



**Commentary on  
NCCD**

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

(Baird, 2009)

By:

Tim Brennan, PhD

Bill Dieterich, PhD

Markus Breitenbach, PhD

Brian Mattson, PhD

June 2009

## Introduction

Controversies can be useful. They can bring issues to light that deserve more attention, demand resolution, and initiate methodological improvements. A recent National Council on Crime and Delinquency, (NCCD) paper by Chris Baird, "A Question of Evidence: A Critique of Risk Assessment Models Used in the Justice System" is a serious criticism of several current developments in the area of risk and needs assessment. Baird directs many of his comments at "modern risk assessments" often known as third generation (3G) or fourth generation (4G) assessment systems. These systems are designed for routine assessment and classification of offenders in correctional agencies. Baird is a respected participant in corrections and has made fine contributions to the field. His Wisconsin and CMC classification systems were highly innovative when they first appeared on the scene about three decades ago. In his paper, he attributes numerous methodological, conceptual and design flaws to the whole category of "newly developed" assessment and classification instruments, and also questions the usefulness of concepts such as "criminogenic needs", "dynamic versus static risk factors" and "protective factors" that appear central to the prevailing "what works" paradigm.

While Baird's paper is largely negative about the major current risk and needs assessment systems, it offers a timely critique of several important conceptual and methodological issues in correctional assessment and raises a number of challenging and in some cases unresolved issues. Thus, his paper is helpful in drawing attention to these controversial issues that deserve to be fully addressed as we move forward in this key area of correctional practice.

Baird's paper pointedly focuses on purported design flaws of the family of recently developed risk/needs models and of the LSI in particular. While we agree with many of the issues he raises, it is important not to overlook some flaws and errors in several of Baird's arguments. His critique includes several overstated arguments, omission of relevant research and in some important instances an absence of key technical data.

Given the wide range of critical issues in Baird's paper it would take more than one discussion to fully respond to all of these issues. Thus, in this paper we offer some perspectives and comments only on eight of the major issues raised. As noted earlier we agree that these issues deserve serious discussion but disagree with some of the conclusions drawn by Baird in his paper.

**Issue 1: Baird assigns all “modern” 3G and 4G instruments into one large category – with little effort to distinguish between them**

Baird makes a consistent and unfortunate categorization error by lumping all current risk and needs assessment and classification instruments into one homogeneous category. This categorization error leads directly to a series of overgeneralizations in which all “modern instruments” are tarred with the same brush. The instruments he critiques, (LSI/CMI, COMPAS, LSI-R, PACT, YASI, etc) in fact, are quite diverse, with diverse design characteristics, measurement approaches, different statistical procedures and even different target populations. Yet, Baird treats them as homogeneous, attributing a series of specific design flaws and conceptual errors to them all. In point of fact, most of Baird's criticisms and examples are specifically aimed at the LSI and then more implicitly attributed to the other 3G and 4G instruments. Many of Baird's attributions to COMPAS are resolved upon closer inspection - as discussed below.

**Issue 2: Baird claims new methods include non-valid and irrelevant factors in their predictive models - with significant damage to predictive accuracy.**

In this criticism of predictive validity - which Baird describes as the “most important” function of assessment (p.3) – Baird offers a classic “straw man” argument in which he attempts to extend to all 3G/4G assessments the weakness he attributes to the LSI. He criticizes the LSI for its simple additive summation of all 54 items to create its overall risk model. As is well known to statisticians, this practice may allow many lesser-predictive factors to enter a predictive model. Such “noise” variables can then blur the boundaries, weaken discrimination between the predictive categories and weaken predictive accuracy. Baird cites specific studies and technical details to demonstrate that the LSI incurs this problem (Austin, Coleman, Peyton, & Johnson, 2003; Flores, Travis, & Latessa, 2004).

However, this particular design flaw does not occur in every 3G/4G assessment instrument, as Baird implies. For example, all the Northpointe COMPAS predictive risk models use statistical procedures to both identify and avoid this problem (Brennan, Dieterich & Ehret 2009). Our general recidivism risk scale and violent recidivism risk scale were developed using advanced regression modeling strategies (Harrell, 2001). The risk scale inputs were selected from a pool of candidates consisting of criminal history variables and COMPAS base scales using methods that protect against over-fitting. Our risk scales are designed to have good discrimination and to be well calibrated; they do not include arbitrary items. For example, the recidivism risk scale includes the following weighted inputs: age at first arrest, number of prior arrests, current age, criminal involvement scale, drug component scale, and vocational/educational scale.

### **Issue 3: Baird claims newer methods do not have demonstrated predictive accuracy**

Baird asserts that the newer models have weak predictive validity yet does not include published peer reviewed studies showing that they do, in fact, achieve substantial validity. Again, with the Northpointe COMPAS we have conducted several large-scale prospective validation tests over the last six years and are now building an accumulating set of peer reviewed published studies demonstrating predictive validity. These extend across multiple jurisdictions and state agencies and achieve predictive validity levels that match or exceed other well known and established models (Brennan, Dieterich & Ehret 2009; Breitenbach, Dieterich, Brennan & Fan 2009 – In press).

Additionally, the single citation to COMPAS in his paper does not relate to the basic COMPAS risk model but instead to an in-house technical report on an experimental short-term re-classification tool. This raises doubts about what version of COMPAS Baird evaluated and whether he had gathered the appropriate research literature.

### **Issue 4: Baird claims modern methods are misguided in merging risk assessment with needs assessment – he recommends separate indices of risk and needs**

On this issue, Baird again focuses on the LSI-R and Y-LSI and cites a study by Flores, Travis and Latessa (2004) to demonstrate the hazards of combining needs and risk items into a composite predictive model as used in the LSI approach. This argument is a variation of the danger of including non-predictive items into a predictive risk model – as discussed above. Baird

suggests “simply separating risk and need factors into different indices will produce better measures of both” (p.5).

Aside from the merits of his argument, he again erroneously generalizes this finding to all other “modern methods.” A review of the COMPAS chart for risk and needs, or a review of any of our other technical reports provides a direct rebuttal to this claim. Specifically, in COMPAS the needs and risk scores are clearly separated with different scoring indices and are built using different statistical procedures. When a “need” enters into any risk scale (e.g. educational failure, drug abuse) it is done explicitly if the specified need factor contributes significant incremental predictive accuracy (Brennan, Dieterich & Ehret 2009).

**Issue 5: Baird argues that the predictive models in modern risk assessments are static and sacrosanct with inadequate attempts to revise them to improve performance**

Baird accuses modern risk assessment models of being too static, rarely “evaluated” or revised (p.4) and treated as though sacrosanct (p.5). He asserts that few attempts are made to revise and “improve” their performance and views this as a “grave concern” (p.6) that applies to all modern methods. Baird again focuses on the LSI prediction model to demonstrate this point, arguing that few researchers ever attempt to “improve” the LSI predictive model.

At Northpointe we have progressively worked to upgrade and improve our risk assessment models – especially as large new prospective data sets with multi-year outcome periods and independent criterion variables are available. Such studies have now occurred in multiple state agencies (e.g. California, New York, Michigan, Georgia, and others) allowing us to systematically explore selected revisions, re-validate and improve the design, factor selection and statistical-mathematical methods of COMPAS predictive models. A series of our

publications and technical reports demonstrate these efforts and a forthcoming paper (Breitenbach, et al. 2009) illustrates our explorations with several innovative predictive methods (Gradient Descent methods, Neural Networks, Support Vector Machines, etc) that can be used as augmentations to more conventional linear regressions and survival analysis procedures. Similarly, a recent study of COMPAS predictive validation with a long term follow-up design (Brennan, et al. 2009) examined several alternative experimental models in addition to the basic COMPAS risk models across separate gender and ethnic groups, and across different offense criterion outcomes. Of 27 separate cells in this design 17 had area under the curve summary measures (AUC's) exceeding 0.70 and the remainder ranged from 0.66 to 0.69.

**Issue 6: Baird argues that the corrections field needs to rethink how the efficacy of predictive validity is measured – and that modern methods omit the most important measures**

Baird is correct in noting the confusion arising in our field from the multiple ways to measure and communicate predictive accuracy (e.g. Correlation, AUC, Sensitivity, Specificity, True Positive Proportions, Percent Correctly Classified, Relative Improvement over Chance, mean squared error - MSE, proportional reduction in error - PRE, phi coefficient, etc). He proposes that one approach using recidivism rates across risk level categories should be used as the standard for evaluating risk assessment systems. We acknowledge the value that this measure has in helping users interpret risk categories. He goes on to suggest that the most “relevant” and clear predictive performance indicators are “sometimes not discussed at all” (p.6) and cites the paper by Flores, et al. 2006) to demonstrate this point.

We agree that this awkward proliferation of methods to assess predictive performance is found in both criminal justice research (Caulkins, Cohen, Gorr, & Wei, 1996; Smith, 1996; Teplin & Swartz, 1989) and in the more general literature on predictive evaluation (e.g., medicine, weather forecasting, etc). In part, this problem results from the several different and relevant aspects of predictive performance that can be measured (Harvey, Hammond, Lusk, & Mross, 1992; Mossman, 1994). However, in criminal justice this problem also results from the serious weaknesses, criticisms and eventual abandonment of the older and most frequently used traditional procedures for evaluating predictive performance in criminal justice (which Baird seems to prefer).

The older traditional methods have been severely criticized on both practical and theoretical grounds (Copas & Loeber, 1990; Hart, Webster, & Menzies, 1993; Smith, 1996). One journal reviewer observed that the traditionally used coefficients (e.g. phi, correlations, RIOC, etc.) are “almost worthless” for the comparative evaluation of risk scales, and stated that a careful analyst may do better to simply disregard them altogether (noted in Smith, 1996, p 107). The current status of this ferment is that the AUC and Receiver Operating Characteristic (ROC) analyses have emerged as perhaps the most unbiased and best general ways to measure overall predictive performance of instruments largely because of their relative independence from base rates and selection ratios. Many of the traditional indices were quite inadequate for predictive efficiency comparisons across studies since they were not sufficiently independent of base rates and selection ratios. Thus, we agree with Baird regarding the confusing number of ways to measure predictive performance, on the difficulty of clearly communicating these to criminal justice decision makers, and on the importance of clearer communication to criminal justice policy makers.

Baird criticizes the use of the area under the receiver operating characteristic curve (AUC) as a measure of predictive accuracy. He suggests the Dispersion Index for Risk (DIFR) proposed by Silver and Banks (1998) or the index of separation (PSEP) described by Altman and Royston (2000) as better methods to evaluate the predictive performance of a risk scale. We don't think it's prudent to suggest, as Baird does, that either of these nascent measures should or could displace the AUC as the preferred measure of predictive performance in criminal justice. We do believe, as Silver and Banks (1998) recommend, that a base rate sensitive measure along the lines of the DIFR would be a useful complement to the base rate insensitive AUC.

The DIFR is a weighted composite measure of the difference between the base rate in the overall sample and the base rates in the different risk classes (for example, low, medium, and high). Naturally, because they are base rate sensitive, DIFR values cannot be fairly compared across studies. In addition the DIFR has no upper bound, and no convention exists for what constitutes an acceptable value. For example, an AUC ranges between zero and one, and a value over 0.7 is considered acceptable for the prediction of recidivism. Thus the DIFR may be appropriate for comparing different models in the same study to each other and can possibly be used to evaluate the quality of different cut-points on a risk scale.

The PSEP is simply the separation (in probability) between the survival rates in the worst and best risk classes. While this measure has the advantage of being bounded it unfortunately ignores proportionality. This can lead to cases where the cuts may be chosen in a trivial manner, for example, all cases are assigned to the high group as this maximizes the spread between the high and low risk group (if all are in the high risk group, the probability for the low risk group is zero hence maximizing the spread between the two).

The AUC provides an overall summary of performance across all the thresholds of a scale. It facilitates comparisons across studies using the same or different risk scale, and it can be used to guide decisions for the placement of cut-points on a scale. Pepe (2003) described ROC analysis as "currently the

best-developed statistical tool for describing the performance” of ordinal and continuous measures of risk (p. 66). The AUC also has its limitations. Its use as a global measure of predictive accuracy has been criticized because it is based on the summary of performance across all decision thresholds, including those cuts that have no skill. In addition, an AUC above the convention of 0.7 may be obtained for scales that don’t perform well (for example, scales that don’t have good dispersion). Finally, as Silver and Banks argue, ROC analysis is best-suited to two-decision problems, such as release or don’t release.

At Northpointe we present findings using many of the practices Baird recommends. For example, our technical reports to clients provide several aspects of predictive performance relevant for criminal justice decision makers. First, overall accuracy is defensibly summarized by the AUC coefficient. Second, if clients are interested in survival estimates we provide survival analysis graphs to illustrate different risk category failures across time. Third, with logistic regression analyses we compute odds ratios to statistically test different failure rates across risk categories. We also produce ROC charts and related rates of false positive and false negative errors at all levels of the risk scales. These can help agencies select cut scores that balance true and false positive errors when determining cuts for supervision levels. We also provide the simple criterion that Baird recommends, i.e., percentage failure rates across multiple classification levels from low, medium and higher risk categories.

We typically include survival analysis to evaluate the predictive power of each specific need and risk factor as well as for the overall risk levels from our predictive models. Survival analysis is a well established statistical method that offers a clear and accurate representation of an instruments performance. Indeed we find in all our studies that higher risk offenders fail earlier and more often than lower risk offenders. In these studies we measure the concordance index (Harrell et al. 1982), a measure essentially equivalent to an AUC for survival models, which specifies the probability of a lower risk offender failing before a higher risk offender. We measure the concordance index on both the continuous scales and on discrete risk levels. This measure, like the AUC, ranges between zero and

one (bounded), is base-rate independent and can be meaningfully compared across studies to indicate comparative efficiency of different instruments. A key procedural issue in conducting survival analysis is the necessity to control for competing risks. For example, an offender might be temporarily returned to prison for some time for a technical violation. During such a temporary stay in prison he cannot recidivate and is thus “not at risk,” hence these temporary time periods must be gapped out in the analysis. While such studies require more effort, they lead to more insights on the performance of a predictive risk model by going beyond a simple yes/no binary risk estimation to introduce the time dimension.

**Issue 7: Baird maintains that for risk assessment models reliability tests and internal consistency are counterproductive and should not be used**

This issue is widely known to most developers of assessment scales and predictive models. Baird correctly points out that internal reliability tests such as Cronbach’s alpha are inappropriate in regard to risk assessment models that are often explicitly multidimensional. He then argues that Cronbach’s alpha is “ideal” (p. 8) for assessing the internal consistency form of reliability for psychological and need constructs such as depression – and presumably other needs and psychological constructs.

Predictive models do not attempt to build one-dimensional scales of a single construct, but aim to optimize predictive accuracy usually by including multiple underlying dimensions, each with separate contributions to predictive power (Brennan, Dieterich, & Ehret 2009). A review of most of Northpointe’s technical reports and peer-reviewed papers indicates that they typically omit Cronbach’s alpha for predictive risk models. However, in evaluating need and psychological scales we do apply Cronbach’s alpha – which Baird acknowledges is “ideal” and appropriate for assessing internal reliability. While on this topic it

may be noted that most of our need scales have Cronbach's alpha scores exceeding 0.70 and thus in the acceptable reliability range. In this case, Northpointe already practices the recommended approach.

### **Issue 8: Baird maintains that modern methods fail to appropriately address inter-rater reliability**

Baird also raises the important issue of inter-rater reliability and criticizes modern correctional risk and needs assessments for weaknesses in this area. Reliability and consistency of staff classification and assessment decisions are the basis for equity, fairness, efficient communication and several other key issues. While the basic question of whether two raters will reach the same score for a particular individual appears simple, the topic is actually quite complex and there are several different formulations of reliability. In terms of methods to assess consistency across "raters" Baird mentions the Kappa coefficient and its particular benefit in correcting chance agreement between raters. It is also true that many modern methods utilize semi-structured and motivational interviewing (MI) and both of these are vulnerable to reliability problems since they require high levels of staff inference, intensive training and high skill levels on the part of the interviewer. If staff skills are deficient then such approaches can be plagued by unreliability.

In considering the problem of low inter-rater reliability it is important to understand the general context across all of the social sciences. Recent reviews reveal that even among trained mental health professionals the consistency of agreement on classification diagnostic decisions is often poor to modest and high reliability is often difficult to achieve. Wood, et al. (2002), for example, in a broad national review reported that across a variety of diagnostic categories and

psychological testing procedures kappa coefficients range from poor ( $K = 0.20 - 0.35$ ); to fair ( $K = 0.40 - 0.55$ ); while on some studies a kappa of 0.61 is hailed as substantial and acceptable (see also Garb 1998). In general, highly structured and rule-based instruments will improve inter-rater reliability. In addition to the design and administrative features of an assessment tool, organizational factors also powerfully impact inter-rater reliability. Specifically, in correctional agencies the levels of staff training, staff competence, supervisory competence, work overload, workload stress, caseload sizes, and so forth, all profoundly impact inter-rater reliability. Even a highly reliable and structured assessment tool thus may be undermined and used inconsistently and with poor reliability within an unfavorable organizational context. Thus, the onus for reliability does not fall on the technical design of the risk and needs instrument alone.

Baird's first criticism is that inter-rater reliability is a particular concern for those assessment methods that require (or allow) repeated subjective decisions and clinical inference by staff in the assessment process. Baird then notes (p.7) that both the LSI and YASI rely on semi-structured interviewing and that this inevitably requires much subjective judgment and clinical inferences by staff, making these tools especially vulnerable to reliability problems. In support of his argument, he cites the study by Austin, et al. (2003) that found "serious difficulties" and reliability problems with the LSI. In this regard, however, we note that the CMC component of the NCCD system also relies on a semi-structured interviewing process (Harris, 1994; Hardyman, 2002) thus making this instrument similarly vulnerable to inter-rater reliability problems.

To minimize inter-rater reliability problems in COMPAS we used several strategies: 1) Data collection methods that minimize or exclude clinical inference and subjectivity by staff. This follows a recommendation by Austin, et al. (2003) for simple methods to minimize staff subjectivity and inference. 2) Mathematical-statistical methods to actually replace or augment human judgment for

classification decisions (where possible). These two approaches are now briefly discussed.

*Automated classifications and reliability:* An extensive body of research across half a century in psychological diagnosis (Grove et al. 2000) has indicated that quantitative methods for diagnostic classification decisions are largely superior to clinical judgment and decision-making. In fact, Quinsey et al 1996, in a respected review of prediction of criminal violence forcefully suggested that actuarial and mathematical methods for classification assignment should be used instead of human clinical judgment. We realize that Quinsey et al's position is controversial and we do not adopt such a strong stance. However, we view our automated and actuarial classification decisions as a "decision support" to staff that can be overridden when staff can provide strong and reasonable justifications with supervisory review.

In COMPAS we use quantitative pattern matching methods to automatically assign offenders to classification categories for both risk levels and for a separate treatment typology. The treatment-explanatory typology is similar in spirit to the classic explanatory-treatment typologies of the I-level (Warren 1971), Megargee's MMPI Typology (Megargee & Bohn 1970) and also to Baird's CMC system. However, it uses more contemporary pattern recognition and quantitative methods in constructing and validating the typology as described in a recent paper by Brennan, Dieterich and Breitenbach (2008). In our reliability studies of classification consistency we use a method specifically designed by McIntyre-Blashfield (1980) to assess the consistency of classification assignments – again using Kappa as a measure of reliability. In these studies the automated pattern matching algorithms achieve Kappa Coefficients ranging from 0.65 to 0.85. These clearly fall in the acceptable to excellent range. It is interesting to note that Kappa coefficients in the Diagnostic and Statistical Manual (DSM)-III of 0.60 and above were regarded with great joy by the

psychiatric community during the reformulation of the DSM and were used to recover the integrity and viability of their discipline (Kirk and Kutchins 1986; Beutler and Malik 2002).

*Data collection methods to improve reliability:* In our data collection strategy we attempt to minimize staff subjectivity and inferences. Specifically, the first third of COMPAS questions are collected from official criminal records – which minimizes staff subjectivity and allows for supervisory verification. Another third of our questions consist of a self-report checklist which does not require a staff rater. Wood, et al. (2002) in their Annual Review of Psychology paper on assessment commented on the strength and viability of self-reports and their treatment utility. However, in this self-report section – since we are working in a correctional environment - we include data verification tests for “faking-good” and “coherency of responses.” We agree with Wood et al (2002) that these tests are important in correctional settings. In the COMPAS software these data verification scales trigger automated warnings to staff when problems are detected. However, about another one-third of the COMPAS instrument requires an interviewer to read aloud a sequence of standardized scripted questions (with fixed response formats) with no comment by the interviewer, except to explain the meaning of a question (as needed). Such standardized administration is widely used in social sciences to minimize rater inference, biases, to obviate training and skill differences among staff and achieve higher reliability. We have also developed a semi-structured interview approach for this section.

We acknowledge that the search to improve inter-rater, internal consistency and other forms of reliability is a constant challenge and that even the above procedures cannot totally negate this issue. This is particularly the case in the real world of large scale correctional agencies and more work can be done in this area – for all current instruments including our own. Baird was entirely right in his decision to place this issue on the table. However, the

pervasive challenge of limited correctional organizational resources will perhaps be a consistent limiting factor in corrections and will impact staff skills, contribute to work overload and impose time constraints for the assessment process.

## **Conclusion**

In this paper we offer a series of objections to many of the critical positions that Baird has taken regarding current developments in risk and needs assessment in corrections. While we agree with Baird that these issues are important, we find that many of the positions Baird has taken are a result of several over-generalization errors. This generalization ignores the multiple differences that exist between the various methods. In other instances it seems that Baird did not have access to important and current technical details of certain methods and criticizes these methods for design flaws that, in fact, had been correctly addressed. In effect, these criticisms are unfounded once the full details are known. Many of Baird's criticisms are aimed specifically at design flaws in the LSI, which he then more implicitly attributes to other systems.

We also note that our present paper – while addressing several of Baird's issues – has not attempted to address all of his positions regarding modern risk assessments. Specifically, we have not addressed his criticisms of the concepts of “dynamic and static factors”, “criminogenic needs” and “protective factors”. These are largely central to the “What Works” paradigm that has become dominant in the last decade and is strongly associated with the work of Andrews and Bonta (1995), the main developers of the LSI, and their various Risk, Need and Responsivity (RNR) principles. These conceptual issues require a careful analysis that is beyond the scope of the present paper.

In the interest of advancing knowledge of effective approaches, Northpointe partners with practitioners to share ideas and to create environments where learning is a two-way street. Practitioners learn the results of scientific experiments to assess the impacts of their work. In turn, researchers learn the challenges inherent in criminal justice practice and come to understand the challenges of producing evidence in these settings and the humility that is required to interpret findings. These lessons help ensure that the production and use of knowledge is done with mutual accountability and respect. In the corrections field today, the partnership between research and practice is perhaps the most promising process for improving our understanding of problem behaviors and for building systemic solutions to address the risk and needs of the people served by these systems.

## References

- Altman, D., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19, 453-473.
- Andrews, D.A., & Bonta, J.L. (1995) *Level of Service Inventory – Revised*. Toronto, Ontario: Multi-Health Systems.
- 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.
- Beutler, L.E., & Malik, M. (2002) *Rethinking the DSN: A psychological perspective*. Washington D.C: American Psychological Association
- Brennan, T., Breitenbach, M., & Dieterich, W. (2008). Towards an explanatory taxonomy of adolescent delinquents: Identifying several social-psychological profiles. *Journal of Quantitative Criminology*, 24, 179-203.
- Brennan, T., Dieterich, W., & Ehret, B. (2008). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36, 21 – 40.
- Breitenbach, M., Dieterich, W., Brennan, T., & Fan, A. (In press). *Creating risk-scores in very imbalanced datasets – Predicting extremely violent crime among criminal offenders following release from prison*. Chapter in forthcoming book.
- Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting criminal recidivism: A comparison of neural network models with statistical methods. *Journal of Criminal Justice*, 24(3), 227-240.
- Copas, J. B., & Loeber, R. (1990). Relative improvement over chance (RIOCI) for 2x2 tables. *British Journal of Mathematical and Statistical Psychology*, 43, 293-307.
- Flores, A. W., Lowenkamp, C. T., Smith, P., & Latessa, E. J. (2006). Validating the Level of Service Inventory-Revised on a sample of federal probationers. *Federal Probation*, 70 (2), 44-78.

- Hardyman, P.L., Austin, J., Alexander, J., & Johnson, K.D. (2004). *Internal prison classification systems: Case studies in their development and implementation*. Washington, DC: National Institute of Corrections.
- Harrell, F.E. (2001) *Regression Modeling Strategies*. New York:Springer.
- Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., & Rosati, R.A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, *247*, 2543-2546.
- Harris, P. M. (1994). Client management classification and prediction of probation outcome. *Crime and Delinquency*, *40*, 154-174.
- Hart, S. D.; Webster, C. D., & Menzies, R. J. (1993). A note on portraying the accuracy of violence predictions. *Law and Human Behavior*, *17*, 695-700.
- Harvey, L.O., Hammond, K. R., Lusk, C. M., & Mross, E. F. (1992). The Application of Signal Detection Theory to Weather Forecasting Behavior. *Monthly Weather Review*, *120*, 863–883.
- Kirk, S. A., & Kutchins, H. (1986). *The selling of the DSM: The rhetoric of science in psychiatry*. New York: DeGruyter Publications.
- McIntyre, R.M., & Blashfield, R.K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research*, *15*. 225–238.
- Megargee, E. & Bohn, M (1970) *Classifying criminal offenders: A new system based on the MMPI*. Sage, Beverly Hills, California.
- Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783-792.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.
- Quinsey, V. L.; Harris, G. T.; Rice, M. E., & Cormier, C. A. *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association,
- Silver, E., & Banks, S. (1998). Calibrating the potency of violence risk classification models: The Dispersion Index for Risk (DIFR). Washington, DC: *American Society of Criminology*.

- Smith, W. R. (1996). The effects of base rate and cutoff point choice on commonly used measures of association and accuracy in recidivism research. *Journal of Quantitative Criminology*, 12(1), 83-111.
- Teplin, L. A., & Swartz, J. (1989). Screening for severe mental disorder in jails: The development of the Referral Decision Scale. *Law & Human Behavior*, 13(1), 1-18.
- Warren, M. Q. (1971). Classification of offenders as an aid to efficient management and effective treatment. *Journal of Criminal Law, Criminology, and Police Science*, 62, 239-258.
- Wood, J. M., Garb, H. N., Lilienfeld, S.O., & Nezworski, M. T. (2002). Clinical Assessment. *Annual Review of Psychology*, 53, 519-543.