



**A Response to**

**“Assessment of Evidence on the Quality of the  
Correctional Offender Management Profiling for  
Alternative Sanctions (COMPAS)”**

(Skeem & Loudon, 2007)

By:

Tim Brennan, PhD

Bill Dieterich, PhD

Markus Breitenbach, PhD

Brian Mattson, PhD

June 2009

## Introduction

In this paper we review and respond to the report “Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions [COMPAS]” (Skeem & Louden, 2007). In this report, Skeem and Louden review the predictive validity, construct validity, and reliability of COMPAS. An important preliminary is that Skeem and Louden only had access to a very limited portion of the available research reports on COMPAS at the time of writing their review. The absence of several longer-term predictive validation studies and peer-reviewed papers was unfortunate given that these studies address several of their central concerns. It appears that many of their conclusions were based on a small-scale 2002 study that focused only on the initial development of predictive models. This specific study was, in fact, part of a much more intensive program of research aimed at the development, improvement and validation of the COMPAS models. In fairness, Skeem and Louden were aware of this limited and incomplete evidence base. They acknowledge that their evaluation was based on a limited set of reports and that readers should interpret their report “with caution” (p.4). Thus, in this response our intention is to update the discussion on the reliability and validity of COMPAS and introduce more current research evidence for COMPAS as well as to address some of the issues raised by Skeem and Louden.

We acknowledge at the outset that most of the evidence for the reliability and validity of COMPAS is found in the results of in-house research studies conducted by Northpointe across a variety of jurisdictions and states. We know that critics may discount this research. However, much of our in-house research is conducted for state agencies. In many cases, competent research divisions within those agencies scrutinize the methods and results very closely. These state-sponsored studies are often subjected to a more thorough vetting than that provided by the editors of peer-reviewed journals since internal research staff has full access to the data and can replicate our analyses, initiate new queries, and require additional verification analyses. We recognize from a scientific standpoint that independent research

evidence and peer-review for the reliability and validity of COMPAS will bolster its standing in the marketplace. Thus, we encourage our clients to form collaborative relationships with independent researchers to pursue independent research opportunities and conduct well-designed validation studies. However, as noted above, peer reviewed papers on reliability and validity issues for COMPAS are also now published (Brennan, Dieterich and Ehret, 2009; Breitenbach, Dieterich, Brennan and Fan, 2009 – In Press). These deal directly with several of the central themes raised by Skeem and Louden and demonstrate that the COMPAS system reaches accuracy levels comparable to, and in some cases better than, most of the current major risk prediction models (e.g. LSI-R). We also note that at least two different university based teams of researchers are currently conducting independent evaluations of COMPAS in different state agencies.

At Northpointe we have an established history of working in partnership with our clients to advance knowledge and practice in the Criminal Justice field. From our early work in jail classification to our recent partnership with the California Department of Corrections and Rehabilitation (CDCR) and the University of Cincinnati, Northpointe leverages the opportunity of public and private partnership to expeditiously test and advance knowledge. The results are shared both in writing and through presentations with others in the field. Findings from our research are also shared with public domain assessment efforts and advance the availability of current information for use in practice. The discussion in this report focuses primarily on the issues of predictive validity, construct validity and the relationship between COMPAS needs scales and outcomes.

From our perspective, while the Skeem and Louden report takes an appropriate look at a broad range of measurement and design issues, its limitation to a small set of early studies is clearly problematic in that it under-represents the scope and depth of our validation work. We also recognize that validation of any assessment instrument is a multi-faceted process and is essentially continuous. Thus, in the last few years our studies have been repeatedly replicated, extended and updated across several large criminal justice agencies, with the benefits of new

and larger samples, multiple dependent variables, alternative statistical predictive methods and longer outcome periods. Our on-going research and development work continues to address many of the issues that Skeem and Loudon raised, and has produced a substantially larger base of empirical evidence, statewide reports, and peer-reviewed papers than was available when their review was written. We will now focus on clarifying the more current and up-to-date evidence of validity and reliability of the COMPAS system.

## **Overall Approach to Validity**

Skeem and Loudon (p.14) mention the tripartite framework offered by Pedhazer & Schmelkin, (1991) for overall construct validation, with an implication that we at Northpointe should follow this approach to scale development and validation. Their approach consists of three broad phases: 1) logical analysis, 2) internal structure analysis, and 3) cross-structure analysis. This construct validation approach is essentially similar to the approach we have used and is described in many of our validation documents. Specifically, we closely followed the validation approach of Millon (1997) with three broad phases and sub-tasks similar to the Pedhazer model:

1) Theoretical-substantive – in this phase of item and scale selection we are strongly guided by current meta-analytic findings in criminal justice regarding the most promising criminogenic factors for predicting recidivism; as well as by extant criminological theory e.g. strain theory, social learning, social control theory, etc.

2) Internal-structural – this empirical phase is dominated by item analysis, internal reliability studies, factor analyses, unidimensionality of scales and examination of potential higher order factors.

3) External criterion – in this phase we mainly focus on prediction of external criterion variables, discrimination of external criterion groups and taxonomic studies of the criminal population. Criterion related validation studies are an on-going focus of this phase.

Thus, in our overall program of scale development and validation we have followed a well known paradigm with a marked similarity to the recommended model proposed by Skeem and

Louden. In terms of overall methods to scale development and validation we have no quarrel with this general approach.

## **Predictive Validity**

COMPAS clearly distinguishes between risk scales (designed to predict recidivism) and needs scales (designed to measure needs and used to inform case plans and identify intervention targets). Our approach of separating risk and needs aligns with current best practices in risk assessment (Baird, 2009; Gottfredson & Moriarty, 2006). Regarding this issue we may have a methodological disagreement with Skeem and Loudon. An interesting recent controversy is pinpointed by the suggestion by Skeem and Loudon that one should combine “all the criminogenic needs and static risk factors ... into a single, total score that would predict recidivism” (p.29) as is the practice of, for example, the LSI-R risk model. However, this practice has drawn considerable criticism for its potential to include irrelevant factors into a risk model (Baird 2009). In his widely distributed paper Baird criticizes the LSI for its simple additive summation of all 54 items (which includes both risk and need items) to create its overall risk model. As is well known to statisticians, this practice may allow many low-predictive or even irrelevant factors to enter a predictive model. Such “noise” variables may 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).

Ultimately, we suggest that the issue of how best to combine risk and needs scales, without introducing irrelevant factors, will be resolved through empirical verification, careful analysis and use of appropriate statistical modeling approaches. We believe that if a need scale has incremental validity and can add some accuracy to a predictive model then it should be included within a risk model.

Skeem and Loudon reference the risk principle in their discussion of risk prediction and needs assessment. Their report uses the terms risk status (relative risk of recidivism) and risk

state (intra-individual dynamic risk of recidivism). We think this is a useful distinction. The suggestion is that risk scales particularly designed to predict risk state should be dynamic (composed of dynamic, criminogenic needs) so that one can measure changes in risk of recidivism over time for specific individuals. The question for research is then to discover those dynamic criminogenic needs that are most relevant for risk prediction (either state or status) and include them in appropriate risk models. In some cases, static measures that have an association with outcomes may not be useful for practical risk prediction models depending on the purpose and context of the decision.

COMPAS has two main risk models: General Recidivism Risk and Violent Recidivism Risk. The Recidivism Risk Scale is an equation originally derived from a regression model that was developed in a sample of pre-sentence investigation and probation intake cases in 2002. It was trained, in that study, to predict any offense arrest within two years of intake assessment. This was the early 2002 document (“Evaluation of reliability and validity of the COMPAS scales: New York Probation Sample”) that Skeem & Loudon used in their critique. It describes the methods used to initially construct our General Recidivism Risk Scale. Unfortunately, and perhaps understandably, it seems that Skeem & Loudon assumed that these few reports represented the totality of our research and that no further work had been done to validate the COMPAS predictive models. We acknowledge that the 2002 study was of relatively small scale and that good practice would require further replications. This is exactly what has occurred, and since 2002 several follow-up validation studies have been completed, peer reviewed and published, with several other studies in preparation for publishing.

We turn now to present some relevant results of our broader program of research pertaining to predictive validation and measurement issues. The Violent Recidivism Risk Scale is also an equation derived from a regression model that was initially developed in 2006 in a sample of presentence investigation and probation intake cases and later validated on prison samples. It was trained to predict violent offenses (misdemeanor and felony) within two years of the time at risk following an assessment. This Violent Recidivism Risk Scale replaced the

original Violence Risk Scale that Skeem and Louden refer to in their report. This newer Violence Recidivism Risk Scale has also now been re-validated in new independent samples from several geographical regions since it was first developed. These studies also address several concerns raised in Skeem and Louden’s report regarding predictive validity. One of the major criticisms was the issue of criterion contamination (i.e. when predictor and criterion variables are not cleanly independent). We strongly reject this criticism and point out that most of our “dependent variables” are new criminal arrests and offences collected quite independently of COMPAS from official state criminal history sources. These criterion variables are new offences collected at a different time and from a different source than the COMPAS predictor variables. A review of the descriptions of criterion variables and predictor variables in our technical reports should clarify this independence. We suspect that this criticism was perhaps a misunderstanding related to a single table in our 2002 report in which – as an experiment – we computed provisional area under the curve (AUC) levels using several diverse offenses as criterion outcomes. This was a minor exploratory data analysis experiment that had no implications for the overall design or for the main results of the 2002 study. It could be eliminated and the results of the study would remain unchanged.

COMPAS also includes a Failure to Appear Risk Index that is used by a limited number of clients for pretrial release cases. We are currently conducting an additional Failure to Appear (FTA) outcomes study for New York Probation to test the predictive validity of the FTA Risk Index in sample of 1,000 pretrial release cases.

While the Skeem and Louden evaluation did not reflect the full breadth and scope of our overall validation research program we stress that Northpointe is committed to vigilantly testing, evaluating, and improving our risk models. During the initial phases of any predictive scale development we typically examine several alternative statistical methods for building predictive models (e.g. logistic regression, survival analysis, random-forest and tree-based methods, etc). Since we often work collaboratively with our clients, we openly discuss the selection of outcome criterion variables to ensure a good fit with their operational needs. We

then incorporate a variety of validation procedures and (in some cases) multiple independent criterion variables to evaluate the criterion validity of the risk models. We typically collect several well-known criterion outcome variables such as age-at-first, total prior violent felony convictions and parole revocations, returns to prison, and so forth, chosen to fit the client agencies needs and to ensure that fundamental associations are present. In this context Skeem and Louden appropriately raise a concern for the possibility of “over-fitting” whenever a predictive model is modified or revised on a given sample. We emphasize that we follow standard procedures to address this issue, and have systematically addressed the need for follow up validation samples and outcomes for any new or updated model. Additionally, where any minor modifications to a model have been introduced we have produced specific technical documentation for our clients and other users. Thus, our General Recidivism Risk and Violent Recidivism Risk scales have been recurrently validated using multi-year prospective outcome studies in new samples as well as for different racial/ethnic and gender groups across several different state systems (Brennan, Dieterich and Ehret 2009). Much of this work was done in the time period following the 2002 study that was the focus for many of Skeem and Louden’s comments.

When implementing COMPAS in a new jurisdiction, our general practice is to incorporate an outcomes study component with at least a year of follow-up for an initial analysis. This is done in a new pilot test to locally evaluate the predictive validity of the risk scales. This cross validation strategy follows a recommendation from Wright, Clear and Dickson (1984) following their finding that a widely used NIC model did not generalize across different jurisdictions. In 2006 we conducted three new pilot tests in the New York Division of Probation and Correctional Alternatives (DPCA), the New York State Division of Parole (NYSDP), and the Michigan Department of Corrections (MDOC). These three pilots all had outcomes studies with multiple follow-up times built into them. In 2008 we conducted additional, more extended outcomes studies at all three sites with longer-term outcomes. We also conducted separate studies in the California Department of Corrections and Rehabilitation (CDCR) and for New York’s DPCA (Brennan, Dieterich, & Ehret, 2009).

Listed below are the summarized results of several follow-up studies of the predictive validity of the two main COMPAS risk assessments conducted in the last two years. These outcome studies report the (AUC) for the General Recidivism Risk and Violent Recidivism Risk scales. The AUC is the most widely used measure of predictive accuracy in criminal justice, psychology, medicine, and related fields. An AUC of .65 to .69 indicates modest to moderate predictive accuracy while an AUC of .70 to .75 and higher indicates moderate to strong predictive accuracy. We note that the available criminal justice risk prediction studies suggest that AUC's for most current risk assessment systems typically range from 0.64 to 0.77 (Flores et al 2006, Brennan, Dieterich and Ehret 2009; Manchak et al 2008, Manchak et al 2009). In addition to the AUC, in our technical reports we also evaluate our risk scales using other scientific criteria, including failure probabilities, odds ratios, and hazard ratios.

Michigan Department of Corrections (n=561)

<u>Outcome</u>	<u>AUC</u>
Any Arrest	.703
Felony Person	.699
Abscond	.660
Return to Prison	.702
	.686

New York State Division of Parole (n=553)

<u>Outcome</u>	<u>AUC</u>
Any Arrest	.679
Felony Person	.630
Abscond	.728
	.652

Division of Probation and Correctional Alternatives—Pilot (n=987)

<u>Outcome</u>	<u>AUC</u>
Any Arrest	.730
Person	.730

Division of Probation and Correctional Alternatives—Study (n=2,328)

<u>Outcome</u>	<u>AUC</u>
Any Arrest	.707
Felony Person	.717
	.742

California Department of Corrections and Rehabilitation (n=20,890)

<u>Outcome</u>	<u>AUC</u>
Return to Prison	.672

Note: For felony arrest, abscond, and return outcomes, the Recidivism Risk Scale is tested. For person arrest the Violent Recidivism Risk Scale is tested.

## On the Issue of Cumulative Improvement of Predictive Models

We also differ with Skeem and Loudon regarding a constraint they appear to impose on the evaluation of models in a context of progressive refinement and improvement. Specifically, they recommend – apparently on the basis of our 2002 report – that any statistical evaluations of COMPAS be restricted to our original models. They write (p.4): “We strongly recommend that UCLA investigators evaluate the utility of the *existing* COMPAS scales in predicting recidivism.” Our position is that in an on-going program of research, in a context in which client agencies desire improvements, and when multiple data sets are generated for the same predictive models, with access to potentially useful additional predictors and with long term outcomes, this situation offers a useful opportunity to explore further improvements and potential revisions. Such improvements may pertain to predictive factors or to the possibility of alternative mathematical approaches. A major example of this was our decision to upgrade our violence risk prediction scale in 2006. We suggest that such opportunities should be used for further revision and updates to scales. However, and here we agree with the cautionary note from Skeem and Loudon, an important proviso is that we must take care to impose procedures to avoid “over-fitting” of any new or modified model and a modified model must be tested on new samples with appropriate cross-validation. As noted elsewhere in this report we have taken great care to minimize or avoid problems of over-fitting.

To restrict any new evaluation study to the 2002 models after a seven year gap and when the models have already been upgraded would be to ignore or deny all of our on-going work to revise and improve models. It would seem more appropriate to focus new evaluations on the current release. We acknowledge that a wish for “stability” of any predictive models will

run counter to a desire for on-going improvement – and that this dilemma is not without controversy. For example, Baird recently criticized most of the current widely used criminal justice risk assessment models for being too static, rarely “evaluated,” revised or improved (p.4), too rigid and being treated as though sacrosanct (p.5). He complains that almost no attempts are made to revise and “improve” the performance of most current risk assessment models and views this as a “grave concern” (p.6). Baird primarily focused on the LSI prediction model to demonstrate this point, arguing that few researchers ever attempt to “improve” the LSI predictive model.

Our strategy at Northpointe is that when the appropriate concerns with over-fitting are addressed by replications on appropriate independent samples we will use such an opportunity to progressively upgrade and improve our risk assessment models where appropriate. This most often occurs when several large prospective data sets with multi-year outcome periods and independent criterion variables are available across multiple sites. These can facilitate a systematic exploration of selected revisions, re-validations and improvements to the design, factor selection and statistical-mathematical methods of COMPAS predictive models. This is not done cavalierly and we last introduced major upgrades in our 2006 work. One recent paper (Breitenbach, et al. 2009) demonstrates our explorations regarding several innovative mathematical predictive models, e.g. Gradient Descent methods, Neural Networks and Support Vector Machines in a comparison to standard models such as logistic regression and survival analyses. Similarly, a recent study of COMPAS predictive validation with a long term follow-up design (Brennan, et al. 2009) examined two alternative models in addition to the basic COMPAS risk models across diverse gender and ethnic groups, for several different offense criterion outcomes. Of 27 separate cells in this design 17 had AUC summary measures exceeding 0.70 with the remainder ranging from 0.66 to 0.69.

### **Construct Validity**

Turning to construct validity we agree with Skeem and Louden that this is relevant for all correctional instruments. This issue clearly applies to the COMPAS needs scales that attempt to measure a single construct, typically constructed as a uni-dimensional scale. The Recidivism Risk and Violent Recidivism Risk scales, in contrast, are regression models developed to predict recidivism. These were constructed to optimize predictive accuracy and not necessarily to measure a single dimensional construct. Unidimensionality and factor structure are not important or relevant evaluative criterion for such regression based risk models.

Skeem and Louden mention several approaches to validity including concurrent and discriminant validity. A key aspect of most forms of validity including both concurrent and construct validity is to cumulatively establish examples where the observed correlations between measures are in the expected theoretical direction, and high correlations are achieved between measures of the same construct. For example, the COMPAS substance abuse measure correlates positively (.44) with the Substance Abuse Subtle Screening Inventory (SASSI) in our MDOC pilot sample. However, construct validity, in particular, is cumulatively established when a measure is found to correlate in the predicted manner with a range of other variables with which it theoretically should correlate. With each new study conducted with COMPAS we are able to add additional findings to this cumulative process.

As one example, research in developmental delinquency (longitudinal research in which anti-social behaviors and attitudes are studied over the life course) consistently finds that youth with early onset of delinquent behavior tend to have more serious delinquency trajectories and more negative emotionality, lower achievement, and problems in social adjustment (Moffit, 2003). Thus, when we consistently find, over multiple studies, that our Criminal Personality, Criminal Attitudes, Social Adjustment and Vocational Educational scales correlate with age-at-first-arrest, just as developmental delinquency research predicts, this adds supporting evidence of COMPAS construct validity. Age-at-first-arrest offers an established and useful external variable to add supporting evidence for the construct validity of the COMPAS needs scales. Additionally, we point out that although age-at-first is located inside the COMPAS system, it is

collected from official records, while the needs scales are scored using a different method (interview and self-report) which negates the danger of criterion contamination or method variance.

While the above correlations with age-at-first-arrest offer only one example of the kind of evidenced that supports construct validity, we are gradually building an accumulating range of evidence of this type to support construct validity from several psychometric studies, including the Michigan Department of Corrections, New York Probation, New York Parole, Georgia Department of Corrections and other sites. While many examples could be given, we may illustrate one approach to demonstrating construct validity using results from a current sample in CDCR in Table 1. This CDCR sample consists of 6,485 Core COMPAS assessments conducted between September 26, 2008 and January 27, 2009. Men comprise 91% of the sample.

**Table 1: Correlations of COMPAS Scales with Criminal History Indicators in CDCR**

	Age-at-First	Prior Arrests	Returns to Custody	Commitments	Assaultive Misconduct
CassPeer	-0.28	-.13	0.17	0.09	0.18
SubAbuse	-0.05	0.23	0.19	0.16	-0.07
Financ	-0.07	0.10	0.11	0.08	0.02
VocEd	-0.22	0.11	0.14	0.06	0.17
FamCrim	-0.19	0.09	0.10	0.05	0.11
SocEnv	-0.18	0.11	0.11	0.10	0.14
Leisure	-0.09	0.10	0.11	0.08	0.08
ResInst	-0.03	0.12	0.15	0.10	0.10
SocAdj	0.02	0.18	0.19	0.12	0.14
SocIsolation	0.04	0.11	0.13	0.10	0.06
CrimAttC	-0.12	0.03	0.05	0.00	0.13
CrimPers	-0.15	0.09	0.13	0.06	0.17

While most of the correlations in these tables are modest, they all reach statistical significance and are largely similar to those found in other published studies using criminal justice samples. It is important to realize that such attenuation is common when using relatively

homogeneous offender and prisoner samples in examining correlations between risk factors and criminal involvement criterion variables.

There are some notable correlation patterns in Table 1 that offer additional evidence of construct validity for the COMPAS scales. For example, we see that age-at-first arrest correlates negatively with the higher-order personality scales Criminal Attitudes ( $p = -.12$ ) and Criminal Personality ( $p = -.15$ ). This comports with findings in developmental research that indicate offenders with early onset are more likely to have high scores on similar personality measures and with serious and persistent criminal involvement (Moffitt, 1993). Similarly, offenders with earlier age-at-first arrest are more likely to have higher scores on scales measuring factors identified as criminogenic in longitudinal developmental studies. These scales include Criminal Associates and Peers ( $p = -.28$ ), Family Crime ( $p = -.19$ ), Vocational/Educational Problems ( $p = -.22$ ), and Social Environment ( $p = -.18$ ) (Farrington, Jolliffe, Loeber, Stouthamer-Loeber, & Kalb, 2001). Again, these correlations are of similar magnitude to those emerging in such studies.

A further pattern in Table 1 is defined by the correlations between the total number of previous arrests (official data) and the scales Substance Use ( $p = .23$ ), Financial Problems ( $p = 0.1$ ), Residential Instability ( $p = .12$ ) and Social Isolation ( $p = .11$ ) (Stouthamer- Loeber, Loeber, Wei, Farrington, & Wikstrom, 2002).

There are additional moderate but significant correlations between assaultive misconduct and the COMPAS scales of Criminal Associates and Peers ( $p = .18$ ), Vocational Educational Problems ( $p = .17$ ), Social Environment ( $p = .14$ ), Social Adjustment ( $p = .14$ ), and Criminal Personality ( $p = .17$ ). In their meta-analysis, Gendreau, Goggin, and Law (1997) report similar findings showing that antisocial attitudes and criminal peers were important individual level predictors of prison misconduct. Again, consistent with prior published research we find significant correlations between the number of returns to custody for a parole violation and the scales Criminal Associates and Peers ( $p = .17$ ), Substance Abuse ( $p = .19$ ), Vocational Educational Problems ( $p = .14$ ), Residential Instability ( $p = .15$ ), and Social Adjustment ( $p = .19$ ).

These findings are again consistent with prior research to identify the most important risk and needs factors associated with reentry failure and recidivism (Nelson, Deess, & Allen, 1999; Herrenkohl et al., 2000; Tolan & Gorman-Smith, 1998).

Skeem and Louden may discount these findings and attribute them to “method variance.” However, all of our criminal history and institutional disciplinary criterion variables are obtained independently from official data sources and thus these significant correlations cannot be attributed to method variance.

However, we agree with Skeem and Louden on the importance of cumulatively establishing a research base for the various kinds of validity of scales. A related component of our recent research is our efforts to build additional data on the correlations between COMPAS and other multiple factor instruments. For example, our current collaborative work in two different states, with the University of Cincinnati on a battery of “Gender-Sensitive” measures has allowed a large number of such correlations to be examined for construct validity implications. This data has allowed us to initiate an examination of convergent and discriminate validity in the context of a multi-method, multi-trait matrix framework. This follows a helpful suggestion by Skeem and Louden to conduct this approach to validation. The preliminary findings are very encouraging.

## **Validity of COMPAS Needs Scales**

Skeem and Louden, in agreement with Andrews et al., suggest that a “risk-needs tool should emphasize criminogenic needs that have been shown to predict future re-offense.” This restriction to factors with demonstrated predictive ability, while generally sensible, is not without controversy. It may require modification when case planning purposes are being considered. For example, certain factors may not reach a globally predictive significance, but nevertheless may be critical in certain individual cases. More generally, the basic concept of criminogenic needs and what purpose they serve in risk and needs assessment and case

planning is not without controversy. Baird (2009) recently, in a broad critique of Andrews and Bonta’s approaches, argues that “the practice of labeling all needs as criminogenic appears to be a misguided effort to merge risk assessment – which uses group data to inform certain fundamental case decisions – with case planning, which must be based on the individual circumstances of each offender” (p. 9).

Our two major risk scales are actuarial tools used to sort individuals into groups of increasing probability of recidivism. These risk scores guide practice decisions such as supervision level. Our need scales generally follow the Andrews approach and their selection was guided by the current meta-analytic literature. Thus, evidence was available from this prior research – subsequently confirmed by our own research program - that these selected scales have an impact on key criminal justice outcomes. Most of our needs scales can be used to guide individualized decisions for case planning, as well as for identifying treatment targets and selecting interventions. Although we view risk scales separately from need scales in terms of function and purpose, both the need and risk scales were chosen because of their practical relevance in criminal justice decision-making. In other words, while we do not use all of the need scales to predict recidivism, we require the need scales to measure individual dynamic factors such as criminal thinking, education, employment, substance abuse, residential stability and other aspects of the person-in-environment that represent potential relevant targets for interventions.

Nevertheless, several of our psychometric studies demonstrate that many of these need scales measure factors that are highly relevant for individual case planning as well as having some predictive power. This is assessed by fitting univariable regression models where each specific need scale predicts future recidivism (again an independent criterion variable, to rule out criterion contamination). Table 2 shows the results of fitting a survival model to each need scale to predict a return to prison for a technical violation in the CDCR sample. In terms of the generalizability of these COMPAS findings, we note that several other studies across our client jurisdictions produce similar evidence to support the relevance of these factors.

The row for Vocational/Education shows the coefficient, hazard ratio, standard error, and t-value from a survival model in which Vocation/Education predicts return to prison for a technical violation. The hazard ratio indicates that for every one-unit increase in the Vocational/Educational decile score, the hazard for return to prison for a technical violation increases by 11%. The contents of the table are sorted by the magnitude of the hazard ratio. Scales with the largest hazard ratio are ranked higher. The top five scales on the basis of hazard ratio are Vocational Educational Problems, Criminal Personality, Social Adjustment, Residential Instability, and Criminal Thinking. If the t-value is greater than 1.96, the effect is significantly different than zero. Thus, all the estimates are significant, but fairly modest in size, although again similar to what other researchers find for similar samples. The significance level is a function of sample size and the sample consists of 6,485 soon-to-be-released inmates (first release to parole). These results demonstrate that the COMPAS need scales measure factors that are predictive of recidivism, and hence, may offer potential intervention targets.

**Table 2: Univariable Survival Model Results: Hazard for Return to Prison for a Technical Violation Regressed on Each Needs Scale (CDCR Sample).**

Scale Decile Score	Coeff	Hazard Ratio	SE (Coeff)	t-value
Vocational/Education	0.101	1.11	0.007	15.47
Criminal Personality	0.081	1.08	0.006	12.66
Social Adjustment	0.076	1.08	0.006	12.45
Residential Instability	0.074	1.08	0.006	12.58
Criminal Thinking	0.057	1.06	0.007	8.59
Leisure and Recreation	0.057	1.06	0.006	9.46
Substance Abuse	0.051	1.05	0.006	7.87
Financial Problems/Poverty	0.048	1.05	0.006	7.87
Social Environment	0.044	1.05	0.006	7.92
Family Criminality	0.043	1.04	0.006	7.12
Social Isolation	0.036	1.04	0.006	5.57

The information provided above is intended to clarify some of the questions raised by Skeem and Loudon in their review of COMPAS regarding predictive validity. We will continue to advance COMPAS as an evidence-based assessment technology to inform and improve decisions in the criminal justice system. We welcome further discussions about the design,

validation and testing of COMPAS and recognize the value of open dialogue in advancing knowledge and practice in the corrections field.

## **Internal Consistency and Inter-Rater Reliability of COMPAS Needs Scales**

On the issue of reliability, Skeem and Louden generally concur with our findings on the internal consistency of the COMPAS scales. These have largely met the standard requirements of alpha levels of 0.70 and above for most of our scales. Our item and factor analytic examinations also generally support the unidimensionality and expected factor structure of these scales. They also raise appropriate questions about inter-rater and test-retest reliability of COMPAS. We agree that this issue is important and have paid considerable attention to optimizing the reliability of our data collection processes through the design of standardized administrative and interviewing procedures, staff training and supervision and related organizational issues (see below). We also have recently initiated new studies of inter-rater reliability in conjunction with several of our client agencies.

While we have done consistent work on internal consistency of our scales we have completed less work on inter-rater reliability and test-retest reliability. Thus, to address this issue we have designed a new study of inter-rater reliability and this is currently under way. We also note that independent studies of test-retest and inter-rater reliability are being carried out by other university-based researchers.

To contextualize this discussion of reliability we first note that Baird (2009) has criticized all modern correctional risk and needs assessments for weaknesses and problems in regard to inter-rater reliability. While the basic question of whether two raters will reach the same score for a particular individual appears simple, the topic is quite complex with several different forms of reliability, as well as many confounding factors that can influence the consistency of raters. In terms of methods to assess consistency across “raters” Baird mentions the Kappa coefficient and its particular benefit in correcting chance agreement between raters. Another

key issue involves administrative procedures and staff skill requirements of many modern assessment methods. For example, the LSI-R utilizes semi-structured and motivational interviewing (MI) and these appear vulnerable to reliability problems by requiring multiple staff inferences, intensive training and high skills on the part of interviewers. In any criminal justice agency, if staff supervision, skills or training programs are deficient then unreliability can be a serious problem.

Additionally, it is important to understand the general context of inter-rater reliability across all the social sciences. Recent reviews suggest 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) in the Annual Review of Psychology 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 has been hailed as substantial and acceptable (see also Garb 1998). In general, highly structured and rule-based instruments tend to improve inter-rater reliability. Additionally, irrespective of the particular assessment tool, organizational factors can powerfully impact inter-rater reliability. In large correctional agencies the levels of staff training, competence, supervisory competence, work overload, workload stress and caseload sizes, all can profoundly impact inter-rater reliability. Even a highly reliable and structured assessment tool may be undermined and used inconsistently in an unfavorable organizational context. Thus, the level of reliability is determined only partly by the technical design of the risk and needs instrument, and also, profoundly, by organizational factors.

As noted above, inter-rater reliability, therefore, is a particular concern for assessment methods that require (or allow) multiple subjective decisions and clinical inference by staff in the assessment process. The design of COMPAS attempts to minimize such requirements. Baird (2009) argues that instruments such as the LSI and YASI that rely on semi-structured interviewing inevitably require many subjective judgments and clinical inferences by staff and may incur serious reliability problems. He cites a study by Austin, et al. (2003) that underscored

the inter-rater reliability problems of the LSI-R, noting “serious difficulties” in this aspect of reliability. In this regard the CMC component of the NCCD system also heavily relies on a semi-structured interviewing process (Harris, 1994; Hardyman, 2002) thus making this instrument similarly vulnerable to inter-rater reliability problems.

In attempting to minimize these problems the design of COMPAS uses several strategies: 1) We use multimodal data collection methods that minimize clinical inference and subjectivity by staff. This follows the findings of Wood et al (2002) and others, and a recommendation by Austin, et al. (2003) for simple standardized methods to minimize staff subjectivity and inference. 2) We utilize mathematical-statistical methods to replace or augment human judgment for classification decisions (where possible). These two approaches are now briefly discussed.

## **Automated Classifications and Reliability**

A separate but related issue regarding reliability is the use of numerical methods in processing the gathered data to reach a classification or predictive decision. This issue is quite separate from the abilities of assessment staff to obtain consistent data from respondents. It pertains specifically to the consistency and validity of such procedures to integrate the collected data into reliable decisions as compared to human or clinical judgment. An extensive body of research across half a century in psychological judgments and psychological diagnosis (Grove et al. 2000) has indicated that quantitative methods for diagnostic classification decisions are largely superior to clinical judgment. In fact, Quinsey et al 1996, in reviewing the 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. We view our automated and actuarial classification decisions as providing “decision support” to staff that can be overridden when staff can provide strong and reasonable justifications and has supervisory review.

Thus, in COMPAS, consistent with Grove et al 2000, we use quantitative pattern matching methods to automatically assign offenders to classification categories for both risk levels and for a separate need-based treatment typology, thus replacing human judgment for this task. 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 to Baird's CMC system. We use contemporary pattern recognition and quantitative methods in constructing and validating the typology, and for case assignment (Brennan, Dieterich and Breitenbach 2008). In reliability studies of classification consistency with this approach we use the kappa coefficient to measure of classification reliability in several split half studies (McIntyre-Blashfield 1980; Gordon 1999). These studies show that the automated pattern matching algorithms in classifying offenders into the typology 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 justify the integrity and viability of their discipline (Kirk and Kutchins 1986; Beutler and Malik 2002).

## **Data Collection Methods to Improve Reliability**

Returning to data collection strategy we attempt to minimize staff subjectivity and inferences by using a multimodal data collection design, as follows: 1) The first third of COMPAS questions are obtained from official criminal records – which minimizes staff subjectivity and allows supervisory verification. 2) Another third of the questions consist of a self-report checklist that does not require a staff rater. We note that Wood, et al. (2002) commented on the strength and viability of self-reports and their treatment utility. Since the assessment occurs in a correctional environment we embedded two automated data verification tests for “faking-good” and “coherency of responses” into COMPAS. We agree with Wood et al (2002) that such tests are particularly important in correctional settings. These verification tests trigger automated warnings to alert staff whenever such problems are detected. 3) Another third of the COMPAS instrument involves a standardized interview in which we use scripted

standardized question (with fixed response formats) that are read aloud sequentially with little or no comment by the interviewer, except to explain the meaning of a question (as needed). Such standardization is widely used in social sciences to minimize rater inference, biases and to obviate training and skill differences among staff in order to achieve higher reliability. However, we realize that in certain situations there are advantages to semi-structured interviewing and related methods and thus, we have also developed a semi-structured interview approach for this section.

To conclude, we agree with Skeem and Loudon on the importance of inter-rater and internal consistency and other forms of reliability and that this is a constant challenge in large busy criminal justice organizations. It is clear that no administrative or interview process cannot totally avoid this issue. We have designed our current administrative and analytical strategies to optimize ease of use, efficiency as well as reliability and validity within the relatively high stress environments of large-scale correctional agencies. The pervasive challenge of limited correctional and staffing resources is one of the more serious, consistent and limiting factors in achieving high quality data. Organizational issues inevitably have a supportive or deleterious impact on staff skills, training and supervision, work overload and time constraints for assessment. Thus, such organizational factors must also enter into the design of workable and efficient assessment techniques.

In closing, this document lays out some agreements and disagreements with Skeem and Loudon and offers updates and new studies that address many of their issues. Their review identifies many measurement issues that are perennial challenges, not just to COMPAS, but to all applied risk and needs instruments used in criminal justice. We suggest, however, that a more complete review that has access to the full scope of our continuing research program would be a fairer statement on the current validation evidence for COMPAS. Thus, we have described additional findings and design procedures from our on-going research program that address most of the key issues in the Skeem and Loudon report. Optimizing reliability and

demonstrating validity of our methods will continue as a priority in the evolution of the COMPAS platform.

Please visit our website to view copies of the reports mentioned in this document at [www.northpointeinc.com](http://www.northpointeinc.com) If you have questions, please feel free to contact us at 303-216-9455 or by email at [info@npipm.com](mailto:info@npipm.com)

## References

- Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime and Delinquency.
- Barnowski, R., & Drake, E. K. (2007). *Washington's Offender Accountability Act: Department of Corrections' Static Risk Assessment*. Olympia, WA: Washington State Institute for Public Policy.
- Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21 – 40.
- Breitenbach M, Dieterich,W., Brennan, T. and Fan, A. (2009 – In press). Creating Risk-Scores in Very Imbalanced Datasets – Predicting Extremely Violent Crime. Ch.15 Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection. Ed. Yun Sing Koh; Publisher: IGI Global
- Farrington, D. P., Jolliffe, D., Loeber, R., Stouthamer-Loeber, M., & Kalb, L. M. (2001). The concentration of offenders in families, and family criminality in the prediction of boys' delinquency. *Journal of Adolescence*, 24, 579-596.
- 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.
- Gendreau, P., Goggin, C. E., & Law, M. A. (1997). Predicting prison misconducts. *Criminal Justice and Behavior*, 24 (4), 414 - 431.
- Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency*, 52(1), 178 – 200.
- Herrenkohl, T. I., Maguin, E., Hill, K. G., Hawkins, J. D., Abbott, R. D., & Catalano, R. F. (2000). Developmental risk factors for youth violence. *Journal of Adolescent Health*, 26, 176-186.
- Moffit, T. E. (2003). Life-course-persistent and adolescence-limited antisocial behavior: A 10-year research review and a research agenda. In B. B. Lahey, T. E. Moffitt, & A. Caspi (Eds.), *Causes of conduct disorder and juvenile delinquency* (pp. 49 – 75). New York: The Guilford Press.
- Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 100 (4), 674-701.
- Nelson, M., Deess, P., & Allen, C. (1999). *The first month out: Post- incarceration experiences in New York City* [Unpublished monograph]. New York.

- Horney, Julie, D. Wayne Osgood, and Ineke Haen Marshall. 1995. "Criminal Careers in the Short-Term: Intra-Individual Variability in Crime and Its Relation to Local Life Circumstances." *American Sociological Review* 60:655-73.
- Stouthamer-Loeber, M., Loeber, R., Wei, E., Farrington, D. P., & Wikstrom, P. H. (2002). Risk and promotive effects in the explanation of persistent serious delinquency in boys. *Journal of Consulting and Clinical Psychology, 70* (1), 111-123.
- Tolan, P. H., & Gorman-Smith, D. (1998). Development of serious and violent offending careers. In R. Loeber & D. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 68-85). Thousand Oaks, CA: Sage.