidivism in probationers with the Youth Level of Service/Case Management Inventory (YLS/CMI). *Criminal Justice and Behavior, 35*(4), 474–483.

Orbis Partners. (2007). *Long-term validation of the Youth Assessment and Screening Instrument (YASI) in New York State Juvenile Probation*. Ottawa, ONT: Author.

Orbis Partners. (2008). *Youth Assessment and Screening Instrument*. Retrieved on January 6, 2009, from www.orbispartners.com.

Schwalbe, C. (2008). Strengthening the integration of actuarial risk assessment with clinical judgment in an evidence-based practice framework. *Children and Youth Services Review, 30*, 1458–1464.

Silver, E., & Banks, S. (1998). *Calibrating the potency of violence risk classification models: The dispersion index for risk (DIFR)*. Washington, D.C.: American Society of Criminology.

Wagner, D. (2008). *Review of alternatives to the Juvenile Assessment Generic (JAG)*. Madison, WI: National Council on Crime and Delinquency.

Wagner, D., & DeComo, R. (1995). *Travis County preliminary findings: Juvenile risk assessment study*. Madison, WI: National Council on Crime and Delinquency.

Wagner, D., Quigley, P., Ehrlich, J., & Baird, C. (1998). *Nevada Probation and Parole risk assessment findings*. Madison, WI: National Council on Crime and Delinquency.

Centennial
Continuing the Struggle for Justice
NCCD

National Council on Crime and Delinquency
426 S. Yellowstone Drive, Suite 250 • Madison, WI 53719
tel 608/831-8882 • fax 608/831-6446 • nccd-crc.org