Criminal Justice System Risk Assessment Models

Quantum Units Education
Affordable. Dependable. Accredited.

www.quantumunitsed.com

In 2009, I published *A Question of Evidence: A Critique of Risk Assessment Models in the Justice System*. This paper identified problems with both the logic and research that support many of the risk assessment models used in the adult and juvenile justice systems. Unfortunately, the issues addressed in that paper remain in force today, further complicated by increased expectations emanating from new methods of analysis. Excitement generated by extremely complex analytical methods has helped propel the belief that mathematical formulae can be applied to a wide variety of circumstances and decision points. As a result, some IT system integrators have taken steps to position themselves to participate and profit from a growing dependence on large-scale quantitative analysis, especially in the area of risk assessment. But success in one area does not mean that such methods can or should be applied to all decision points. While these methods may work well in predicting macro trends, there is little evidence that they can improve predictions of individual behavior. Such predictions would be riddled with false positives and false negatives. No methodology, no matter how sophisticated, can be so precise when it comes to predicting the behavior of individuals. And yet, even before the introduction of these seemingly sophisticated methods of analysis, the social sciences began to adopt language that incorporates terms implying such precision. Risk classification has become risk "prediction"; correlations are frequently referred to as "effect size"; and needs that correlate, albeit modestly, with recidivism are called "criminogenic," implying causation.

The first problem is that the role group data can play in individual decision making is, by definition, limited; and, as some including NCCD have noted, misapplication is fraught with both ethical and logical problems. NCCD has long supported the idea that group data have a place in decision making, but agencies need to exercise great care with how these data are used at the level of individual cases. Risk assessment is, for example, now used at sentencing and in parole board hearings. Such application is not entirely new, but growing confidence in the power of numbers has led to a reduction in safeguards previously considered crucial. The original Minnesota Sentencing Guidelines, for example, limited risk factors to those related to an individual's prior record. Today, many risk instruments contain socioeconomic factors and rather crude measures of an individual's psychological condition, and, as many have pointed out, use of such instruments at these decision points may be discriminatory and unconstitutional.

Applying statistical probabilities to individuals is raising questions in other fields as well. A 2014 article by Paul Kalanithi in the *New York Times* described the problem doctors face in applying information gleaned from group data to individual patients. It is one thing to know where patients with certain characteristics and medical histories fall on a survival curve; it is quite another to project how long a specific patient will survive. Kalanithi concludes that it is "impossible, irresponsible even, to be more precise than you can be accurate." Yet that is exactly what assessment models used in adult and juvenile corrections systems do when "criminogenic needs" are identified as those that should be addressed in case planning and that "non-criminogenic needs" are unimportant. This practice, quite simply, implies a level of precision that cannot be justified.

Throughout the nation, over-reliance on statistics has, in fact, created a climate where abuse is commonplace. Examples of data being misunderstood, misused, and misreported abound. Several books, including *Proofiness: The Dark Art of Mathematical Deception*, detail how data have been cleverly manipulated to support particular viewpoints by the business community, politics, and science.

The social sciences, including corrections, have been victimized by such misrepresentation in the past, and the results have sometimes proved devastating. The massive growth in incarceration, for example, was driven to a large extent by seriously flawed analysis of the relationship between crime and incarceration rates, leading to policies that have not only cost taxpayers billions of dollars but have decimated African American communities throughout the nation. Policymakers, both liberal and conservative, are only now coming to grips with just how damaging this movement has been.

This over-reliance on statistical models is every bit as dangerous as failing to use data to help drive decision making. While NCCD has long supported the use of risk assessment and continues to do so, much of what has been published in the field over the last 15 to 20 years on risk assessment is based on suspect logic and poorly conducted research.

This series of briefs, titled *A Question of Evidence: Part Two*, will focus on adult and juvenile justice assessment models currently in use throughout the United States, Canada, and Europe. A recent review of the research behind these models indicates that much of what has been produced over the last two decades may well be far less evidence-based than their developers claim. Further, products of this research are often misapplied, thereby raising expectations far beyond what can be legitimately accomplished with risk assessment.

The purpose of this series is to clarify issues facing administrators charged with selecting risk assessment models. Administrators need to ensure that decisions are based on the best available information. These briefs address problems discovered during a thorough review of available risk models.

I acknowledge that some issues raised will be controversial given the level of acceptance particular concepts have attained. However, I feel that issues raised here and by the above-noted authors, as well

as by others with extensive experience in the field, are a clear indication of the need to inject clarity and transparency into the rhetoric currently dominating literature on risk assessment in adult and juvenile justice. A short description of each brief is presented below.

The Generations Myth

The justice field has come to understand assessment models in terms of the "generation" each is said to represent. It is implied, and sometimes explicitly stated, that each succeeding generation offers greater "predictive" capacity. This brief explores the origins of those claims and discusses the promotional strategies that led to widespread acceptance of the "generations" terminology and associated claims.

Criminogenic Needs

Criminogenic needs (often referred to as "dynamic risk factors") have dominated the literature on risk assessment for years. While assessing needs is a critically important component of assessment, much of what is advocated conflates the roles of group data and the actual treatment needs of the individual. This brief identifies flaws in the logic employed to support the use of criminogenic needs in risk assessment and discusses the appropriate role of needs assessment in case planning and service delivery.

Developing and Validating Risk Assessment Instruments for Justice Agencies

This brief explores the research behind many current models, discusses methods commonly used to measure "predictive power," and outlines what is required to measure the efficacy of various approaches to risk assessment.

Structured Professional Judgment Models

While other fields have moved to more structured decision-support systems, the justice field has seen the development and promotion of less structured approaches to risk assessment, commonly referred to as structured professional judgment (SPJ) models. This brief discusses the research behind this approach, including concerns with validity, reliability, equity, and utility of SPJ models.

Summary and Recommendations

This final brief summarizes the major problems identified throughout the series regarding risk assessment models, then goes on to suggest four steps toward remedying those problems.

About 25 years ago, Andrews, Bonta, and others began describing different "generations" of risk assessment instruments. Although this terminology may have initially been intended only to delineate differences in development methods, it quickly gained footing as a means to promote new risk assessment models.

Culturally, "generations" terminology implies an improvement in design and functioning from earlier models to the latest model. After all, a generation-6 smartphone must offer advantages over the generation-5 smartphone, or there would be no inducement to buy the new model.

Developers of risk assessment tools for adult and juvenile corrections have, in fact, claimed that later-generation models are superior to the earlier instruments already in use (see, for example, Andrews, Bonta, & Wormith, 2006). But a review by this author found no real evidence that these risk assessment models perform better than older models. Instead, our review of efforts to promote the generation-3 and-4 instruments uncovered inconsistencies in labeling, a failure to account for all components of some models, flawed logic behind cited evidence, and claims of enhanced predictive power that are not supported by evidence.

Although there is some variance in how each generation is defined, generation-1 decision making is most often described as clinical judgment or the absence of instrumentation. Generation-2 instruments are defined as those relying only on static factors—those factors that do not change over time. Generation-3 instruments introduced the inclusion of dynamic risk factors, also referred to as "criminogenic" needs. Developers claimed that inclusion of these factors allowed risk levels to change over time based on changes in an individual's circumstances.

Generation-4 models purportedly address "responsivity," meaning that the system provides information on individual capacities and learning style, identifying programs and strategies that are likely to produce success in the community.

As cited in the manual that accompanies the LSI-CMI, a generation-4 tool, "Andrews, Bonta, and Wormith (2006) examined the predictive validity of different generations of risk instruments. The second-generation risk assessments (such as the PCL-R, Wisconsin, and SFS) had predictive validity in the range of .26 to .46. The third-generation assessments (LSI-R) had an average predictive validity for general recidivism of .36 and the fourth-generation [instrument] (LSI/CMI) had a predictive validity of .41. Accordingly, with the improvement in each generation there was improvement in the predictive power of the instrument" (p. 3).

Putting aside the fact that the authors compare ranges to averages and present no information regarding the power of the analyses cited, the data presented do not support their conclusion. The best results (.46) came from a generation-2 assessment; the generation-3 average falls squarely in the middle of the range cited for generation-2 assessments, and the one generation-4 instrument correlated with outcomes (not described) at a lower rate than at least one of the generation-2 instruments. In sum, these comparisons do not provide any evidence of improved predictive power over generations.

Claims supporting the validity of other higher-generation instruments are equally questionable. Independent studies have found that some have little demonstrated validity (see, for example, Flores, Travis, & Latessa, 2004; Skeem & Eno Louden, 2007). In 2010, the Canada Department of Justice reported, "For the most part, these instruments have been adopted without proper validation and reliability studies" (Hannah-Moffat & Maurutto, 2003). The Illinois Juvenile Justice Commission found little data to support another risk assessment model, the YASI, issuing a report stating, "While the YASI is often cited as highly valid and reliable, the author of this report was unable to substantiate such claims and could not locate any peer-reviewed articles in which these properties were assessed" (Illinois Criminal Justice Information Authority, 2010). In essence, these models were adopted by agencies in the United States and Canada before their validity was established, sometimes replacing well-validated tools already in place.

Recent research comparing results from simple actuarial instruments—those likely to be defined as generation-2 models—found that these instruments actually separated high-, moderate-, and low-risk offenders more accurately than risk assessment systems claiming to be generation-3 or -4 (Baird et al., 2013). A review panel of researchers, including developers of later-generation risk assessments, agreed in a response to the NCCD report that later-generation instruments were not more valid than simple, actuarial models—a significant step back from earlier claims (Andrews et al., 2006).

Adoption of these tools appears to be, in essence, based more on promotion than evidence. The following presents a brief synopsis of what occurred.

When the LSI instruments emerged, there was a concentrated effort to demonstrate their superiority over the most widely used system at that time, the National Institute of Corrections (NIC) model, which was based on a system developed for the state of Wisconsin.

As articulated in the NIC manual, agencies adopting the NIC model were encouraged to validate the risk instrument on their own system-involved population and make all appropriate changes as soon as available data permitted. Many validation studies were conducted, some of which resulted in changes to the initial risk instrument to reflect differences in law, policy, practice, and populations in each jurisdiction.

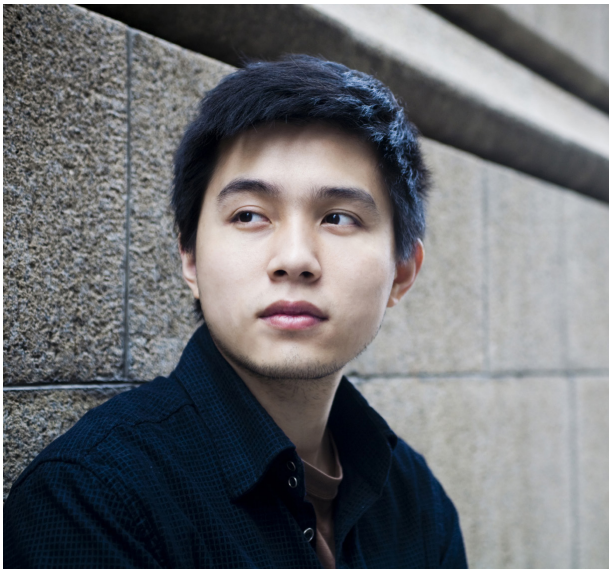
The NIC manual also explained, in detail, the role of an "assaultive offense" item on the initial risk scale. This item was not a risk factor, but included a weighting that, in effect, established a matrix that considered both risk and prior violence to establish an initial level of community supervision. In Wisconsin, every person convicted of an assaultive offense, regardless of risk level, was placed at the highest level of community supervision for the first six months of probation or parole. Each agency using the NIC model was free to adopt this policy or to delete the assaultive offense item from the scale.

The NIC model also included a reassessment risk instrument that focused on behavior observed since the last assessment. This instrument allowed individuals rated high or moderate risk at admission to move to lower supervision levels over time. Over 50%

of all cases rated moderate or high risk moved to lower risk levels over the course of their supervision period (Baird, Heinz, & Bemus, 1979).

Finally, the NIC model included a separate needs assessment. The three objectives of this assessment were (1) to ensure that specific needs were considered for every case; (2) to add consistency to the manner in which needs were assessed; and (3) to provide direction for case planning. Finally, as a 2004 National Institute of Justice survey noted, most agencies adopting the NIC model included CMC (Case Management Classification), a component of the Wisconsin system devoted to what is now called "responsivity" (Hubbard, Travis, & Latessa, 2001).

Despite the fact that the NIC model included all of these elements, virtually all comparisons made by LSI supporters focused solely on the initial risk instrument, which LSI proponents labeled a generation-2 instrument. This was a serious misrepresentation, given that the model also included a risk reassessment and a needs assessment. Moreover, this misrepresentation encouraged perceptions that the Wisconsin model was "not dynamic" and that the LSI and other later-generation risk assessment instruments offered improvements over the Wisconsin model.



Another error is more concerning. In comparing the relative validity of the NIC initial risk assessment to generation-3 and -4 instruments, nearly every study included points generated by the policy factor on the Wisconsin scale, "prior assaultive convictions." Hence, they rated all individuals with a prior assault as high risk, when, in fact, many were not. This incorrect analysis seriously diminished the relationship between risk scores on the Wisconsin scale and recidivism. (For an excellent discussion of this effect, see Eisenberg, Beryl, & Fabelo, 2009). As a result, other instruments appeared to produce higher correlations with recidivism, allowing LSI supporters to claim greater predictive capability.

Other errors were made as well. In comparing results from their own Ohio risk assessment and the Wisconsin system, Latessa and colleagues (Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009) used a version from a Canadian province that *combined* scores from the *Wisconsin risk reassessment and the needs assessment*. Comparing data from an initial risk assessment and a combination of a reassessment and needs assessment is neither meaningful nor useful. The developers' selection and labeling of this as the "Wisconsin model" seems to demonstrate that little care was taken to ensure that comparisons were legitimate. Despite being alerted to this mistake over three years ago, the Ohio report is, at this writing, unchanged and available online.

In the decade following development of the LSI family of instruments, the corrections field was deluged with articles on the LSI. Nearly all repeated the generations language in their introductions. Most of these studies were based on small samples of cases, limiting their value, and nearly all used correlations as the only measure of validity (Vose, Cullen, & Smith, 2008). Furthermore, *any* level of correlation, no matter how modest, was presented as evidence of validity. Even when important reviews questioned the level of evidence behind generation-3 and-4 models, their use continued to expand (Skeem & Eno Louden, 2007; Illinois Criminal Justice Information Authority, 2010).

LSI developers also published papers emphasizing the predictive power of "dynamic risk factors" (sometimes called criminogenic needs), often using obtuse logic to support the contention that they were, in fact, better "predictors" than static factors such as prior measures of criminal behavior. When studies found poor relationships between LSI classifications and recidivism, they were generally downplayed; the lack of validity was often attributed to problems with training and/or implementation that resulted in a lack of fidelity with the model as defined. A few studies that analyzed the relationship of individual risk factors to outcomes, however, indicated that the lack of validity may well be due to design issues, as a number of scale factors simply were found to have little relationship to recidivism (Flores et al., 2004, Austin, Coleman, Peyton, & Johnson, 2005); Baird et al., 2013). Despite these problems, the LSI evidence opened the door for other commercially available assessment models, some of which had even less evidence of validity or reliability.

The LSI promotional effort was very successful. By 2014, the NIC model all but disappeared from the correctional landscape, despite the fact there is little, if any, evidence that the LSI produced results equal to or better than those produced by the NIC model.

Although generations labeling seems entrenched in corrections lexicon, it is clear that it has been used to imply superiority where none exists. So-called generation-3 and -4 instruments are not more dynamic, claims of greater validity are simply not true, and claims of additional capacities are, at best, highly suspect. The developers of the YLS/CMI, for example, claim the system addresses "responsivity" and that this makes it a generation-4 instrument (Andrews et al., 2006). Responsivity means the system matches interventions with an individual's characteristics and learning styles. But there is nothing in the assessment model that does this. Responsivity could perhaps be addressed in training, but the model itself does not contain anything that identifies the learning style or capacities of an individual. While this information would indeed be of value to case planning, it requires far more in-depth analysis than that provided by the YLS/CMI.

There is nothing inherently wrong with using generations terminology to identify real advancements in practice or products. But when differences are not improvements and the terminology is used principally to promote a product, real damage is possible. Resources may be wasted on making unwarranted changes, staff expectations may be raised unrealistically, and case plans may be less effective, meaning that we miss a real opportunity to create a positive impact in our communities and improve public safety. It is especially troubling that some of these systems are finding their way into sentencing and release decision making, assuming a far greater role than is warranted given their limitations. There is a pressing need to clean up the morass of flawed concepts, inconsistencies, false claims, and marketing jargon that permeates corrections. We must be absolutely clear about what risk assessment is and what it can and cannot do for the justice field.

Needs assessments were first introduced in the late 1970s. The objectives of these assessments were to ensure that needs were assessed for every case and to create greater consistency in service planning. Early needs instruments made no claims that needs assessed caused criminal behavior.

If a factor was significantly correlated to subsequent criminal behavior and assisted in accurately classifying individuals to different levels of risk, it was included on the agency's risk instrument. In general, few factors defined as "needs" met this test. Factors with the highest relationships to recidivism most often included substance abuse, employment issues, peers/associates, and school/behavioral issues.

The term "criminogenic needs," meaning needs seen as causing criminal behavior, emerged in the 1980s. Typical lists of criminogenic needs generally encompass four to eight needs categories or domains (known colloquially as the "Big Four," "Big Six," or "Big Eight"), including parenting/family relationships, education/employment, substance abuse, leisure/recreation, peer relationships, emotional stability/mental health, criminal orientation and thinking, and residential stability.

There are serious problems with identifying needs as criminogenic and with the way that various risk models define and measure needs thought to be criminogenic. There were also flaws in the logic that developers used to stress the importance of the role that criminogenic needs play in risk assessment.

Problem 1: Some needs assessed in various risk instruments have little or no relationship to recidivism.

Although nearly all current risk assessment models are described as actuarial, many in truth are not. In actuarial science, scale content (and item weights) is determined by data analysis with the objective of including only those factors that, in combination, best separate cases into different levels of risk. Scale construction is based on actual cases with observed outcomes (Gottfredson & Snyder, 2005).

However, for many generation-3 and -4 risk assessment models, their content was determined by individuals who developed the model, often guided by prior research and/or crime theory. Most of these

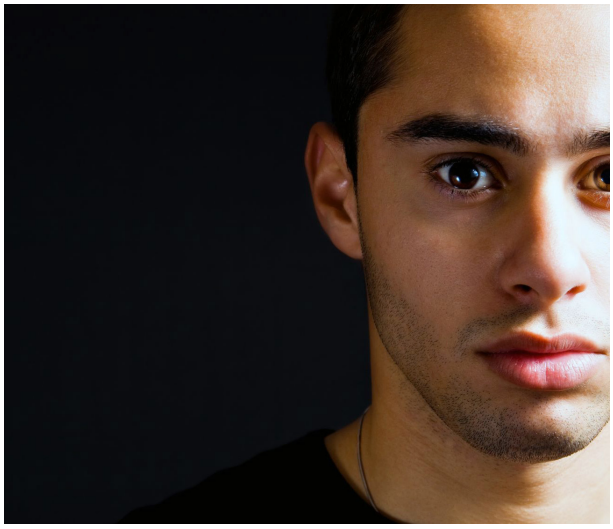
models were then tested for validity, but such analysis was rarely used to revise these instruments.

As a result, several instruments currently in use contain many more factors than true actuarial scales and some of these items have little relationship to outcomes (Flores, Travis, & Latessa, 2004; Baird et al., 2013). Including these factors in a risk score can dramatically change the proportion of cases categorized as high, moderate, and low risk and substantially decrease the degree of discrimination attained between recidivism rates for cases at each level.

No one demonstrated this effect more conclusively than James Austin when he and his colleagues compared results using eight factors from the LSI-R with results from the entire 54-item scale. Much better discrimination was obtained when classifications were based on the eight best scale factors, and the proportion of cases placed at each risk level changed dramatically (Austin, Coleman, Peyton, & Johnson, 2003). More recent research on models used in the juvenile justice system produced similar findings (Baird et al., 2012).

Problem 2: A general statistical relationship between a need and recidivism does not mean that need is "criminogenic" for an individual offender. Still, several models link these needs directly to case planning for individual offenders. Such inference conflates the appropriate use of individual and group data.

Most would agree that any one of the Big Eight factors could contribute to criminal behavior in individual cases. However, the existence of a need does not mean it is always related to, let alone that it causes, criminal behavior. Correlation does not equal causation, yet some developers have made this leap in logic. Hoge and Andrews stated that "risk factors are those identified as *causally* linked with criminal activity" (1996, p .6). This is not true. Risk factors, whether



identified through statistical analysis or reliance on previous studies, are those factors that *correlate* with recidivism.

There is nothing in these models that identifies which needs are criminogenic for a specific offender. For example, association with a particular peer group could lead one youth into delinquent behavior, while for another youth, association with particular peers may simply be an artifact of his or her delinquent behavior. Both would score the same on this risk factor, yet putting effort and resources into changing the peer group of the latter youth would, in all probability, have little effect on his or her delinquent behavior. Yet some risk assessment models label this "peer relationships" need as criminogenic, implying a claim of causality that far exceeds what can legitimately be concluded from the assessment conducted.

The practice of labeling a need as criminogenic without an in-depth clinical assessment to establish causality appears to be an effort to merge risk assessment—which uses group data to inform certain fundamental case decisions—with case planning, which must be based on each person's individual circumstances. Labeling a need as criminogenic when it has little or nothing to do with criminal behavior could lead to ineffective, even harmful, interventions and unnecessary expense.

Conversely, there is also a problem when researchers label needs other than the Big Eight, Big Six, or Big Four as "non-criminogenic" (see, for example, Vincent, Guy, & Grisso, 2012). The lack of relationship between a need and recidivism in the general correctional population does not mean the need is unrelated to (or even the underlying cause of) the criminal behavior of an individual.

A need like "lack of self-esteem" is a prime example. It is one of several factors often identified as non-criminogenic (Taxman, Shepardson, Delano, Mitchell, & Byrne, 2006; Vincent et al., 2012). While seldom a cause of delinquent behavior, self-esteem issues can and do occasionally lead to serious violence. Some acts of horrific violence committed by young people who felt bullied or simply dismissed by peers and authority figures have been linked to self-esteem issues.

In essence, using statistical information (group data) to define what is to be considered in case planning, treatment, and services for an individual represents a misapplication of data. *The presumption that relationships gleaned from group data can be readily applied to individual offenders (particularly when these relationships are modest to begin with) far exceeds any legitimate interpretation of the research.*

Problem 3: There are flaws in the logic used to assert that criminogenic needs represent the most powerful predictors of recidivism.

Andrews, Bonta, and others have stressed the importance of criminogenic needs, basing their views at least in part on the belief that changes in these needs over time are predictive of changes in delinquent or criminal behavior. Consider the following statement regarding the predictive capacity of the LSI instruments when put in the context of actual practice:

> "Dynamic predictive validity is demonstrated when changes in total scores predict changes in the probability of criminal behavior" (Andrews, Bonta, & Wormith, 2008).

This reasoning is both misleading and illogical. All LSI instruments include a substantial number of criminal history items. Scores on these items can, of course, increase if a person is rearrested during the supervision period. Hence, the total LSI score at reassessment will, in most instances, increase when new criminal behavior is observed. In these cases, the change in the LSI score did not *predict* criminal behavior; the change occurred because new criminal behavior was detected. In one study of the YLS/CMI, the delinquency history score increased at reassessment for more than 60% of cases in the sample (Raymour, Kynch, Roberts, & Merrington, 2000). Thus, the increase in total risk scores correlated well with recidivism in large part *because* recidivism led directly to the increase in the risk scores. Calling this evidence of "dynamic predictive validity" is a misrepresentation.

Similar issues can occur in other domains. Substance abuse, especially for minors, is often, in itself, an offense. Youth who continue to use drugs (or alcohol) are thus committing new offenses, so any change in risk scores for these individuals may well be the *result of* a crime, not *predictive of* a crime. The probation officer may know of continued substance abuse because the youth was re-arrested.

Finally, the very idea that assessments conducted well into the supervision period can be "predictive" defies logic. The behaviors and attitudes assessed six, 12, or 18 months into a probation term are clearly enmeshed with outcome measures. Naturally, youth who continue to abuse drugs, consort with peers who commit delinquent acts, etc., are more likely to recidivate than those who do not. But these behaviors are co-occurring, not predictive. Those who do well on supervision are more likely to succeed; those who do poorly, recidivate.

These developers thus have employed circular logic to promote their models. This is not an exercise in prediction. There is a reason that bets (predictions) cannot be placed after a horse race begins. If bets could be placed at the seventh furlong, when a good portion of the race has been completed, predictive accuracy would undoubtedly improve. Predictions, by definition, are made before an event occurs, not well into the event.

Still, reassessment plays a critically important role in corrections. At intake, risk is based on group probabilities because there is no experience with the individual, at least in the current timeframe. If properly designed, re-assessment instruments can shift emphasis from group probabilities to the actual behavior of each individual. Appropriate revisions to supervision requirements, treatment plans, and services provided can be made based on each person's response to supervision. Reassessment and consequent changes to case plans and supervision requirements are crucial to an individual's success as well as community safety.

In summary, including needs that have little relationship to outcomes on a risk assessment likely has significant implications on the instrument's power. Risk instruments should contain only those factors that, in combination, produce the greatest degree of discrimination between recidivism rates for individuals at different risk levels.

Further, assuming that needs that are statistically related to recidivism are criminogenic for a specific individual far exceeds any legitimate interpretation of statistical inference. Needs should be assessed separately for case planning and service-delivery purposes. Combining the two in a single scale conflates the roles of group and individual data.

Articles on risk assessments from the last few decades include numerous claims that these systems are actuarial models. In truth, most are not. True actuarial systems are developed in the following manner: Once a study cohort is identified, a wide range of variables is selected for inclusion in data analysis. Factors analyzed may be based on prior research, theory, speculation, or simple curiosity. A dependent (or outcome) variable—some measure of recidivism over a specific period of time—is also identified. Statistical analysis is then conducted to determine which factors are related to recidivism. Those not related are eliminated. The objective is to determine which combination of risk factors most accurately classifies the cohort into different levels of risk based on actual outcomes.

If a study cohort is large enough, it is divided into construction and validation samples. This is important, because the best results obtained are almost always attained with construction samples. Once an instrument is developed, testing it on a second, or validation, sample provides a better estimate of accuracy in actual practice.

If the study cohort is not sufficient to create two samples, the instrument may be implemented in the agency for which it was developed. Then, to determine how well the risk tool works in actual practice, a prospective validation should be conducted using new cases. There are both advantages and disadvantages to prospective validation. One advantage is that it provides data not only from a separate sample, but from a different time period as well. The principal disadvantage is that prospective validations take much longer to complete.

In essence, actuarial systems are produced by data analysis. However, most systems currently in use were not developed through analysis, but rather constructed by researchers or clinicians. Because developers cite prior research and theory in factor selection, these systems contain variables thought to be related to recidivism (Vincent, Guy, & Grisso, 2013). However, subsequent validation studies have shown that many factors contained in these instruments are not related to continued criminal or delinquent activity. As noted earlier, Flores, Travis, and Latessa (2004) found this to be true of the YLS; Austin and colleagues found it true of the LSI-R (2003). In a comprehensive study of instruments used in juvenile justice, NCCD found that several instruments, including the YLS/CMI, PACT, COMPAS-Youth, and the YASI, all contain factors with little or no relationship to recidivism (Baird et al., 2013). Table 1 lists factors from the analysis of PACT that demonstrated little or no correlation with recidivism.

Table 1: Correlations for Selected PACT Risk Factors and Recidivism for Probationers in Florida

Risk Factor	Correlation
Prior Weapon Referrals	0.00
Prior Felonies Against Persons	0.02
Escapes	0.00*
Commitment Orders/One Day or More	0.03
Gender	0.04
History of Mental Health Issues	0.04

*Actual value is 0.002.

Factors not related to recidivism introduce substantial "noise" and dilute the relationship between overall risk scores and outcomes. The result of including such factors is, that while most of these instruments contain enough real correlates with recidivism and demonstrate a modest ability to classify cases, accuracy could be improved by removing factors unrelated to outcomes. In many instances, the level of improvement attained using true actuarial techniques is substantial (Baird et al., 2013). Researchers need to return to true actuarial development methods to ensure optimal classification of cases. If cases are not accurately classified, all other goals of case management may be seriously compromised.

The research field has also generally abandoned the most important means for evaluating the validity of risk assessment instruments. As noted by Gottfredson and Snyder (2005), two measures should be used to establish the validity and utility of risk assessment systems: (1) the degree of discrimination observed between recidivism rates for cases at different levels of risk and (2) the distribution of cases throughout the risk levels. An earlier National Institute of Corrections publication (Baird, 1991) stipulated the same criteria for evaluating the efficacy of risk assessment instruments. Silver and Banks (1998) not only identified these criteria as critical but actually developed a summary statistic that assesses how well a cohort is partitioned into different risk groups and the extent to which group outcomes vary from the base rate for the entire cohort. Their work was predicated on the position that distribution and the level of discrimination attained are critical to understanding the power of any system. While measures of specificity, sensitivity, association, and false positives/false negatives are useful, they simply are general measures of validity that do not accurately convey the utility of a system in everyday decision making. Yet claims of validity are frequently based solely on these measures. Most analyses published between 1995 and 2010 did not report recidivism rates for different risk groups.

In recent years, risk assessment validity has been based almost exclusively on two measures: the AUC (area under the curve) or simple correlations between risk scores and recidivism. Both are general measures of validity and, while useful measures, do not take into account two factors that are enormously important: the overall recidivism rate for the study cohort and the distribution of cases across risk levels.

The AUC has become particularly popular in recent years. Supporters cite the fact that this measure does not consider base rates or the distribution of cases across risk levels as strengths of the AUC, noting that this allows for easy comparisons of results

across systems. These "strengths," however, are in reality serious weaknesses. The AUC represents the chance that a true positive (i.e., a recidivist) selected at random will have a higher risk score than a true negative (a non-recidivist) also randomly selected. However, there are many scenarios where a high AUC can be attained for a system that produces low levels of discrimination and has extremely limited utility. This is especially true when there are few true positives (i.e., rates of recidivism are low). Hence, when one instrument clearly produces a higher level of discrimination between risk levels than another, AUC values for the two scales can be similar. This allows supporters of risk instruments that are based on prior research and theory to insist their instruments are as accurate as actuarial models.

This is precisely what a group of respondents did to challenge the conclusions of a recent study of instruments used in the juvenile justice field (Baird et al., 2013). Interestingly, these reviewers made no claim that the later-generation instruments were better, only that they were equivalent in terms of predictive accuracy, a step back from earlier claims made by Andrews, Bonta, and Wormith (2006). Their view is that tools with similar AUCs should produce approximately equal classification results; it is only a matter of selecting the proper cut points. There is little evidence, however, that this is true. In the study cited above, different cut points were used for all of the tools analyzed in an effort to optimize classification results. Improvement was noted for only one system, which was an anomaly due to a scoring system that produced a very narrow range of scores and the fact that two risk factors accounted for virtually all of the discrimination attained.

To understand how misleading the assumption that similar AUCs translate into equal classification tools is, consider one of the examples used to support this argument. AUCs for a risk assessment instrument used in Georgia and an actuarial instrument developed using data from the same jurisdiction were .64 and .67, respectively. This fact, combined with similar comparisons from other jurisdictions, led the respondents to conclude: "Fundamentally, this study provides evidence that tools that differ in their length, format, and foci can achieve similar levels of predictive utility." However, other measures of predictive utility clearly demonstrate that sole reliance on the AUC is problematic. Table 2 compares discrimination results as well as distribution across risk levels for the same two assessments. The original tool placed only 1% of cases in the high-risk category: In effect the system identified only two risk categories and placed nearly nine of every 10 youth at the lowest risk level. The actuarial system produced a much better distribution across risk levels and effectively separated cases into low-, moderate-, and high-risk categories. As a result, the DIFR statistic developed by Silver and Banks to measure the power of a risk prediction model was much higher (.61 vs. .40) for the actuarial system.

In sum, despite the instruments' similar AUC values, classification results from the actuarial system are clearly superior. The actuarial system has far greater utility, despite producing only a slightly higher AUC. This group of respondents appear to have given no consideration to other measures of predictive utility in reaching their conclusion.

The level of discrimination attained by different tools simply cannot be ignored. The primary purpose of risk assessment is to assign cases to different risk levels. Either implicitly or explicitly, the assigned risk level plays a role in case decision making, ranging from assigning a supervision level in the community to informing a decision to incarcerate a young person. Given the importance of these decisions, risk assessment systems must optimize differences in outcomes observed for cases at different risk levels. It is clear that assessments with similar AUC values often produce very different classification results.

The standard for measuring the efficacy of a risk assessment model should be the level of discrimination attained between outcomes for cases at each risk level. AUC values may be helpful in scale construction, but they fail to accurately convey how well a risk model operates in actual practice.

Table 2: Comparing Assessments With Similar AUCs

Risk Level	Actuarial Risk Instrument		Original Risk Instrument	
	% at Level	Recidivism Rate	% at Level	Recidivism Rate
Low	32%	17.0%	88%	25.3%
Moderate	44%	37.1%	11%	52.4%
High	24%	49.1%	1%	57.5%

Structured professional judgment (SPJ) instruments are used at various points in justice decision making, including release to parole and sanctioning of adjudicated youth. In other disciplines, these tools are called "expert" or "consensus" systems. SPJ models identify factors to consider in assessing risk (usually based on prior research and/or theory), but items are not scored. At the end of every assessment, workers simply assign a level of risk, presumably based on the overall profile derived from factors assessed. It is one of the most troubling developments in risk assessment in recent years.

Rating risk factors but not summing them to create an overall risk score has long proved problematic. As far back as the 1950s, assessment studies demonstrated that while highly trained clinicians could reliably rate individual risk factors, they had little success in predicting outcomes. In one seminal study of this issue, Bleckner (1954) found that simply summing the ratings given individual factors by clinicians produced far more accurate predictions of outcomes than individual clinicians could provide. A study of SPJs in child welfare found they were neither valid nor reliable and were significantly outperformed by additive actuarial models (Baird & Wagner, 2009). Given this history, coupled with the relative success of actuarial tools in the justice field, it seems strange that anyone would suggest SPJs have advantages over risk assessment systems where factors are weighted and scored to produce a risk level.

Several assessments serve as examples of problems with the SPJ approach. Some used in the adult criminal justice system (e.g., the PCL-R) were not developed specifically for corrections. These systems were originally intended for use in the mental health system to assist in diagnosing psychopathy and gradually made their way into corrections. Recent studies have raised issues concerning reliability and validity (see, for example, Singh, Frazal, Ralitza, & Buchanan, 2014; Yang, Wong, & Coid, 2010). Their use in release decisions is especially controversial and currently is being challenged in court in California.

One SPJ assessment, the SAVRY, was developed specifically for the juvenile justice system to assess a youth's potential for violence. The SAVRY is particularly important because it has gained considerable support in recent years, due in part to the MacArthur Foundation's Models for Change initiative. The SAVRY is used in the United States and several European countries.

Vincent and colleagues claimed that "the SAVRY has reported the best predictive accuracy of any instrument based on available research" (Vincent, Terry, & Maney, 2009). The lead author cited, as evidence for this claim, a "meta-analysis" of 11 studies of the SAVRY (Yang et al., 2010). However, a detailed review of those studies reveals the following.

1. Though the study is presented as a meta-analysis, the review really represents a basic compendium of very small studies. The average sample size of studies cited in this article is 113 cases; no study included more than 176 cases, and four studies included fewer than 100 cases.
2. These studies were from a variety of countries including Canada, Spain, the Netherlands, England, and Germany. Policies and practices undoubtedly varied greatly across these countries, seriously diminishing the value of any attempt to combine results. Only one study was conducted in the United States, which presented results indicating poor predictive validity.
3. Validity measures presented in the meta-analysis were limited to correlation coefficients and ROC values. No data were provided regarding the ability of SAVRY to discriminate by risk level.
4. Follow-up periods varied substantially. In some studies, a standard follow-up period was not used.

These issues are important. First, small study samples offer little in terms of knowledge advancement. When small samples are divided into three or four different risk categories, not to mention gender or racial/ethnic groupings, the number of cases in each category is too small to produce stable and meaningful statistics. The total combined sample from all 11 SAVRY studies was 1,239, a figure frequently exceeded by single studies of other risk assessment models (see, for example, Gottfredson & Snyder, 2005; and Baird et al., 2013). The Gottfredson and Snyder study sample

alone comprised more than 9,500 cases. Secondly, combining the results of small studies from agencies with widely disparate policies, procedures, and offender populations is problematic at best. Third, the US study, the second largest cited, found the SAVRY produced a correlation of .15 with recidivism. It is not uncommon to find a single factor (such as prior delinquencies) that correlates at a higher level. Finally, results from Canada far exceeded those produced in other countries, raising questions regarding transferability.

SAVRY proponents who served as consultants on the Models for Change initiative generally ignored the work of many other researchers, citing problems with short actuarial risk models that simply do not exist. For example, in *Risk Assessment in Juvenile Justice: A Guidebook for Implementation*, they state that without evidence or citation, "unfortunately, there is no brief risk tool currently available that can adequately identify low-risk youth" (Vincent, Guy, & Grisso, 2013, p. 36). This is not true. The guidebook fails to consider findings from a major study conducted in 2005 for the Office of Juvenile Justice and Delinquency Prevention by highly respected researchers (Gottfredson & Snyder, 2005). These researchers developed a nine-

item scale that effectively divided cases into five risk categories with substantial differences in outcomes, accurately identifying low-risk cases. In addition, the National Council on Crime and Delinquency (NCCD) developed an 11-item scale that outperformed longer instruments in identifying low-risk youth (Baird et al., 2013).

Further, some studies conducted to measure the validity of the SAVRY focused on an analysis of "scores" derived by summing items from SAVRY subscales (Gretton & Abramowitz, 2002; Catchpole & Gretton, 2003). While studies of total or subscale "scores" may indicate some relationship between SAVRY scores and outcomes, they say little about "summary risk level" assigned by workers. Some studies have, in fact, found surprisingly low correlation between scores and summary ratings, an indication that subjectivity may play a significant role in assigning overall risk ratings (Gretton & Abramowitz, 2002).

Inter-rater reliability citations used to support the SAVRY are also frequently based on very limited studies. In one study, SAVRY ratings were completed by two teams, each composed of two staff members with advanced degrees (Lodewijks, Dorleleijers, & Ruiter, 2008). Thus, assignments to risk levels were agreed upon by two individuals. This does not represent what occurs in most US probation agencies, where assessments are completed independently by dozens, if not hundreds, of staff with very different levels of experience and education. The small size of the study (N=25), combined with the fact that it was completed on cases from a European country with different policies, procedures, and laws, renders the reliability results rather meaningless to any US jurisdiction. Further review of the analysis of individual scale factors conducted by these researchers demonstrates just how spurious relationships derived from small samples can be: Several items shown in prior research to be related to recidivism and violence were inversely correlated with outcomes in this study. This is counterintuitive and at odds with violence theory at the foundation of the SAVRY. This is very likely an artifact of the small sample size.

Models for Change consultants have also been selective regarding results produced by the SAVRY in Louisiana. A 2011 publication linked reductions in the use of residential placement to the SAVRY (Models for Change Knowledge Brief, 2011). It may be likely that these reductions are related more to increases in community-based programs introduced as part of the program. No data on the relationship between SAVRY ratings and outcomes are presented.

The Models for Change consultants did, to their credit, conduct an inter-rater reliability study of the SAVRY, concluding that ratings among staff members were highly reliable. But this finding stands in sharp contrast to other findings regarding SPJs (Baird & Wagner, 2000; D'Andrade, Benton, & Austin, 2005). The difference in findings may well be attributable to study methodology.

There is no perfect way to conduct reliability studies. Results obtained from such studies always will be mere estimates of the level of reliability attained in the field. Replicating the actual field assessment process as closely as possible probably produces the best estimates. Whereas NCCD used case files from four jurisdictions in its study of SPJs in child welfare (Baird, Wagner, Healy, & Johnson, 1999), Models for Change consultants used vignettes created specifically for the study. Using case files to measure reliability may well produce a more accurate estimate of what will be achieved in actual practice. Case files represent information typically collected and available to assessors.

A plethora of research indicates that SPJs do not provide the degree of structure needed to ensure reliability among raters in large, diverse agencies (D'Andrade et al., 2005). Agreement among independent raters is often well below acceptable standards, indicating the SPJ process is simply too subjective to provide a high level of consistency among staff members.

In sum, evidence supporting the SAVRY is not well established and is based almost entirely on very small studies that often fail to report on the actual levels of discrimination achieved. There also is research showing that SPJ models lack the level of validity needed to improve decision making. Most actuarial models allow workers to override the risk level derived through scoring if they know something about the case that indicates a higher or lower designation. Overrides often require supervisory review and can be tracked to help ensure fidelity to the system. This builds in flexibility to include judgment of the staff member conducting the assessment, while placing needed control on the use of subjective judgment. SPJs lack the structure needed in justice decision making and represent a step backward in practice. Agencies should avoid using these tools.

In this series of briefs, I described problems with risk assessment models that have emerged over the last two or three decades. The sources of these problems are varied, ranging from poorly designed research, flawed logic, and misrepresentations of older, well-established risk assessment systems to the proliferation of for-profit vendors that sell and support risk assessment models. To summarize, the major issues identified include the following.

1. Most newer systems are not truly actuarial. These systems were frequently marketed before being adequately tested for validity, reliability, or equity.
2. Language associated with the goals and objectives of risk assessment has changed significantly, suggesting a level of precision that far exceeds what can be legitimately inferred from available research.
3. For many years, the principal measures of validity, the degree of discrimination attained between observed outcomes for cases assigned to different risk levels, were largely ignored. Criticisms have had some effect and, lately, there has been something of a return to standard measures of validity.
4. In promoting the use of new "generation 3" and "generation 4" risk assessment models, developers misrepresented existing models and ignored important research conducted for county, state, and federal agencies. Conclusions that the new systems offered greater "predictive validity" and better reflect changes over time have been thoroughly refuted.
5. The emergence of the for-profit sector in the development of risk assessment systems represents a major change in the justice and corrections landscape. Historically, risk assessment research was conducted by universities, nonprofit research organizations, and state research offices, often funded by grants from the federal government. Risk instruments were generally developed for individual agencies, reflecting the laws, policies, and populations of each jurisdiction. Furthermore, because they were developed with public funding, most systems were in the public domain. Today, many in the research community promote risk assessment models that are "transferable" or "generalizable." There are two major problems

with this approach. First, as noted above, many models were marketed before being adequately tested for validity or reliability. Second, even if valid and reliable, these instruments will not perform optimally in all jurisdictions as they fail to reflect local policy, law, practice, and population differences.

The ability to purchase a validated risk model can be an attractive alternative to the time, cost, and effort associated with system development, and could be more effective if developers encouraged agencies to make changes based on follow-up research in each jurisdiction. However, firms know that supporting various renditions of a model is difficult (and therefore costly), so customization is not always encouraged.

These developments come at considerable cost to the efficacy of decision making. In many instances, follow-up research indicates these systems do not work as intended (Flores, Travis, & Latessa, 2004; Baird et al., 2013). But these findings seem to have little impact, as marketing continues without revisions to existing models. Such results are often ignored by the research community and attributed to lack of fidelity to the model.

While these briefs paint a bleak picture of current practice, remedies need be neither difficult nor costly. Risk systems can be easily streamlined and improved, given the amount of data now available in most agencies. However, to stem the proliferation of instruments that fail to optimally discriminate between high-, moderate-, and low-risk offenders, both researchers and correctional administrators must be clear about the objectives of risk and needs assessments. The following steps would improve assessment practice and add the clarity that is desperately needed if assessment is to optimally guide decision making in corrections.

1. First, justice agency officials must better understand the roles of risk and needs assessments. Marketing strategies are often cloaked in research and statistical terminology, and few administrators have the technical background needed to effectively evaluate claims of validity and reliability. Having research expertise on staff would be valuable, but smaller agencies seldom have this "luxury." As an alternative, federal agencies such as the National Institute of Corrections, the National Institute of Justice, and the Office of Juvenile Justice and Delinquency Prevention could establish guidelines for assessment practice, including guidance on interpreting research results. This could significantly increase understanding at the agency level and improve both decision making and outcomes. These guidelines are sorely needed. Vendor claims of "predictive accuracy" will be more difficult to evaluate as statistical modeling becomes more complex and larger firms, with greater marketing potential, enter the field.
2. The research community must become far more self-regulating. Peer review needs to improve, and uncritical acceptance of ideas published in prior articles should end. Although beyond the

scope of this paper, the volume of error found in journal articles and other publications is astounding.[1] Critiques should play an essential role in knowledge development and transfer. Recently, however, criticizing accepted views has opened researchers to attacks from both developers and users,[2] which discourages challenges to widely accepted assessment and treatment protocols. As a result, the justice field has been dominated by a few individuals whose work has attained a status that few researchers are willing to challenge. At the same time, excellent publications by other respected researchers have been basically ignored, rarely cited by vendors marketing specific models (e.g., Gottfredson & Snyder, 2005; Gottfredson & Moriarty, 2006; Austin, Coleman, Peyton, & Johnson, 2003; and Fabelo, Nagy, & Prins, 2011).

3. The inconsistencies, flawed logic, and inadequate research that have permeated the field over the last two decades must be addressed. In this era of evidence-based practice, justice and corrections officials rely on the research community to provide guidance in selecting both assessment and treatment programs. When issues of professional advancement and financial interest enter the equation, the potential for abuse increases. While these issues will never be eliminated, everyone needs to understand they exist. These briefs identify problems with the research that supports many widely used approaches to assessment, flaws in the design of many models, and statements meant to provide guidance that are simply inaccurate. Given the degree to which emphasis on evidence-based practice influences policy and practice, it is essential that the evidence provided is accurate, unbiased, and open to critique.

4. Finally, proponents of new analytical methods (e.g., neuro-networking and random forest analysis) are making unprecedented claims of predictive accuracy. These claims should

[1] Many statements are made in journals and reports without citation or evidence presented to substantiate their content. In other instances, the evidence presented does not support the claims made. Unfortunately, many such statements are repeated in subsequent publications. Several such statements are presented in this series.

[2] For example, one researcher's well-constructed critique of research used to support multisystemic therapy was met with a deluge of unwarranted criticism from the industry (Littell, 2006).

be viewed with caution and fully vetted by the research community. Often, the results are no better than those produced by well-designed studies that use traditional methods of scale development. Other fields—medicine, for example—also have found that the new analytical techniques provide no benefits over standard scale development methods (Altman & Royston, 2000). These methods sometimes produce highly suspect results, overstated by omitting large segments of the target population, and can result in predictive models that have little decision-making utility. Furthermore, developers of these models are often reluctant to share their algorithms with users, claiming they are "intellectual property." As a result, these systems become "black box" models, as users are not provided with the criteria used to assign risk levels. There are two major issues with black box systems. First, workers cannot determine if an override is appropriate because the basis for assignment to a risk level is unknown. Second, no one can effectively challenge any decisions influenced by risk assessment, again because information used to make the decision is unknown. When these tools are used in sentencing or release decisions, their constitutionality can and should be challenged. As prior briefs have noted, many existing models contain factors that have no relationship to recidivism or to violence. In addition, black box models could well contain factors that discriminate against specific groups of people (e.g., people of color), or factors that should have no role in sentencing, placement, or parole decision making. Transparency should be a requirement, and no jurisdiction should allow black box systems any role in decision making.

Baird, C. A Question of Evidence, Part Two.
(2017, January). National Council on Crime & Delinquency.