

A Comparison of Risk Assessment Instruments in Juvenile Justice

August 2013

Chris Baird
Theresa Healy
Kristen Johnson, PhD
Andrea Bogie
Erin Wicke Dankert
Chris Scharenbroch

National Council on Crime and Delinquency

This study was funded by *Grant 2010-JR-FX-0021 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, US Department of Justice*. Points of view in this report are those of the authors and do not necessarily reflect the official position or policies of the US Department of Justice.

Acknowledgments

A study of this magnitude requires a tremendous amount of commitment from participating agencies. The National Council on Crime and Delinquency would like to thank the following individuals for their remarkable efforts in making this study possible. From the Arkansas Division of Youth Services, thank you to Ron Angel, Elbert Grimes, Lisa Hutchinson, and Mickey Yeager; from the Arizona Administrative Office of the Courts, Chad Campbell, Amy Stuart, and David Redpath; from the Arizona Department of Juvenile Corrections, John Vivian, PhD, and Tasha Fox; from the Nebraska Office of Probation Administration, Corey Steele and Amy Latshaw; Peg Barner with the Nebraska Office of Juvenile Services; from the Virginia Department of Juvenile Justice, Beth Stinnett; from the Florida Department of Juvenile Justice, Mark Greenwald and Julie Lauder; Dee Bell and Josh Cargile from the Georgia Department of Juvenile Justice; Lisa Wamble and Earl Montilla with Solano County Probation Department; Cherie Lingelbach from Oregon's Youth Authority; and Torri Lynn, who represented the county directors from across the state of Oregon. We also extend our gratitude to the members of the project's advisory board for their extensive voluntary efforts. Thank you all for your dedication and hard work on this study; it could not have been done without you. Your contributions were essential to this effort in moving the field of juvenile justice risk assessment forward.

This report does not necessarily reflect the views or positions of any of the agencies that participated in the study. Dissenting opinions from the advisory board are included at the end of the report in the discussion section.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	i
I. INTRODUCTION	1
A. The Historical Background of Risk Assessment in Juvenile Justice	2
II. RESEARCH METHODOLOGY	5
A. Goals	5
B. Research Questions	6
C. Risk Assessment Instruments Evaluated	6
D. Participants	7
E. Advisory Board	10
F. Measures	11
1. Reliability	12
a. Calculating Reliability	12
b. Methods Used to Study Reliability	14
2. Validity	19
a. Construction and Validation of Revised Risk Assessments	23
3. Equity	24
4. Cost	25
III. FINDINGS	25
A. Findings by Risk Assessment Instrument	26
1. The Georgia CRN	26
a. Summary of Findings	34
2. Solano County JSC and Girls Link Risk Assessments	34
a. Summary of Findings	37
3. Florida PACT	38
a. Summary of Findings	45
4. Virginia YASI	46
a. Summary of Findings	51
5. Nebraska and Arkansas YLS/CMI	51
a. Summary of Findings	55
6. Arizona AOC Risk Assessment Instrument	55
a. Summary of Findings	60
7. Arizona DJC DRI	61
a. Summary of Findings	65
8. Oregon JCP Assessment	66
a. Summary of Findings	68
B. Comparison of Results Across Jurisdictions and Assessments	68
1. Reliability	68
2. Validity	72
3. Equity	76
4. Revised Risk Assessment Instruments Constructed in the Study	78
5. Efficiency and Cost	82

TABLE OF CONTENTS (continued)

IV.	DISCUSSION	86
A.	Instruments Developed for General Use	87
1.	Overall Results	87
2.	Developmental Methods.....	88
3.	Other Design Issues	91
4.	Would Simpler Models Transfer Better Among Agencies?	94
5.	Are Complex Scoring Algorithms or Classification Methods Needed or Beneficial?	98
B.	Risk Instruments Developed for a Specific Agency	100
C.	Comments From Advisory Board Members and Authors' Responses	103
1.	Best-Practice Implications of the Study Findings: Comments on the Validity, Reliability, and Equity of Commonly Used Juvenile Risk Instruments, by James Howell, PhD, and Aron Shlonsky, PhD.....	103
2.	Youth Risk Assessment Approaches: Lessons Learned and Questions Raised by Baird et al.'s Study (2013), by Jennifer Skeem, PhD, and (in alphabetical order) Robert Barnoski, PhD, Edward Latessa, PhD, David Robinson, PhD, and Claus Tjaden, PhD.....	108
a.	Overview.....	108
i.	Context and Purpose.....	108
ii.	Summary of Key Points.....	109
b.	Conclusions Supported by Data	110
c.	Open Question: Does Reduction-Oriented Risk Assessment Add Value?.....	119
3.	Authors' Responses to Comments.....	121
V.	LIMITATIONS.....	132
VI.	CONCLUSION.....	134
	REFERENCES.....	135

APPENDICES

Appendix A:	Risk/Needs Assessment Systems
Appendix B:	Validation Results by Site
Appendix C:	Reliability Results by Site
Appendix D:	Expert Scorer Qualifications
Appendix E:	Staff Perceptions
Appendix F:	Administrator Advice

EXECUTIVE SUMMARY

Juvenile justice service staff began exploring the use of actuarial risk assessments that classify offenders by the likelihood of future delinquency with earnest in the 1970s, but actuarial risk assessments have been used by public social service agencies in the United States since 1928. The value and utility of a valid, reliable, and equitable risk assessment within a broader practice reform effort was made clear to justice agencies in 1998 when the Office of Juvenile Justice and Delinquency Prevention (OJJDP) published the Comprehensive Strategy for Serious, Violent, and Chronic Juvenile Offenders. OJJDP's reform effort illustrated how juvenile justice agencies could better ensure the effectiveness and appropriate targeting of services by implementing both an actuarial risk assessment to accurately, reliably, and equitably classify youth by the likelihood of future delinquency and an equally effective needs assessment to identify an intervention and treatment plan tailored to an individual's needs. This approach built upon the efforts of the National Institute of Corrections' Model Probation/Parole Management Project that combined actuarial risk assessment, individual needs assessment for effective treatment planning, regular reassessments of risk and needs and risk-based supervision standards, and workload-based budgeting.

Other models of risk assessment were introduced over subsequent decades, and researchers began categorizing and comparing them as generations of risk assessments. The first generation of risk assessments were not actuarial—individual workers assigned risk levels without the aid of actuarial instruments. Generation 2 instruments were statistically derived, but relied heavily on static criminal history factors to assess risk. They tended to be developed using local data for specific jurisdictions, typically consisted of fewer than a dozen factors (e.g., the California Base Expectancy Tables developed in the 1960s), and focused on identifying groups of offenders with distinctly different risks of future offending. Many of today's instruments, often referred to as generation 3 or generation 4, have expanded beyond the singular objective of risk assessment to classify individuals by risk of delinquency. These instruments often contain dozens of factors (for example, the Correctional Offender Management Profiling and Alternative Sanctions [COMPAS] Youth risk assessment instrument). They frequently divide risk factors into two groups: "static" and "dynamic" (see, for example, Schwalbe, 2008; Hoge, 2002). Static factors are generally measures of prior delinquency. Dynamic factors are commonly referred to as "criminogenic needs" and represent conditions or circumstances that can improve over time (Andrews, Bonta, & Wormith, 2006). In addition, protective factors and references to "responsivity" have been added to generation 4 instruments. Responsivity is intended to reflect an individual's readiness for change and gauge a youth's ability to respond to particular treatment methods and programs (Andrews, 1990). Generation 4 instruments contain anywhere from 42 to approximately 150 factors.

These variations in methodology and philosophy raised questions about which types of instruments most accurately and effectively help jurisdictions differentiate between low-, moderate-, and high-risk youth. Many evaluations of risk assessments based validity on correlation coefficients or other measures of association. Those that examined the degree of difference in recidivism rates observed for youth identified as low, moderate, or high risk often found little differentiation; results could vary

substantially by race, ethnicity, and gender. Few jurisdictions conducted local validation studies to ensure a risk assessment's validity and reliability, and now one foundation-funded reform effort is telling agencies that local validation is not required if an instrument has been validated in three agencies or for similar populations.

Perhaps the most significant change in the last few decades has been the emergence of commercially available risk assessment systems. Prior to this development, risk assessment studies were generally conducted by universities, nonprofit research organizations, or research units within government agencies. Claims made about the validity and reliability of some of these tools have been challenged by other researchers (Skeem & Eno Louden, 2007; Baird, 2009). In response to concerns about the classification and predictive validity of several risk assessments voiced by juvenile justice practitioners and researchers, OJJDP funded a proposal submitted by the National Council on Crime and Delinquency (NCCD) to evaluate commonly used risk assessments by comparing their predictive validity, reliability, equity, and cost. NCCD is a nonprofit social research organization, and its researchers conducted the study of eight risk assessments in 10 jurisdictions in consultation with an advisory board of juvenile justice researchers and developers of commercial juvenile justice risk assessment systems included in the study.

The 10 jurisdictions use a variety of risk assessment instruments, ranging from commercially available systems to models developed for use by a specific agency. The seven agencies that use risk assessment models created for general use include the Arkansas Department of Human Services, Division of Youth Services; Florida Department of Juvenile Justice; Georgia Department of Juvenile Justice; Nebraska Department of Health and Human Services, Office of Juvenile Services; Nebraska Office of Probation Administration; Solano County, California; and the Virginia Department of Juvenile Justice. The three that were validated on and for local populations are Arizona and Oregon tools (Table E1).

Table E1			
Sites and Risk Assessments Evaluated for Inter-Rater Reliability and Validity			
Site Agency	Risk Assessment Instrument	Who completes risk assessment protocol?	What decisions does it inform?
Arizona Administrative Office of the Courts (AOC)	Risk/needs system validated for Arizona youth placed/referred to juvenile court	Probation officers	Supervision type and level, services
Arizona Department of Juvenile Corrections (DJC)	Dynamic Risk Instrument (DRI), validated for secure care/committed population	Secure commitment facility staff	Placement decisions, treatment planning, case planning, release decisions
Arkansas Department of Human Services, Division of Youth Services (DYS)	Youth Level of Service/Case Management Inventory (YLS/CMI) for youth in secure commitment	Case coordinators and service managers	Establishment of treatment goals, program placement

Table E1			
Sites and Risk Assessments Evaluated for Inter-Rater Reliability and Validity			
Site Agency	Risk Assessment Instrument	Who completes risk assessment protocol?	What decisions does it inform?
Florida Department of Juvenile Justice (DJJ)	Positive Achievement Change Tool (PACT)	Probation officers	Supervision levels, services
Georgia Department of Juvenile Justice (DJJ)	Comprehensive Risk/Needs (CRN) assessment, an early derivative of COMPAS Youth	Probation/commitment assessment specialists	Supervision levels, commitment and placement decisions
Nebraska Department of Health and Human Services, Office of Juvenile Services (OJS)	YLS/CMI for youth in secure commitment	OJS evaluation coordinators	Supervision levels, commitment decisions
Nebraska Office of Probation Administration	YLS/CMI	Probation officers	Supervision levels, case planning
Oregon Juvenile Justice	Juvenile Crime Prevention (JCP) assessment developed for youth referred to juvenile justice system	Probation officers, detention workers, and prevention workers	Direct service supervision, case plan
Solano County, California	Gender-specific risk assessments in JAIS for youth referred to probation	Probation officers	Risk informs supervision and service intensity, needs assessment case plan
Virginia Department of Juvenile Justice (DJJ)	Youth Assessment and Screening Instrument (YASI) for youth on probation, facility or parole	Probation officers and facility staff	Supervision levels, number of probation contacts, case plan

Inter-Rater Reliability Testing

Inter-rater reliability is a necessary quality in an assessment because it helps ensure that different caseworkers, faced with the same case information, will reach the same scoring and recommendations for key decision thresholds such as risk of future delinquency. If assessment items are not reliable, it is unlikely that they will be predictive.

We measured the inter-rater reliability of risk assessment items by asking a sample of officers/caseworkers to review case files for 10 youth, observe a videotaped interview of each youth, and score a risk assessment (or risk/needs assessment) for each youth. The number of raters varied by site between five and 69, with most sites engaging 25 or more workers in testing (selection was random in some sites but voluntary in others). We used multiple measures to assess inter-rater reliability, as each has limitations that are important to understand. Percent agreement is and has been our primary measure for comparison across items and assessments because it is easy to understand and transparent; the limitation is that it does not control for the likelihood that caseworkers would randomly reach the same response by chance.

In a comparison of assigned risk level by each assessment for 10 test cases, most tools achieved high percent agreement between workers. Fewer instruments achieved high levels of agreement with an expert score (five of the 10), intra-class correlation coefficient with risk score at or above .80 (five), and kappa above .6 (three). Of most interest is that only three of the risk assessments had positive indications of inter-rater reliability across every measure: Arizona's homegrown AOC assessment, Solano County's gender-specific assessments, and Virginia's YASI. Overall, prior delinquency history and other similar static risk factors demonstrated higher levels of inter-rater agreement than dynamic factors; this was especially true for more subjective measures such as youth attitudes.

Validity and Equity Testing

In order to effectively target limited resources, a risk assessment needs to result in valid and equitable classifications. Testing the predictive validity and equity of the risk assessments involved sampling a cohort of youth on probation or released from a facility (i.e., post-commitment). Recidivism was tracked over a 12-month follow-up period for all sites but one (where only nine months of outcomes were available). Outcome measures were obtained from agency databases and include subsequent arrests, subsequent adjudications, and subsequent placement in a juvenile facility. Exceptions were two sites for which recidivism was limited to return to a correctional facility for youth released from facilities. Findings showed that several of the evaluated risk assessment systems failed to provide the level of discrimination needed by probation and correctional service staff if they are to optimize decisions regarding supervision requirements.

Three systems, the Oregon JCP, Solano County's Juvenile Sanction Center risk assessment for boys, and the YASI model used in Virginia, demonstrated considerable capacity to accurately separate cases into low, moderate, and high risk levels with progressively higher recidivism with each risk level increase. The area under the curve (AUC) and Dispersion Index for Risk (DIFR) scores for these risk assessments were also acceptable. Other instruments evaluated suffered from a lack of distinction between risk categories by outcomes examined. The AUC and DIFR were also insufficient for several risk models.

In all jurisdictions where sample size allowed, NCCD conducted additional analyses to determine if a simple actuarial risk instrument would provide better classification results. This effort was restricted by available data, but better results were obtained in most instances using simple construction scale methods such as analyses of correlations and regression models. In two agencies with large study cohorts available, cases were divided into construction and validation samples and results from the validation samples presented. This step is recommended because results from a construction are generally the best that will be attained. When tested on an independent sample, the level of discrimination attained tends to decline. In this exercise, we found minimal "shrinkage." The combined results of all analyses conducted suggest that limiting factors on a risk assessment to those with a strong, significant relationship to outcomes will result in a more accurate risk classification.

Some members of the advisory board claim that little difference was shown in predictive efficacy of many of the instruments tested in this study. They base these conclusions primarily on a comparison of AUC values. Their viewpoint, comments from other advisory board members, and our responses appear in the “Discussion” section of the report. In short, risk assessments should be evaluated based on how the information informs practice; thus, we assessed predictive validity using multiple measures, with recurrence of delinquency by risk classification level as our primary measure. The reasoning for this approach is further described in the body of the report.

Implications for Practice

The proper use of valid, reliable risk assessments can clearly improve decision making. Results of this study show, however, that the power of some risk assessment instruments to accurately classify offenders by risk level may have been overestimated. The first step in remedying this situation is to ensure that everyone working in the field of juvenile justice understands the importance of valid, reliable, and equitable risk and needs information. Although the study provided fodder for many areas of policy and practice, as well as future research and development, researchers, practitioners, and advocates should focus attention on the following points.

- A. Jurisdictions must be able to ensure that the risk assessment completed by field staff to inform case decision making is reliable, valid, and equitable. Decisions about youth are based on the level of risk assigned. Thus, the primary measure of validity must be the level of discrimination produced. This study clearly demonstrates that similar AUCs do not translate into similar classification capability. Jurisdictions should expect reliability testing and validation studies when assessment models are transferred to other jurisdictions and would benefit from making evaluation of assessments part of a more comprehensive approach to evidence-based practice.
- B. National standards could provide juvenile justice administrators with clear guidelines for assessing the reliability, validity, and equity of existing models. Such standards could also help agencies develop the capacity to construct instruments for their populations and understand how valid risk and needs information can help them monitor and improve practice. National standards could be established to help ensure due diligence, such as ensuring reliability testing and validation studies before and after risk assessment instruments are transferred to other jurisdictions and emphasizing measures that are most applicable for practice conditions and easier for administrators to understand. Measures emphasized over the last decade have significant shortcomings and fail to convey that which is most important to correctional administrators: the level of discrimination in outcomes attained between risk levels and the proportion of cases assigned to each risk level. The purpose of risk assessment is to classify offenders into groups with substantially different probabilities of future offending; measures such as correlations (frequently depicted as effect size) and AUC, while useful, are not by themselves adequate measures of validity. Likewise,

while correlations are not adequate measures of reliability, they sometimes are the only measure reported.

- C. Risk assessment should focus solely on identifying cases most and least likely to be involved in future offending, e.g., limiting the list of contributing factors to items significantly related to delinquency in the expected direction. Simple, straightforward, actuarial approaches to risk assessment generally outperform more complicated approaches.

Risk assessment should be a simple process that can be easily understood and articulated. This study's findings show that simple, actuarial approaches to risk assessment can produce the strongest results. Adding factors with relatively weak statistical relationships to recidivism—including dynamic factors and criminogenic needs—can result in reduced capacity to accurately identify high-, moderate-, and low-risk offenders.

INTRODUCTION

This study examined the validity, reliability, equity, and cost of nine juvenile justice risk assessment instruments. Though many researchers and practitioners believe that risk assessment is critical to improving decision making in the juvenile justice system, the range of options currently available makes the selection of the most appropriate instrument for each jurisdiction a difficult choice. This study was designed to provide a comprehensive examination of how several risk assessments perform in practice.

Further, the study helps establish an agenda for both researchers and practitioners to explore questions relating to risk assessment construction, evaluation, and use in practice and offers possible solutions to issues identified in the study. Using data available in each agency's data extract, additional analyses were undertaken to determine if validity and equity could be improved using actuarial scale construction methodology. These analyses are presented and discussed in the "Discussion" section of the report.

The study is premised upon the need for valid, reliable, and equitable risk assessment instruments in juvenile justice. Broadly defined, risk assessment refers to the process of estimating an individual's likelihood of continued involvement in delinquent behavior. A risk instrument can inform crucial decisions, including whether and where youth will be incarcerated, how they will be supervised in the community, and in which programs they will participate. A valid, reliable, and equitable assessment of risk, when used in concert with sound clinical judgment and effective delivery of appropriate services, can be essential to treatment, reentry, and rehabilitation. Accurate assessment can also help juvenile justice agencies allocate resources to youth who need them most, which can then impact the safety and well-being of communities.

The study examined the Positive Achievement Change Tool (PACT), the Youth Assessment and Screening Instrument (YASI), the Youth Level of Service/Case Management Inventory (YLS/CMI), the

Comprehensive Risk and Needs Assessment (CRN, a derivative of Correctional Offender Management Profiling and Alternative Sanctions [COMPAS] Youth), the Juvenile Sanctions Center (JSC) risk assessment instrument, the Girls Link risk assessment instrument, the Arizona Administrative Office of the Courts risk assessment instrument, the Arizona Department of Juvenile Correction Dynamic Risk Instrument (DRI), and the Oregon Juvenile Crime Prevention (JCP) assessment.

The National Council on Crime and Delinquency (NCCD) conducted the study between 2011 and 2013. The study was supported by the Office of Juvenile Justice and Delinquency Prevention (OJJDP) and overseen by an advisory board that approved study design and implementation.

A. The Historical Background of Risk Assessments in Juvenile Justice

Recent literature on risk assessment describes four different generations of risk assessments. Variations in methodology and philosophy have developed over time, and the objectives of risk assessment have expanded beyond classification. This expansion has raised questions about which types of instruments most accurately and effectively help jurisdictions differentiate between low-, moderate-, and high-risk youth; whether the instruments are consistently completed by line staff; and whether the instruments equitably assign youth to risk levels by race and gender.

Early approaches to risk assessment are generally known as “generation 1” and “generation 2.” In generation 1, risk levels were assigned by individual workers without the aid of actuarial instruments. Generation 2 instruments were statistically derived, but relied heavily on static criminal history factors to assess risk. They tended to be developed using local data for specific jurisdictions, typically consisted of fewer than a dozen factors (e.g., the California Base Expectancy Tables developed in the 1960s), and focused on identifying groups of offenders with distinctly different risks of future offending. Juvenile risk instruments were first developed in the 1970s.

Many of today's instruments, often referred to as generation 3 or generation 4, have expanded beyond the singular objective of risk assessment. These instruments are meant for general use, as they most often have not been constructed for a particular jurisdiction's population. They often contain dozens of factors (e.g., the COMPAS Youth risk assessment instrument). They frequently divide risk factors into two groups: "static" and "dynamic" (see, for example, Schwalbe, 2008; Hoge, 2002). Static factors are generally measures of prior delinquency. Dynamic factors are commonly referred to as "criminogenic needs" and represent conditions or circumstances that can change over time (Andrews, Bonta, & Wormith, 2006).

In addition, protective factors and references to "responsivity" have been added to generation 4 instruments. Responsivity is intended to reflect an individual's readiness for change and to gauge a youth's ability to respond to particular treatment methods and programs (Andrews, 1990). This change has sometimes resulted in the addition to these instruments of "protective factors"—conditions, circumstances, or strengths that may help a youth overcome obstacles to success in the community. Generation 4 instruments now contain anywhere from 42 to approximately 150 factors.

Some generation 3 and 4 instruments incorporate risk factors identified in prior research studies and in one or more theories of criminal or deviant behavior. The YLS/CMI, for example, includes "those items that previous research had indicated were most strongly associated with youthful criminal behavior" and were also based on the "General Personality and Social Psychological Model of Criminal Conduct" (Andrews & Bonta, 2003). Similarly, the COMPAS Youth risk assessment instrument is based on theories of criminal/deviant behavior (Brennan, Dieterich, & Ehret, 2009).

A principle of risk assessment is that services should be targeted to the highest-risk cases (Andrews & Bonta, 1995). Therefore, if a risk assessment instrument does not accurately identify high-risk cases, the instrument does not achieve its primary purpose. It does little good to identify service needs unless services are directed toward youths who are truly "high risk." If changes to risk assessment instruments have resulted in diminished capacity to accurately discriminate among high-

moderate-, and low-risk youth, then decision making in juvenile justice has been adversely affected, regardless of other features added to the instruments.

Evaluations of individual instruments have been conducted in recent years. In some instances, researchers have found little differentiation between low, moderate, and high risk levels. For example, a 2007 study of YASI in the state of New York found only a 3.8% difference in outcomes between moderate- and high-risk cases (Orbis Partners, 2007).¹ A study of the PACT risk assessment instrument in Florida found little difference in recidivism rates among moderate-risk, moderately high-risk, and high-risk probationers (Baglivio & Jackowski, 2012). In another study, eight particular factors from the Level of Service Inventory-Revised (LSI-R) produced far better discrimination than the entire 54-item scale for offenders in Pennsylvania (Austin, Coleman, Peyton, & Johnson, 2003). A recent validation of the adult COMPAS model in California found that better discrimination could be attained using four simple risk factors (Farabee, Zhang, Roberts, & Yang, 2010).

In addition, relatively few published studies of these risk assessment instruments have included an item analysis—that is, an analysis of how well each of the factors individually corresponds to the risk of recidivism. Of those that did include item analysis, some found only modest and often insignificant relationships between risk factors and outcomes. For example, a study of the LSI determined that a substantial number of factors in the instrument demonstrated little or no relationship to recidivism (Flores, Travis, & Latessa, 2003).² A review of COMPAS (Skeem & Eno Loudon, 2007; p.17, 19) found “no evidence that the original COMPAS basic scales, higher-order scales, or risk scales predict recidivism ... available data provide no evidence that the original COMPAS risk scales

¹ The 3.8% difference cited above reflects the difference in arrest rates 12 months from the date of the assessment. The difference in adjudication rates was 3.7%. Twelve-month rates were reported for arrests and adjudications because those were the recidivism measures available in the jurisdictions that participated in this study. While overall base rates for these two outcomes increased substantially at 24 months in New York, differences observed between cases rated moderate and high risk increased only slightly, from 3.7% to 5.9% for adjudications and from 3.8% to 6.6% for arrests. Even at 24 months, the differences remain well below those found in many effective classification systems.

² The Youth Level of Service Inventory (LSI) is an adaptation of the YLS used in adult corrections. The YLS/CMI also includes a case-planning component.

predict reoffending of any sort.” The original COMPAS was validated by estimating the relationships between COMPAS scores and *prior* criminal activity.

Some studies of generation 3 and 4 instruments used small samples and overstated their conclusions. For example, in one recent publication reporting on results of 47 studies of LSI validity, only correlations were reported (Vose, Cullen, & Smith, 2008). The correlation coefficients obtained varied substantially, and coefficients as low as 0.137 were cited as evidence of validity. (The threshold for validity was a statistically significant relationship between the total LSI-R score and some measure of recidivism.) Of these studies, 22 used samples of less than 200 cases. In another meta-analysis of 22 LSI validation studies, eight studies used samples of 100 or fewer offenders, and only two examined samples of 300 or more cases. Only correlations were reported, and the average correlation between LSI scores and recidivism was 0.24 (Campbell, French, & Gendreau, 2007).

Few of the existing evaluations include studies of reliability and equity. The existing studies tend to focus on a single risk instrument at a time and use various methods to examine validity, reliability, and equity, making comparisons across instruments as well as generalizations about the field difficult. This study examines the validity, reliability, and equity of nine instruments using the same methodology to review validity, reliability, and equity.

II. Research Methodology

A. Goals

This project was designed to provide the field with an objective study of the validity, reliability, and equity of different approaches to risk assessment. A second goal was to review methods currently used to evaluate validity and reliability and to discuss the strengths and weaknesses of each. A third goal of the research was to provide the field with clear and relevant information on each instrument’s capacity to estimate risk across all major race/ethnicity groups. Ensuring that risk assessment

instruments equitably classify all youth could help reduce the incidence of minority overrepresentation in the juvenile justice system. A fourth goal of the study was to report basic cost parameters about each of the risk assessment instruments reviewed.

B. Research Questions

This study posed the following questions.

1. Is each risk assessment instrument sufficiently reliable (i.e., inter-rater reliability) to ensure that decisions regarding level of risk and identified service needs are consistent across the organization?
2. What specific risk assessment items are associated with less reliability? What items are rated reliably by staff?
3. Is each risk assessment instrument valid? Specifically, what degree of discrimination is attained between assigned risk levels? Could the instrument be improved by adding or deleting specific factors and/or altering cut-off scores?
4. Is each risk assessment instrument valid for population subgroups: White/Caucasian, Black/African American, Hispanic/Latino, females, probationers, and youth in aftercare status? Could equity be improved by adding or deleting specific factors or altering cut-off scores?
5. What costs are associated with each risk assessment instrument?

C. Risk Assessment Instruments Evaluated

The following risk assessment instruments were reviewed for this study:

- Youth Level of Service/Case Management Inventory (YLS/CMI), Multi-Health Systems;
- Positive Achievement Change Tool (PACT), Assessments.com;
- Comprehensive Risk/Needs Assessment (CRN), a derivative of COMPAS Youth, Northpointe, Inc.;
- Youth Assessment and Screening Instrument (YASI), Orbis Partners, Inc.;

- The Juvenile Sanctions Center (JSC) risk assessment, available in the public domain;
- A risk assessment developed for the Girls Link program in Cook County, Illinois, available in the public domain;
- The Oregon Juvenile Crime Prevention (JCP) Assessment, Oregon;
- The Arizona Department of Juvenile Corrections Dynamic Risk Instrument (DRI), Arizona DJC; and
- The Arizona Juvenile Risk and Needs Assessment, Arizona Administrative Office of the Courts.

D. Participants

Participants included juvenile justice agencies that had implemented risk/needs assessments during the past 10 to 12 years. They represent the range of agencies that use risk/needs assessments: county probation, state probation, and state juvenile justice systems responsible for incarcerated youth and those in aftercare. Brief profiles of each agency are provided below.

- Arkansas Department of Human Services, Division of Youth Services (DYS), is a statewide agency responsible for youth in secure commitment. DYS uses the YLS/CMI.
- Florida Department of Juvenile Justice (DJJ) uses the PACT instrument. DJJ is a statewide system that works with juveniles on probation, in secure care, and in aftercare.
- Georgia Department of Juvenile Justice (DJJ) uses the CRN instrument, an early derivative of the COMPAS Youth. DJJ works with youth on probation, in secure commitment, and in aftercare in 142 of 159 “dependent” counties (17 other counties have their own “independent” court services).
- Virginia Department of Juvenile Justice (DJJ) is a statewide agency that works with juveniles on probation, in secure care, and in parole. DJJ uses the YASI.
- Nebraska Office of Probation Administration is a statewide agency that works with youth on probation and uses the YLS/CMI.
- Nebraska Department of Health and Human Services, Office of Juvenile Services (OJS) is a statewide agency that works with youth in secure commitment. OJS uses the YLS/CMI.

- Solano County, California, Probation Department, Youth Division, uses the JSC risk assessment for boys and the Girls Link instrument for girls.
- Arizona Department of Juvenile Corrections (DJC) is a statewide agency that at the time of the study used the DRI, a risk assessment instrument developed specifically for the secure care commitment population in the state.
- Arizona Administrative Office of the Courts (AOC) is responsible for setting policies related to youth referred to juvenile court, including youth placed on probation. AOC uses a risk assessment developed and validated specifically for cases referred to juvenile court in Arizona.
- All county-based juvenile justice departments in the state of Oregon. In addition, the state oversight body, the Oregon Youth Authority, which is responsible for youth offenders and other functions related to state programs for youth corrections, was involved. The JCP risk assessment instrument was developed specifically for Oregon.

A comparison of risk assessment instrument use across sites is presented below, including who completes each instrument, when the instrument is completed, and what decisions are informed by the results.³ For copies of each instrument, see Appendix A.

³ Based on interviews with site administrators in September 2012.

<p>Table 1</p> <p>Use of the Risk Assessment at Each Site</p>					
Site	Risk Assessment Instrument	Who completes instrument?	When?	What decisions does it inform?	Shared with courts?
Arizona AOC	Risk and Needs Assessment	Probation officers	At referral and when probation is ordered	Supervision type and level, services	Varies
Arizona DJC	DRI	Secure commitment facility staff	Intake	Placement decisions, treatment planning, case planning, release decisions	Yes, courts can view results on website
Arkansas	YLS/CMI	Case coordinators and service managers	Within 1–3 weeks of commitment to DYS custody	Establishment of treatment goals, program placement	Yes
Florida	PACT	Probation officers and contracted staff (in some instances)	Intake, new violations, and re-assessments every 90 days	Supervision levels, services, risks, needs	Yes
Georgia	CRN	Juvenile probation parole specialists (probation) and assessment and classification specialists (commitment)	Within 30 days of disposition (probation) and prior to the 10th business day after disposition (commitment)	Supervision levels, commitment decisions and placement, custody decisions	Yes, for committed youth
Nebraska OJS	YLS/CMI	OJS evaluation coordinators	After adjudication or if commitment is anticipated	Supervision levels, commitment decisions	Yes
Nebraska Probation	YLS/CMI	Probation officers	Pre-disposition investigation, placed on probation, or new juvenile (if not done previously)	Supervision levels, case planning	Yes
Oregon	JCP	Probation officers, county detention workers, and juvenile crime prevention community agencies	Intake, program referral, or after adjudication (in small number of counties)	Direct service supervision, case planning	Varies by jurisdiction
Solano County, California	JSC; Girls Link	Probation officers	Every six months after the initial assessment	Risk assessment informs supervision levels; risk and needs assessments inform services and case planning	Yes
Virginia	YASI	Probation officers and secure commitment facility staff (as of 7/1/12)	Predisposition reports, when probation is ordered, at time of commitment, and six-month reassessment	Supervision levels, number of probation contacts, commitment case planning	Yes

E. Advisory Board

An advisory board consisting of researchers, a former head of a juvenile corrections department, and purveyors of the various risk assessment instruments examined in the study helped oversee study design and completion. The advisory board met seven times during the course of the project (once in Phoenix, Arizona; once in Baltimore, Maryland; and five times via web-based conference) to review and approve all proposed methods of data collection and analysis, all materials used to conduct reliability testing, and all findings and results. Advisory board members reviewed a draft report and were given an opportunity to include a dissenting opinion on any aspect of the analysis and final report. The advisory board consisted of the following individuals:

- David Gaspar, Senior Program Manager, NCCD; former director, Arizona DJC; former president of Council of Juvenile Correctional Administrators; former member of the Board of Governors of the American Correctional Association;
- Sean Hosman, JD, CEO, Assessments.com;
- James Howell, PhD, Managing Partner, Comprehensive Strategy Group;
- Edward Latessa, PhD, Professor and Director, School of Criminal Justice, University of Cincinnati;
- David Robinson, PhD, Director of Implementation and Development–Assessment, Orbis Partners, Inc.;
- Aron Shlonsky, PhD, Factor-Inwentash Chair and Associate Professor, University of Toronto School of Social Work;
- Jennifer Skeem, PhD, Professor, Departments of Psychology and Social Behavior and Criminology, Law, and Society, University of California, Irvine; and
- Claus Tjaden, PhD, Founder and Senior Partner, Martinez Tjaden, LLP.

Midway through the project, Robert (Barney) Barnoski, PhD, formerly with the Washington State Institute for Public Policy (retired) and adjunct faculty at Washington State University, was added to the board at the request of Mr. Hosman.

F. Measures

The risk assessment instruments evaluated in this study range from simple additive scales contained on a single page to risk assessments completed as part of a more comprehensive system that includes needs assessment. The study sites also differed in philosophy, policies, and procedures. These types of differences have the potential to result in substantial variance in services provided to youth in the justice system and in recidivism rates reported in each jurisdiction. Study methods, described below, attempted to account for these differences.

It is important to recognize the role of base rates (in this report, recidivism rates serve as the base rates). Base rates are the overall recidivism rates observed in each state or county studied and can have a profound impact on the ability to construct valid risk assessment instruments (Gottfredson, 1987). Of the seven probation agencies represented in the study, five had remarkably similar rates of new adjudications reported in the 12-month follow-up period. Oregon, however, reported a rate that was less than half those reported in Nebraska, Arizona, Georgia, Virginia, and Florida. The Solano County, California, cohort used for this study had a much higher base rate than other participating jurisdictions.

Base rates can be affected when an agency screens out or diverts low-risk offenders because the practice essentially results in assessing a higher-risk group, i.e., higher recidivism rates are often observed for those who enter the system. On the other hand, if an agency assesses all cases referred to juvenile court, recidivism rates for these offenders will generally be lower than those observed for an agency that systematically screens out low-risk offenders.

1. Reliability

a. *Calculating Reliability*

Inter-rater reliability is a measure of consistency among workers responsible for completing assessment instruments. Inter-rater reliability can be evaluated by a straightforward measure: percent agreement among raters. This measure is intuitive and has been used extensively in other fields, such as studies of assessment instruments used in child welfare (see, for example, Coohey, Johnson, Renner, & Easton, 2013). Percent agreement among raters indicates how often raters arrive at the same score for each risk factor and for the overall score. Additionally, percent agreement with scores from local experts who have extensive experience administering the assessment provides an indication of the degree to which raters' selections were correct (assuming the expert correctly scores items on the instrument). Inter-rater reliability is calculated with the following formula:

$$\text{Average percent agreement for item } i = \frac{(a_1 + a_2 + \dots + a_n)}{(r_1 + r_2 + \dots + r_n)}$$

where a is the number of raters who agreed with the most common response for item i on each vignette, n is the total number of cases completed for item i , and r is the number of raters on each vignette for item i .

Percent agreement with expert scores is calculated by summing the number of ratings that matched the expert rating across the study cases, then dividing by the total number of ratings.

$$\text{Percent agreement with expert for item } i = \frac{(e_1 + e_2 + \dots + e_n)}{(r_1 + r_2 + \dots + r_n)}$$

where e is the number of raters who agreed with expert score for item i on each vignette, n is the total number of cases completed for item i , and r is the number of raters (excluding the expert) on each vignette for item i .

Kappa, a standardized measure regularly used to measure item inter-rater reliability, tests whether levels of agreement exceed agreement that might occur by chance. A kappa of 0 means that actual agreement is equal to the agreement that would be expected to occur by chance. A positive

kappa indicates a level of agreement greater than what can be accounted for by chance. A kappa of 1 represents perfect agreement among raters. Cohen's kappa applies to two raters, whereas Fleiss' kappa is the recommended statistic for categorical measures and more than two raters (Landis & Koch, 1977a). Using the standardized kappa approach facilitates the comparison of reliability across risk assessment instruments.

The kappa, however, has limitations. Fleiss' kappa can vary with changes in prevalence rates even in the presence of a high rate of actual agreement (Uebersax, 1987; Rodella, 1996). The standardized kappa is also limited to its assumptions about the role and likelihood of chance, for example, that raters make decisions by chance or that workers in practice settings would score the risk assessment by randomly selecting responses.

Kappas are calculated using the following formula:

$$\kappa = \frac{P - P_e}{1 - P_e}$$

where $P = \frac{1}{N} \sum_{i=1}^N P_i$; $P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1)$; $P_e = \sum_{j=1}^k p_j^2$; and $p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}$ and N is the number of cases, n is the number of raters, and k is the number of risk categories.

Another standardized measure of agreement is the intra-class correlation coefficient (ICC). The ICC attempts to minimize the effect of rater patterns by accounting for the magnitude of difference between raters' scores. In other words, the ICC measures the degree to which two raters have the same or *nearly* the same ratings. The ICC can be applied to risk scores and levels.

The ICC compares the variance of different ratings of the same case to the total variation across all ratings and all cases. The ICC attempts to account for the absolute differences in rater patterns, which can minimize the effect of rater patterns on the coefficient. One limitation of this measure is that it is possible to obtain a high ICC when the level of actual agreement is low; for

example, when one rater consistently ranks more severely than another. In addition, high correlations can be attained when the number of rankings possible is limited (e.g., a scale that only included three levels: high, moderate, and low risk).

ICC can be calculated using a two-way, mixed-effects analysis of variance (ANOVA) model, with raters as a fixed factor and agreement defined as absolute. The formula is:

$$ICC(2,1) = \frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$$

where BMS is the between mean square for cases, EMS is the error mean square within raters, JMS is the mean square within raters, k is the number of raters, and n is the number of cases.

To accommodate the strengths and weaknesses of each of these measures, inter-rater reliability was tested in this study using percent agreement, kappa, and ICC. Statistics were computed in the Statistical Package for the Social Sciences (SPSS) or the R software package.

b. Methods Used to Study Reliability

Examining inter-rater reliability typically begins with the construction of cases by: (1) procuring case files from each site and redacting identifying information; (2) creating case vignettes augmented with videotaped interviews and any other information required to complete a risk assessment; or (3) using a hybrid of actual cases and augmenting them with information needed to serve as the basis for risk ratings. In all instances, raters assess cases using the same information.

The current study included videotaped interviews augmented by file information including the offense, prior delinquency, and other factors not covered in the interview. The approach was constant across all sites, and information provided to participants was sufficient to score most of the factors contained in each risk assessment instrument. The cases used did present some limitations for some of the sites, particularly sites that exclusively serve youth committed to secure care facilities. The

interviews used were with youth recently placed on probation and did not include youth placed in state facilities. In addition, because these interviews were used across sites, not all questions could be posed in the exact manner in which staff in all sites were trained and/or are accustomed.

Reliability test cases consisted of videotaped interviews with 10 youth who were clients of a private service provider in the southern United States. To protect youth identity, each youth chose an alias to use during the interview and was instructed to not disclose any personally identifying information such as real name, date of birth, sibling names, and/or addresses. All youth were involved in the juvenile justice system at the time of the interview, and all volunteered to be interviewed. Each youth and/or his/her parent/guardian signed forms consenting to the use of the videos for purposes of this study and other training exercises.

The interview questions were designed to gather information related to the various items/domains from each of the risk assessment instruments included in the study. To ensure that videotapes and supporting documentation contained all information necessary to score all instruments, NCCD researchers identified similarities and differences across the instruments. Questions posed to youth in the interviews reflected a compilation of questions and/or items across all instruments in the study. To the extent possible, all questions from all instruments in the study were included in the interviews.⁴ In addition, any questions/items that could not be incorporated into the interview were provided to study participants in an electronic file. The file listed each youth's offense history along with other information not covered by the interview questions (e.g., number of stays in detention).

Interviews were conducted by two individuals: the executive director of a private, youth-serving agency who had more than 12 years of experience in juvenile justice; and a senior NCCD staff member who was not a member of the project team. This individual had more than 10 years of

⁴ While every effort was made to capture the necessary information in the interviews, not all risk assessment questions could be addressed.

experience in juvenile justice and routinely interviews justice system-involved girls. An NCCD senior researcher assigned to the study observed the interview process to ensure that all youth understood participation was voluntary and they could opt out at any time.

Cases included in the study consisted of interviews with six boys, ranging in age from 13 to 17, and four girls, who ranged in age from 14 to 18. Six of the youth were Black/African American, two were White, one was Hispanic/Latino, and one was an American Indian youth.

In small sites, all staff who routinely complete risk assessments participated in the study as raters. In larger sites, approximately 50 randomly selected staff participated. The two exceptions were Arizona DJC and Virginia DJJ. In Arizona, specialized juvenile justice practitioners complete various portions of the risk assessment, which is embedded in a larger comprehensive assessment of youth needs and functioning in a variety of domains. Each specialist is responsible for scoring items related to his/her specialty and none of the specialists were trained to complete all portions of the assessment. Because of this, a limited number of DJC staff members who were trained and familiar with scoring most of the assessment instrument participated in the study. In Virginia, current agency priorities prohibited a random selection of workers from participating. Rather, Virginia staff who routinely complete the risk assessment were asked to volunteer and per office help ensure appropriate representation. Roughly 16% (80 out of approximately 500 staff) volunteered to participate.

Overall, the majority of raters were women; the average age of raters was 39 years; and most were White. Staff had spent an average of nearly 12 years working in the juvenile justice field, and nearly all had earned a post-secondary degree (Table 2).

Table 2							
Reliability Study Participant Raters Demographics							
Site	# Staff	Gender		Average Years of Juvenile Justice Experience	Average Age in Years	Most Prevalent Race	Percent With Post-Secondary Degree
		Male	Female				
Arizona AOC	46	43.5%	56.5%	13.7	43	White (63.0%)	100.0%
Arizona DJC	6*	33.3%	66.7%	15.6	42	White (66.7%)	50.0%
Arkansas	18*	44.4%	55.6%	14.9	37.5	Black (61.1%)	83.3%
Florida	51	33.3%	66.7%	12	43	White (51.0%)	100.0%
Georgia	54	38.9%	61.1%	8.6	37	Black (53.7%)	96.3%
Nebraska OJS	48	18.7%	81.3%	6.5	33	White (81.3%)	100.0%
Nebraska Probation	28	28.6%	71.4%	6.9	34	White (82.1%)	100.0%
Oregon	46	41.3%	58.7%	14.8	41	White (91.3%)	91.3%
Solano County	27*	18.5%	81.5%	9.5	40	White (40.7%)	100.0%
Virginia	76	22.4%	77.6%	14.5	40	White (60.5%)	100.0%
TOTAL	400	31.5%	68.5%	11.7	39	--	97.0%

Note: Reliability study participant raters were asked to complete a survey; table results are based on staff who completed the survey.

*All staff participated.

All cases in the reliability study were scored by an expert or a team of expert scorers in each jurisdiction to create an answer key for each case in the study. Staff scores were then compared to expert scores to provide a measure of the degree to which staff scored the risk assessment instrument correctly (i.e., consistent with the expert scoring). Experts in each jurisdiction consisted of staff with extensive training and/or knowledge of the risk assessment instrument and its use in the jurisdiction. Some sites identified one expert and others used a team of experts to score each case in the study. Most experts had been with the jurisdiction for an extended period of time, were former field staff, and had extensive experience training the assessment. Expert scorer qualifications can be found in Appendix D.

Reliability testing was conducted between December 2011 and April 2012. Materials included an online version of each instrument, a set of 10 videotaped interviews, and an offense history file to accompany each videotape. Advisory board members and representatives from each site in the study were given the opportunity to review all materials prior to the study start date. Videotaped cases and associated offense history files were posted to a secure website created specifically for this study. The site also contained a link to the online version of the instruments. Prior to the study start, participants were trained via a web-based conference to access the secure website, view the videos and offense history files, and complete the online version of the instrument. Workers were given four weeks to complete risk assessments for all 10 cases. Risk and needs assessment items were tested for inter-rater reliability.

To provide context for comparing the relative inter-rater reliability results, a minimum threshold of 75% agreement was established. The threshold is artificial but easy to understand—it can be interpreted as three of four people agreeing. This threshold was applied to percent agreement with risk levels, risk items, and expert scores.

While researchers do not all agree on acceptable thresholds for the ICC⁵ or kappa,⁶ the following ranges offer guidelines for interpreting results.

Table 3			
ICC and Kappa Inter-Rater Reliability Thresholds			
ICC		Kappa	
0 to 0.2	Poor	<0.2	Poor
0.3 to 0.4	Fair	0.21 to 0.4	Fair
0.5 to 0.6	Moderate	0.41 to 0.6	Moderate
0.7 to 0.8	Strong	0.61 to 0.8	Good
>0.8	Almost perfect	0.81 to 1.0	Very good

⁵ http://www.statstodo.com/ICC_Exp.php

⁶ www.medcalc.org/manual/kappa.php

2. Validity

A validation study measures risk assessment instrument performance on a population that differs from the one used to construct the instrument, or in cases where no construction sample is available, to measure the instrument's performance in the agency where it was implemented. In general, validity can be understood as the extent to which an instrument measures what it is intended to measure. For this study, validity can be understood as the extent to which risk classification and items contained in a risk assessment instrument relate to observed recidivism.

Several methods exist for measuring validity. A useful and intuitive measure is the level of separation in recidivism results attained between groups at various risk classifications and whether offenders are grouped into risk classifications of meaningful size (Gottfredson & Snyder, 2005). The combination of separation (or discrimination between recidivism rates for each classification) and the distribution of cases across the risk continuum is a meaningful measure of the risk assessment instrument's performance in practice.

The Dispersion Index for Risk (DIFR) is a measure of risk assessment accuracy that adjusts for sample size and evaluates classification levels (Silver & Banks, 1998). The DIFR assesses the performance, or "potency," of a risk assessment instrument by assessing how an entire cohort is divided by risk and the extent to which group outcomes (e.g., low, moderate, high) vary from the base rate for the entire cohort. In essence, it weights the distance between a subgroup's outcome rate and the cohort's base rate by the subgroup size to estimate the potency of an instrument. Because this measure considers proportionality and differences in outcome rates among several subgroups, it is a useful measure of the efficacy of a multilevel classification system. The DIFR formula is:

$$DIFR = \sqrt{\sum_{i=1}^k \left(\ln\left(\frac{P}{1-P}\right) - \ln\left(\frac{p_i}{1-p_i}\right) \right)^2 * \frac{n_i}{N}}$$

where k is the number of subgroups in the risk classification model, P is the total sample base rate of the outcome, N is the total sample size, p_i represents the base rate of each of the k subgroups, and n_i is the size of each k subgroup.

A limitation of the DIFR is that it measures relative distance in outcomes between groups. Thus, an instrument used in a jurisdiction with a low rate of recidivism may have a higher DIFR score than an instrument used in a jurisdiction with a higher rate of recidivism, even when the latter instrument provides far greater separation in absolute terms.

Another measure of validity is the receiver operating characteristics (ROC) curve. The ROC tests accuracy by plotting the true positive rate (sensitivity) and true negative rate (1 – specificity) for each risk score (Zweig & Campbell, 1993). The ROC curve represents the range of sensitivities and specificities for a test score.

The area under the curve (AUC) is a single measure used to compare ROC curves (Liu, Li, Cumberland, & Wu, 2005; Zweig & Campbell, 1993). It represents the probability that the value of a positive case (future delinquency) will exceed the value of a negative case (no future delinquency).⁷ A strength of the AUC is that its results are easy to interpret—the greater the AUC, the greater the accuracy of the instrument. The AUC is limited, however, in that it is possible to have a high AUC when the vast majority of people are classified to a single risk level. Attempts to standardize interpretations of the AUC vary (Royston, Moons, Altman, & Vergouwe, 2009; Tape, n.d.) and must take the base outcome rate into account (Rice & Harris, 2005).

An AUC of 1 represents a perfect test (i.e., 100% accurate); an AUC of 0.5 represents a worthless test (only accurate 50% of the time, the same as chance). No standard exists for interpreting the strength of the AUC; however, some researchers have employed the following point system, much like the one used in traditional academic grading (Tape, n.d.):

⁷ The AUC equals the probability that a randomly selected youth who has committed a new offense (a positive) will score higher than randomly selected youth who did not recidivate (a negative).

- 0.90–1 = excellent
- 0.80–0.90 = good
- 0.70–0.80 = fair
- 0.60–0.70 = poor
- 0.50–0.60 = fail

Some researchers have suggested that a point system that mirrors academic grading is too stringent. In at least one study, researchers suggested that risk assessment instrument AUCs of 0.70 or higher are acceptable (Van Der Put et al., 2011); others have suggested that, in general, AUCs greater than 0.75 are strong (Dolan & Doyle, 2000). A recent study (Schwalbe, 2007) examined AUCs from 28 studies of risk assessment and found that the average AUC was approximately 0.64. None of these studies, however, suggested thresholds for interpreting the strength of the AUC.

The primary measure of validity used in this study was the degree to which each instrument was able to discriminate between groups of youth with higher and lower rates of recidivism and the distribution of cases across the risk continuum. In addition to discrimination and distribution, two summary statistics were used to provide overall estimates of scale validity: AUC and DIFR. As noted, both of these measures have limitations; however, because the AUC is frequently used as a primary measure of validity in other studies of risk assessment and because the DIFR considers both the degree of discrimination attained and the proportion of cases at each risk level, they were included to augment overall understanding of the relative validity of each risk instrument.

The validation study was based on samples of youth assessed in the sites between 2007 and 2009, with some variation. Recidivism was observed for a 12-month follow-up period, except in Arkansas, where follow-up was limited to nine months.

Several measures of recidivism were used based on data available in each jurisdiction. In most instances, recidivism measures included new arrests, new adjudications, and subsequent placement in a correctional facility. However, data on each of these measures were not available in every participating agency. The primary outcome used in the current study was subsequent adjudication;

however, in two jurisdictions, the only measure available for youth in the commitment cohort was “return to an institution.” Differences in outcomes are noted in the report when relevant.

Sample sizes ranged from 119 in Arkansas to more than 27,000 in Florida. The sample period, size, and outcomes used in each site are shown below in Table 4. (Note that Virginia was in the process of phased implementation; sample size reflects less than one third of youth placed on probation during the sample timeframe.)

Table 4					
Sample Descriptions					
Risk Assessment Instrument	Year of Implementation	Sample Period	Sample Size	Sample Description	Outcomes Examined
Arizona AOC Risk Assessment Instrument	2000	July 2007 – June 2008	7,589	Probation start	Complaint, petition, adjudication
Arizona DJC DRI	2007	July 2007 – June 2008	1,265	Releases from secure care	Commitment
YLS/CMI					
Arkansas*	2008	July 2008 – September 2009	119	Releases from secure care	Commitment
Nebraska Probation	2002	June – December 2009	1,077	Probation start	Offense, offense with sanctions, criminal offense with sanctions
Nebraska Commitment	2002	2008–2009	597	Releases from secure care	Petition, adjudication, commitment
PACT					
Commitment	2006	July 2007 – June 30, 2010	11,154	Releases from secure care	Arrest, adjudication, commitment
Probation	2006	July 2007 – June 30, 2009	27,369	Probation end	Arrest, adjudication, commitment
CRN					
Probation	2001	2008	5,695	Probation start	Arrest, adjudication, commitment
Commitment	2001	2008	469	Releases from secure care	Arrest, adjudication, commitment
Oregon JCP	2000	2007–2008	12,370	All youth assessed with JCP	Offense, adjudication

Table 4					
Sample Descriptions					
Risk Assessment Instrument	Year of Implementation	Sample Period	Sample Size	Sample Description	Outcomes Examined
Public Domain Risk Assessments					
JSC Boys	2007	May 2007 – December 2009	880	Probation start	Offense, adjudication
Girls Link	2007	May 2007 – December 2009	261	Probation start	Offense, adjudication
Virginia YASI	2008	July 2008 – June 30, 2009	1,919	Probation start	Arrest, conviction

*The standardized follow-up period for Arkansas sample cases was nine months; thus, the outcome was re-committed or not within a standardized nine-month period. Ultimately, because of limitations in the number of cases available in Arkansas and the short follow-up period, little could be deduced regarding this application of the YLS/CMI. For all other sites, outcomes were observed for a standardized, 12-month follow-up period.

a. *Construction and Validation of Revised Risk Assessments*

One question explored by the current study was whether *longer* risk assessment instruments (instruments that include additional goals and objectives and, hence, more items) might introduce “noise” into risk assessment and, as a consequence, reduce discriminatory power. Therefore, using data available in each jurisdiction, simple actuarial instruments were constructed to determine if classification results attained with the risk assessment instrument currently in use might be improved.

To develop revised instruments in sites with sufficient sample sizes, the cohort was divided into construction and validation samples. Because classification results are nearly always more robust for the sample from which a risk assessment instrument has been constructed (because the instrument is essentially tailored to that sample), revised instruments were developed using a construction sample and tested on a validation sample. If sample sizes were not adequate, the risk instrument was constructed on a single sample. For additional detail, see Appendix B.

The analysis did not, however, address the question of what type of instrument might best transfer to other jurisdictions. Therefore, to test the idea that simple actuarial systems might transfer better than more complex instruments and/or systems derived via non-actuarial methods, the study

simulated the use of the JSC instrument in two of the largest sites, Florida and Georgia. Because the simulation was not part of the original study design, results are not included in the findings section and are instead included in the discussion section.

3. Equity

The goal of many agencies is that a risk assessment instrument will work equally well for different racial and ethnic groups and across genders. Validity of these instruments should be established for these population subgroups and taken into consideration when determining policies and procedures that affect individual youth. Efforts to improve equity can help reduce disproportionate minority contact in the juvenile justice system. Youth of color enter all levels of the system at higher rates than White youth (Short & Sharp, 2005; NCCD, 2011; Hartney & Silva, 2007; Pope, Lovell, & Hsia, 2002).⁸ It is critical for assessment processes to treat all groups equitably.

Youth from different racial and ethnic groups labeled high, moderate, or low risk often have different rates of recidivism. When high-risk offenders from one group have similar (or lower) recidivism rates compared to moderate-risk offenders from another racial or ethnic group, the potential for biased decisions increases. The potential consequence is that risk classifications assigned to youth do not accurately represent the overall base expectancy rates used to define and differentiate risk levels.

In addition, there is growing evidence that separate instruments may be required to optimize classification results for girls (Van Voorhis, Salisbury, Wright, & Bauman, 2008; Ereth & Healy, 1997). Recent efforts to improve assessment instruments for girls have attempted to address growing concerns that standard assessment protocols fail to identify issues critical to providing care for girls in the juvenile justice system (Shepherd, Luebbers, & Dolan, 2013).

⁸ A recent study of detention procedures in Cook County, IL, for example, found that Black/African American youth were 46 times more likely to be detained than White youth (NCCD, 2011).

Equity is the degree to which a risk assessment instrument measures outcomes the same way across subgroups (i.e., what “high risk” means for boys and girls and across major race and ethnicity groups). Because equity is an essential measure of instrument validity, validity results presented for each instrument evaluated are delineated by gender, race, and ethnicity. Discrimination, distribution, AUC, and DIFR were employed to assess the equity of each instrument for subgroups and are included when sample sizes were sufficient.

Additional information on all outcome measures across jurisdictions, including comparisons by race and ethnicity and risk assessment item analyses, can be found in Appendix B.

4. Cost

Given the overall objective of this project, sufficient resources to conduct an extensive cost/benefit analysis were not available. Nevertheless, the estimates provided might help agencies determine the best approach considering their needs and circumstance.

Administrators from each of the study jurisdictions provided cost information for each instrument via phone interviews. An NCCD interviewer with experience working in a governmental agency and conducting research interviews conducted all phone interviews using a standardized interview protocol developed for this study. Interviews were conducted in September 2012.

III. FINDINGS

The following section reviews findings from the examination of the reliability, validity, equity, and cost. In addition, it describes results of revisions made to risk instruments currently in use to determine if classification results might be improved.

To construct revised instruments, samples of 2,000 or more cases were divided into construction and validation samples. (The results from validation samples are considered to be a

better indicator of how the instrument will perform in practice.) However, dividing fewer than 1,000 cases into three or four risk levels further delineated by race, ethnicity, and gender can be problematic. The number of cases in these breakdowns is often too small to produce stable, representative statistics. Both construction and validation samples were used in Florida, Georgia, and Arizona AOC (probation).

A. Findings by Risk Assessment Instrument

1. The Georgia CRN

The CRN was developed by Tim Brennan, PhD, of Northpointe, and Claus Tjaden, PhD, of Martinez Tjaden, LLC. In its original form, it was a derivative of COMPAS Youth, a risk assessment instrument developed by Northpointe. The CRN was tailored for Georgia DJJ to aid in making decisions related to security as well as to assess youth criminogenic need factors. It is composed of 27 scales across the following five domains: usual behavior and peers, personality, substance abuse and sexual behavior, school and education, and family and socialization. These domains are used to classify youth as low, medium, or high risk. The centerpiece of the CRN is the interview process with the youth, though additional collateral information is also considered.

The most recent validation of the CRN was conducted in 2006 by Tjaden. The study found that the CRN effectively classified youth by risk level in that high-risk youth reoffended at a higher rate than youth classified as low risk. The CRN was found to have moderate predictive ability, as evidenced by an AUC value of 0.61.

More than 150 items are included in the CRN, but two factors—the age of the youth at first adjudication and the number of prior adjudications—account for two thirds of the possible point total. Of the 150 remaining items, about one third contribute to risk scoring. Combined, they account for only three of nine possible risk points. Theoretically, these variables can account for a maximum of

60% and a minimum of 14% of the total risk score. On average, they comprise 34% of the total score. An automated scoring process statistically transforms these factors once, transforms them again into normalized values, and then finally aggregates the resulting values into a three-point scale.

The remaining items on the CRN are collected to assess needs for case planning. Each factor is rated on a four- to five-point Likert scale reflecting either the observed frequency or the severity of a behavior or characteristic. The CRN was piloted in Georgia in 2001 and fully implemented in 2002.

The overall results of the validity study are presented in Table 5.

Table 5 Georgia CRN New Adjudications by Risk Level Current Cut Points				
Risk Level	Probation Cases (N = 5,698)		Committed Youth (N = 469)	
	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated
Low	88%	25.3%	25%	23.9%
Moderate	11%	52.4%	37%	43.3%
High	1%	57.5%	39%	45.9%

As shown in Table 5, there is substantial separation in rates of recidivism reported for low- and moderate-risk cases, but little difference between rates observed for moderate- and high-risk youth. This is true for both probation cases and for youth released from state facilities. Second distribution of probation cases across risk levels is limited. Most (88%) of all probation cases are rated low risk, while only 1% are rated high risk. In essence, only two levels of risk are being identified. The limited distribution of cases across risk levels reduces the value of the system to probation, and probation cases represent about three fourths of all cases assessed.

The overall AUC for the current CRN was 0.64. The DIFR score was 0.40, reflecting the problems with distribution noted above.

A change in the cut points used to assign risk levels significantly improves the utility of the Georgia CRN. Current cut points are as follows: 3–5 = low risk; 6–7 = moderate risk; and 8–9 = high risk. Modifying these to 3–4 for low risk, 5 for moderate risk, and 6 and above for high risk produces the following results for all cases (Table 6).

Table 6 Georgia CRN New Adjudications by Risk Level Revised Cut Points* (N = 7,412)		
Risk Level	Percent at Each Level	Percent Adjudicated
Low	59%	22.5%
Moderate	18%	36.2%
High	23%	46.9%

*Cut points were altered so that 3–4 = low, 5 = moderate, and 6 or more = high.

While the revised cut points still put a high percentage of cases at the low risk level (particularly for the probation subgroup), the change produces better distribution across risk levels and greater separation of outcomes for moderate- and high-risk cases. In essence, the revision identifies a moderate-risk group that the current cut points do not distinguish.

The current cut points lead to equity issues as well. Although few cases were classified as high risk, both Black/African American and Hispanic/Latino youth who were classified as high risk had lower recidivism rates than those placed at the moderate-risk level. When the entire sample is considered, moderate-risk males had higher rates of recidivism than high-risk males.

When results were broken down by gender, similar patterns emerged. The current instrument does not distinguish well between high- and moderate-risk cases for either boys or girls. Because current cut points lead to equity issues, and at a minimum changes to cut points are needed, additional discussion of equity would be premature.

The combined predictive power of the youth's age at first adjudication and number of prior adjudications was tested without including other risk factors to give a better understanding of how the CRN's risk classification functions. As noted above, these two factors account for 66% of the average risk score in Georgia. When they are combined and the remaining risk factors left out of risk formula, they produce the classifications that clearly distinguish between low-, moderate-, and high-risk cases (Table 7).

Table 7 Georgia CRN New Adjudications by Risk Level Adjudication Score + Age Score N = 7,412		
Risk Level	Percent at Each Level	Percent Adjudicated
Low	46%	21.5%
Moderate	41%	34.8%
High	13%	48.9%

The factors of youth's age at first adjudication and number of prior adjudications account for virtually all of the predictive power of the CRN instrument (see Appendix B for CRN scoring methods). They also create a much larger moderate-risk group and produce slightly better overall discrimination than what was attained simply by revising the original cut points. While the additional 150 items may provide data for case-planning purposes, they impact risk classification very little.

Selecting factors with the highest correlations with outcomes allowed for the creation of a simple additive risk index to test whether results attained with age and the number of adjudications might perform better. Differences in risk factors for boys and girls led to the creation of two revised risk assessment instruments. These are presented on the following pages.⁹ Results by risk level are presented in Table 8.

⁹ As stated earlier in this report, this analysis was undertaken to investigate the potential for improving the risk assessment instrument. The analysis was limited to variables collected and categorized by the risk model currently in place. Further improvements are possible with the introduction of additional factors to the test.

Georgia Department of Juvenile Justice
Revised Risk Assessment for Community-Placed Boys

	<u>Score</u>
1. Age at first adjudication	
a. 15 or older, or no prior adjudications	0
b. 14 or younger	1
2. Number of arrests prior to current arrest	
a. None.....	-1
b. One or two.....	0
c. Three or more.....	1
3. Most serious current offense was property related	
a. No.....	0
b. Yes.....	1
4. Youth had conflicts with teachers	
a. No.....	0
b. Yes, either known or suspected	1
5. Number of classes youth failed	
a. None.....	-1
b. One or two.....	0
c. Three or more.....	1
6. Number of times youth suspended since first grade	
a. 0–3 times	-1
b. 4–6 times	0
c. 7+ times	1
7. Youth argues or fights with other students	
a. No.....	0
b. Yes, either known or suspected	1
8. Characteristics of youth’s friends	
a. None apply	0
b. One or more apply (mark all that apply and add)	
___ At least some of youth’s friends are gang affiliated	1
___ More than half of youth’s friends have been arrested	1
9. Characteristics of youth	
a. None apply	0
b. One or more apply (mark all that apply and add)	
___ Youth does not participate in any sports, church, creative, or school activities	1
___ Youth has used marijuana at least once in the last three months.....	1
___ Youth has used alcohol at least one time per week for the last three months	1
Total Risk Score	_____

<u>Risk Score:</u> ___ -3–0 ___ 1–4 ___ 5–12	<u>Risk Level:</u> Low Medium High
--------------------------------------------------------------	----------------------------------------------------

**Georgia Department of Juvenile Justice
Revised Risk Assessment for Girls**

	<u>Score</u>
1. Number of arrests prior to index arrest	
a. None.....	-1
b. One or two.....	0
c. Three or more.....	1
2. Number of prior adjudications for property offenses	
a. None.....	0
b. One or more.....	1
3. Age at index arrest	
a. 11 or under, 17 or older.....	-1
b. 12 to 16.....	0
4. Number of times youth suspended since first grade	
a. 0–3 times.....	-1
b. 4–6 times.....	0
c. 7+ times.....	1
5. Youth had conflicts with teachers	
a. No.....	0
b. Yes, either known or suspected.....	1
6. Youth participates in activities	
a. Youth participates in at least one sport, church, creative, or school activity.....	0
b. Youth does not participate in any activities.....	1
7. Youth’s parent(s) knows who youth’s friends are	
a. Yes.....	0
b. No, either known or suspected.....	1
8. Family characteristics	
a. None applicable.....	0
b. One or both apply (mark all that apply and add)	
___ Youth’s mother was ever arrested.....	1
___ Youth’s mother was ever in jail or prison.....	1
9. Youth was raised by a single parent	
a. No.....	0
b. Yes.....	1
10. Youth’s friends have been arrested	
a. None or some of youth’s friends.....	0
b. More than half of youth’s friends.....	1
Total Risk Score	

<u>Risk Score:</u> ___ -3 to -1 ___ 0–3 ___ 4–10	<u>Risk Level:</u> Low Medium High
------------------------------------------------------------------	----------------------------------------------------

Table 8				
Georgia Revised Risk Assessment Instrument				
Risk Level	Boys		Girls	
	Percent at Level	Subsequent Adjudication	Percent at Level	Subsequent Adjudication
Low	32%	17.0%	23%	11.7%
Moderate	44%	37.1%	54%	21.0%
High	24%	49.1%	23%	33.9%
Base Rate	33.4%		21.8%	
Sample Size	2,506		2,005	
AUC	0.67*		0.64*	
DIFR	0.61		0.46	

*AUC significantly different from 0.50.

Note: Results for boys reflect the validation sample; results for girls are based on single sample.

As the data presented in Tables 8 and 9 illustrate, the revised instrument worked well across all major ethnic and racial groups in Georgia and both genders. Both AUCs and DIFR scores increased across the board for the revised scale. Results delineated by gender are presented in Table 8 and by race/ethnicity in Table 9.

Table 9								
Georgia Revised Risk Assessment Instrument New Adjudications by Risk Level Boys								
Risk Level	All Cases		Hispanic/Latinos		Whites		Black/African Americans	
	% at Level	Subsequent Adjudication	% at Level	Subsequent Adjudication	% at Level	Subsequent Adjudication	% at Level	Subsequent Adjudication
Low	32%	17.0%	40%	6.8%	40%	13.9%	26%	22.2%
Moderate	44%	37.1%	34%	39.4%	45%	30.4%	44%	42.9%
High	24%	49.1%	26%	41.4%	15%	47.7%	30%	50.7%
Base Rate	33.4%		27.0%		26.4%		39.7%	
Sample Size	2,506		111		1,014		1,346	
AUC	0.67*		0.73*		0.68*		0.64	
DIFR	0.61		1.12		0.63		0.50	

*AUC significantly different than 0.50.

Note: Reflects validation sample.

Results for the revised instrument for girls were not as strong, although substantial separation of outcome rates by risk level was attained (Table 10).

Table 10 Georgia Revised Risk Assessment Instrument New Adjudications by Risk Level Girls						
Risk Level	All Cases		Whites		Black/African Americans	
	Percent at Level	Adjudication	Percent at Level	Adjudication	Percent at Level	Adjudication
Low	23%	11.7%	33%	10.4%	14%	14.7%
Moderate	54%	21.0%	52%	16.2%	57%	24.3%
High	23%	33.9%	15%	27.0%	29%	37.0%
Base Rate	21.8%		15.8%		26.6%	
Sample Size	2,005		833		1,080	

Note: The number of Hispanic/Latino girls in the sample was too small for independent analysis. These girls are, however, represented in the “all cases” statistics. Reflects results from a single sample.

a. Summary of Findings

A minor change—altering the cut-off points used to assign risk levels—was found to improve both the distribution of cases across risk levels and the power of the instrument. The CRN’s large number of factors and complex scoring system did not appear to help the instrument produce better results. Simple additive scales using variables selected from the CRN produced better classification and gender equity results than the current instrument.

2. Solano County JSC and Girls Link Risk Assessments

The Solano County Probation Department (Juvenile Division) uses public-domain, gender-specific risk assessment instruments. The instruments are composed of eight to 10 items, which include youth’s age at first referral to juvenile court, school discipline/attendance, substance use, peer

relationships, and parent/sibling criminality. The instruments are embedded in the Juvenile Assessment and Intervention System™ (JAIS).

The risk assessment instrument for boys was created by the National Council of Juvenile and Family Court Judges (NCJFCJ) in partnership with NCCD as a model instrument for NCJFCJ's JSC, an OJJDP-supported initiative (Wiebush, 2002). The basic elements of the JSC risk assessment instrument for boys have been validated in more than a dozen agencies across the United States; the items in the instrument are a composite of those that appear in those instruments (Wiebush, 2002). The risk assessment instrument for girls was developed by NCCD in 1997 for the Cook County (Chicago), Illinois, Girls Link program. Both risk instruments are available in the public domain.

At the time of the current study, the JSC and Girls Link instruments were used at two points in Solano County. First, the county employed a paper copy of the instrument to screen youth coming through juvenile court. Youth who scored in the low risk category—with the exception of those adjudicated for specific felony offenses—were typically placed on informal probation. Minors at the low risk level were routinely contacted by the agency (i.e., at least once every three months) once they completed the court process and were assigned to a caseload. Offenders who scored moderate to high risk received a full risk and needs assessment.

This practice of conducting a pre-screen and then conducting the full assessment for only some affected the study in two ways. First, because most low-risk offenders do not enter probation, the number of cases at the moderate and high risk levels was disproportionately high. Second, this policy also produced an artificially high base rate relative to other probation departments represented in this study. The ideal solution would be to obtain the risk scores for diverted cases and include them in the study to determine if they reoffended; however, these records were not available. Still, of the 1,141 cases in the study sample, about 15% were low risk. In most cases, these were youth committed to probation for felony offenses.

The follow-up period used to test validity was either 12 months from the date of admission to probation or 12 months from the date of assessment if the two dates did not align. Table 11 presents the overall combined results for both the boys' and girls' instruments.

Table 11 Solano County JSC and Girls Link Risk Assessment Instruments New Adjudications by Risk Level		
Risk Level	Percent of Cases at Level	Percent Adjudicated
Low	15%	20.0%
Moderate	47%	42.4%
High	39%	63.4%
Base Rate		47.2%
Sample Size		1,141

As Table 12 illustrates, the instruments produced substantial separation in outcomes by risk level. The county's policy of screening out low-risk offenders resulted in a somewhat skewed distribution of cases across risk levels, which may have lowered the DIFR scores. The overall AUC for this risk assessment instrument was 0.68 and the DIFR computed for all cases was 0.68.

As illustrated, the instruments produced strong results for major race and ethnicity groups represented in the Solano County population.

Table 12 Solano County Probation Department JSC Risk Assessment Instrument Recidivism by Risk Level for Boys								
Risk Level	All Cases		Hispanic/Latinos		Whites		Black/African Americans	
	Percent at Level	Adjudication	Percent at Level	Adjudication	Percent at Level	Adjudication	Percent at Level	Adjudication
Low	15%	18.8%	13%	25.0%	21%	20.9%	11%	13.6%
Moderate	43%	47.9%	43%	51.0%	40%	37.0%	44%	53.1%
High	43%	64.4%	43%	65.4%	39%	61.5%	44%	64.0%
Base Rate		50.7%		53.3%		43.1%		53.6%
Sample Size		880		240		202		394

Despite differences in overall rates of reoffending among racial and ethnic groups, the JSC instrument effectively separated youth in each cohort into low, moderate, and high risk groups. AUCs ranged from 0.65 to 0.68, and DIFR scores ranged from 0.55 to 0.73.

Because the risk assessment instruments are actuarial scales, efforts at improvement represented customization to better reflect the probation population in Solano County. While slightly better discrimination was attained with minor changes to the boys' instrument, these gains were offset by distribution issues. The results of the revised instrument are available in Appendix B.

The Girls Link instrument did not produce results comparable to those produced for the boys' instrument. The degree of separation attained for outcomes, while substantial, was lower than that attained for boys.

Minor revisions to the Girls Link instrument improved both the level of discrimination attained and the distribution across risk levels. A comparison of results, pre- and post-customization, is presented in Table 13.

Table 13 Solano County Girls Link Risk Assessment Instrument Comparison of Current and Revised Instruments (N = 261)				
Risk Level	Current Risk Assessment Instrument		Revised Risk Assessment Instrument	
	Percent at Level	New Adjudication	Percent at Level	New Adjudication
Low	16%	23.8%	23%	13.6%
Moderate	59%	29.0%	49%	28.3%
High	25%	42.2%	29%	64.0%

a. Summary of Findings

The boys' JSC risk assessment instrument is an effective classification instrument, as evidenced by the degree of separation in outcomes by risk level. For the Girls Link instrument, the degree of separation attained for outcomes, while substantial, was less than that attained for boys. As a result,

modifications reflecting girls sentenced to probation in Solano County substantially improved results. Both instruments worked well across the major racial and ethnic groups in Solano County.

3. Florida PACT

The Florida PACT is a derivation of the Washington State Juvenile Court Assessment (WSJCA), a risk assessment instrument developed in the state of Washington in the 1990s through the Washington State Institute for Public Policy in cooperation with the Washington Association of Juvenile Court Administrators. Slightly different versions of the PACT are used in a number of states and county agencies throughout the country. The WSJCA was validated by the Washington State Institute for Public Policy in 2004. The 27-item pre-screen risk assessment had moderate predictive ability in estimating the likelihood of recidivism with an AUC of 0.64 (Barnoski, 2004).

The PACT was designed to assess juvenile offenders' risks, needs, and protective factors. It incorporates an automated criminal history domain, additional mental health and substance abuse questions, and a case-planning module. The full PACT assessment includes a pre-screening component consisting of 44 items to provide workers with a social and criminal history of each juvenile. The pre-screen determines the risk level assigned to each individual. The full assessment is composed of 126 items across the following 12 domains: criminal history, gender, school, use of free time, employment, relationships, family and living arrangements, alcohol and drugs, mental health, attitudes/behaviors, aggression, and skills. With this information, the PACT is designed to obtain risk factor information as well as assess offenders' needs in order to provide targeted treatment interventions.

The PACT instrument has been validated several times since its original implementation (Baglivio, 2009; see also Baglivio & Jackowski, 2012; Winokur-Early, Hand, & Blankenship, 2012). Each of

these three validation studies found the instrument had a moderate ability to appropriately classify repeat offenders, with AUC values of 0.59, 0.63, and 0.59, respectively.

The PACT instrument in Florida is used for all youth in the juvenile justice system. Ten items are scored, and the sum is used to establish a record of referrals (criminal history) risk level; scores from another 11 items are totaled to reach a social history score. Workers consult the following matrix to assign the youth to his/her risk classification.

**Overall Levels of Risk to Re-Offend
Based on Record of Referrals and Social History Risk Scores**

Record of Referrals Risk Score	Social History Risk Score		
	0 to 5	6 to 9	10 to 18
0 to 5	Low	Low	Moderate
6 to 8	Low	Moderate	Moderate-High
9 to 11	Moderate	Moderate-High	High
12 to 31	Moderate-High	High	High

The current analysis focused on youth placed on probation or sentenced to juvenile facilities. Results for each group are reported separately, consistent with prior evaluations published by the State of Florida.

The most recently published evaluation on probation cases (2012) used a follow-up period that began when the probation episode was closed. To keep the Florida evaluation consistent with those conducted in other sites, the outcome measures used in this study were those observed in the 12-month period following admission to probation. Results are presented in Table 14.

Table 14						
Florida PACT: Probation Cases New Adjudications by Risk Level						
Risk Level	All Cases		Boys		Girls	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	67%	30.0%	66%	31.1%	70%	26.8%
Moderate	18%	44.4%	18%	45.2%	17%	41.7%
Moderate/High	10%	48.8%	11%	49.8%	9%	44.9%
High	5%	57.5%	6%	57.4%	4%	58.1%
Base Rate	35.9%		37.0%		32.3%	
Sample Size	27,369		20,621		6,748	

Most (85%) of all youth placed on probation were classified as low or moderate risk. Only 17% of boys and 15% of girls scored moderate/high or high risk, and most of these fell in the moderate/high range. The PACT pre-screen instrument produced some discrimination between outcomes recorded for low- and moderate-risk cases, but only minor separation between the moderate and moderate/high levels, despite the fact that relatively few cases are placed in these risk categories.

Overall, high-risk youth had an 8.7% and 13.1% higher rate of adjudication than youth at the moderate/high and moderate risk levels, respectively, demonstrating a moderate level of discrimination. The overall AUC for PACT was computed for two subscales: the criminal history risk scale and the social history risk scale. Results from the two scales were combined in a matrix to determine the risk level assigned. The AUC for the criminal history score was 0.59; the social history

scale produced an AUC of 0.63. As noted earlier, AUCs under 0.6 are generally considered poor. The overall DIFR was 0.37.

As illustrated in Table 15, the highest level of discrimination (as well as the most meaningful distribution of cases across risk levels) was found for Black/African American youth. Still, little difference between outcomes was recorded for moderate-risk and moderate/high-risk Black/African American youth. This was true for White youth as well. For Hispanic/Latinos, little difference in recidivism rates was shown for those rated low or moderate risk.

The breakdowns by race/ethnicity also revealed some overlap in outcomes among risk levels. Both moderate-risk Black/African Americans and Whites had higher rates of recidivism than moderate/high-risk Hispanic/Latinos (Table 15).

Table 15						
Florida PACT: New Adjudications by Risk Level						
Probation Sample						
Risk Level	Hispanic/Latinos		Black/African Americans		Whites	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	70%	24.5%	54%	34.3%	70%	28.1%
Moderate	17%	26.4%	22%	46.7%	17%	44.8%
Moderate/High	9%	42.7%	15%	50.9%	8%	47.9%
High	4%	50.0%	9%	63.9%	5%	52.7%
Base Rate	29.2%		40.4%		33.8%	
Sample Size	3,885		4,426		11,664	

When the PACT risk assessment instrument was examined for performance for girls, little difference arose in outcome rates by risk level, particularly between the moderate and moderate/high risk (the equivalent of moderate and high risk in other risk assessment instruments) levels. In addition, overlap in recidivism rates by risk level occurred for every race/ethnicity group, particularly

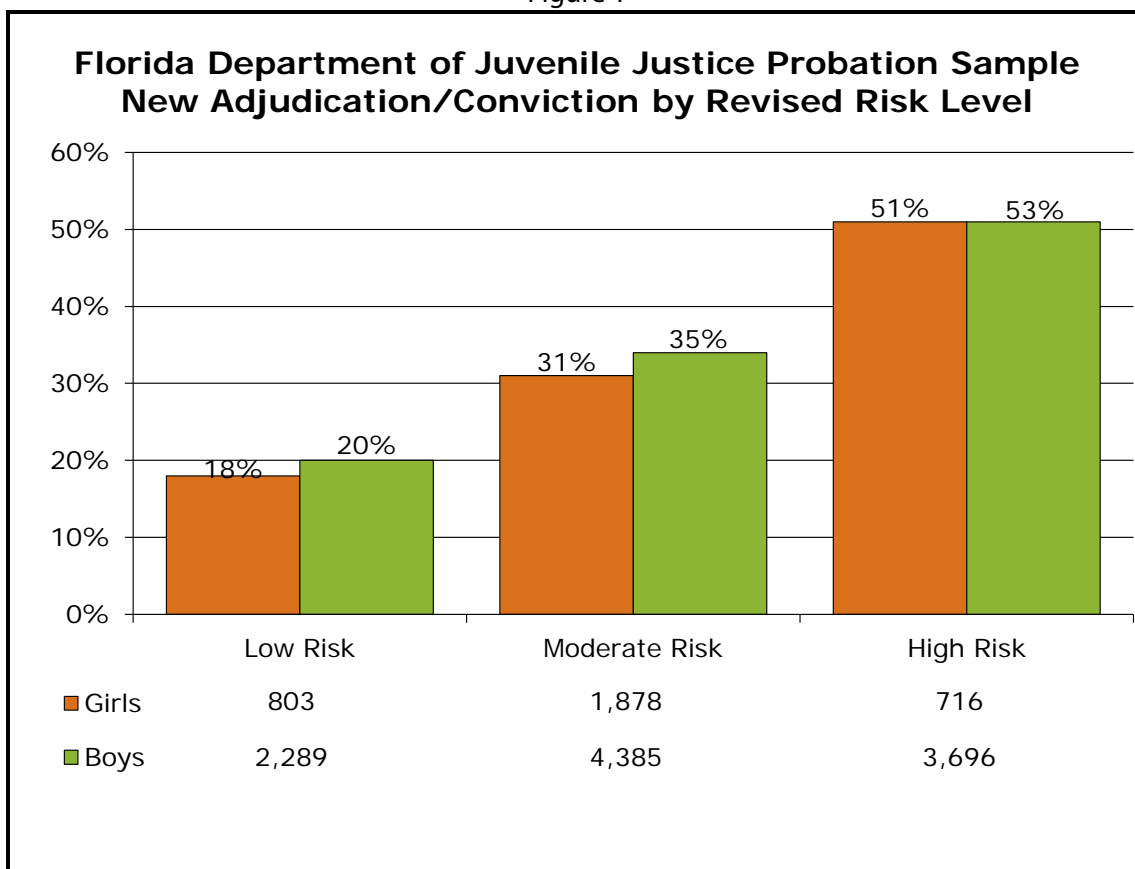
Hispanic/Latina girls. The recidivism rate for moderate/high-risk Hispanic/Latina girls was lower than the rate of recidivism reported for moderate-risk girls of other races/ethnicities (Table 16).

Table 16				
Florida PACT				
New Adjudication by Current Risk Level by Race/Ethnicity				
Girls' Probation Sample				
Race/Ethnicity	N	%	New Adjudication	
			N	%
TOTAL SAMPLE	6,748	100.0%	2,179	32.3%
Black/African American				
Low	1,971	67.7%	553	28.1%
Moderate	510	17.5%	199	39.0%
Moderate/High	278	9.6%	134	48.2%
High	151	5.2%	98	64.9%
Subgroup Total	2,910	100.0%	984	33.8%
White				
Low	2,119	71.8%	573	27.0%
Moderate	490	16.6%	231	47.1%
Moderate/High	228	7.7%	101	44.3%
High	114	3.9%	56	49.1%
Subgroup Total	2,951	100.0%	961	32.6%
Hispanic/Latina				
Low	505	71.9%	108	21.4%
Moderate	122	17.4%	41	33.6%
Moderate/High	52	7.4%	19	36.5%
High	23	3.3%	14	60.9%
Subgroup Total	702	100.0%	182	25.9%

Finally, analysis was undertaken to determine if the PACT's classification power could be improved through an actuarial approach using data currently collected in PACT. The revised instruments increased the balance of the distribution of cases across the categories and increased the

level of separation attained in recidivism rates for cases classified as high, moderate, or low risk (these instruments are presented in Appendix B). Figure 1 outlines the overall results obtained with the revised risk assessment.

Figure 1



The use of the PACT for youth placed in correctional facilities was also examined. Outcomes were analyzed for a 12-month period following release. Although balanced case distribution across risk levels was found for the probation sample, 71% of all cases were classified as either moderate/high or high risk. The new adjudication rates for each of these groups were 45.6% and 49.1%, respectively. Hence, nearly three fourths of youth released from Florida facilities were classified

into two groups with a 3.5% difference in outcomes. These rates of adjudication were reflected in a low AUC (0.58) and a low DIFR (0.28). Table 17 presents these results, delineated by gender.

Table 17						
Florida PACT: New Adjudications by Risk Level Youth Released From Institutions						
Risk Level	All Cases		Boys		Girls	
	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated
Low	13%	29.0%	12%	30.8%	14%	19.9%
Moderate	16%	39.6%	16%	42.5%	18%	25.3%
Moderate/High	33%	45.6%	33%	48.1%	29%	29.8%
High	38%	49.1%	38%	52.0%	39%	33.1%
Base Rate	43.9%		46.6%		28.9%	
Sample Size	11,154		9,449		1,705	
AUC:						
Criminal History	0.58*		0.58*		0.57*	
Social History	0.52*		0.54*		0.52*	
DIFR	0.28		0.28		0.23	

*AUC significantly different than 0.50 (asymptotic significance ≤ 0.05 ; lower bound of confidence interval greater than 0.50).

When evaluated by race/ethnicity, some of the same issues encountered with the probation sample emerged. Moderate- and moderate/high-risk Hispanic/Latinos and moderate-risk Whites all had lower rates of subsequent adjudications than low-risk Black/African Americans. Moderate/high- and high-risk Whites and Hispanic/Latinos had lower rates of recidivism than moderate-risk Black/African Americans (Table 18).

Table 18 Florida PACT New Adjudications by Risk Level by Race/Ethnicity for Youth Released From Institutions						
Risk Level	Black/African Americans		Whites		Hispanic/Latinos	
	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated
Low	11%	34.5%	16%	24.1%	11%	29.5%
Moderate	15%	48.0%	19%	32.9%	16%	31.5%
Moderate/High	36%	51.0%	28%	40.0%	32%	33.3%
High	39%	53.7%	38%	44.7%	41%	42.9%
Base Rate	49.9%		37.9%		36.5%	
Sample Size	5,571		4,093		1,174	

Because PACT items were found to bear little statistical relationship to subsequent recidivism and did not perform equitably across different races and ethnic groups (particularly for boys), a revised risk assessment instrument could not be constructed. For additional information, see Appendix B.

a. Summary of Findings

While the current PACT instrument produced some separation of outcomes by risk level, it did not perform as well as several other instruments tested in this study. In some instances, analysis showed less than a 5% difference in recidivism across three risk levels. Equity problems, particularly for youth placed in facilities, were also in evidence. For probation cases, better results were obtained using simple actuarial scales developed using data collected by the PACT instrument.

4. Virginia YASI

Like the PACT, the YASI evolved from the WSJCA instrument designed in Washington in the 1990s. YASI was implemented in several states including New York, Illinois, and Mississippi in the 2000s and Virginia starting in 2008.

YASI consists of pre-screen and full-screen assessments. The full assessment is composed of 87 items across the following 10 domains: legal history, family, school, community and peers, alcohol and drugs, mental health, aggression, (pro-social and antisocial) attitudes, (social and cognitive) skills, and employment and free time. The YASI generates risk and protective scores in each of these areas.

The pre-screen consists of 32 risk items from the full screen. The pre-screen component is designed to assess a youth's risk level while obtaining a brief social and legal history. For each of its domains, the YASI provides a rating of static and dynamic risks and protective factors, which are designed to help predict recidivism as well as point to behavior patterns that ostensibly need to change in order to reduce future problems. Scores in these areas range from low to very high, using a six-point rating system. The final component of the YASI is a case supervision plan to be used by juvenile justice personnel that builds on problem areas identified in the assessment. The YASI is a product of Orbis Partners, Inc.

The YASI pre-screen is the risk assessment examined for this study. In 2007, Orbis Partners, Inc. conducted a validation study of the YASI in New York (Orbis Partners, 2007). Study results indicated an AUC value of 0.62 for 12-month and 24-month outcome measures.

Limitations were inherent to the evaluation of Virginia's YASI data for the current study. The data were collected during the early stages of YASI implementation in Virginia, and less than one third of cases admitted to probation had YASI scores available. It is possible, therefore, that some selection bias was introduced. However, administrators in Virginia report that the areas first selected for implementation were representative of the entire state. Further, it was difficult to align dates of assessments with probation admission dates because of the timing of implementation. To optimize

the sample size, all assessments conducted up to 90 days prior to the start of probation to 90 days after admission were included. Applying these parameters meant that events used to rate behaviors and possibly even new arrests and adjudications could have occurred prior to the assessment, conflating assessment results and outcomes. Similar issues were encountered in other jurisdictions (Solano County, for example) but follow-up periods could be adjusted to reflect 12 months from the date of assessment rather than admission. In Virginia, outcome rates were computed by department staff, so this adjustment was not possible.

To evaluate the extent to which these limitations affected the study, outcomes were compiled from the portion of the sample assessed prior to the start of probation with those of cases assessed after the admission date. Slightly less separation occurred in outcomes across risk measures for these two samples; however, the percentage of cases assigned to risk levels varied substantially. Comparisons are presented in Table 19.

Table 19		
Virginia YASI		
Percentage of Cases at Each Risk Level by Timing of Assessment		
Risk Level	Pre-Probation	Pre-Admission
Low	22%	45%
Moderate	48%	40%
High	30%	15%
Sample Size	908	1,011

This variance could reflect differences in characteristics of cases from counties that assess youth before and those that assess after probation admission, but the size of the difference suggests other factors may have been at least partially responsible.

Table 20 outlines overall results of the validation study delineated by gender.

Table 20 Virginia YASI New Adjudications by Risk Level						
Risk Level	All Cases		Boys		Girls	
	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated	Percent at Level	Percent Adjudicated
Low	34%	11.1%	27%	14.4%	53%	6.3%
Moderate	44%	27.3%	46%	28.2%	37%	24.2%
High	22%	41.7%	27%	44.5%	10%	21.2%
Base Rate	25.0%		28.9%		14.4%	
Sample	1,919		1,405		507	

For boys, the distribution demonstrated step-wise increases in recidivism rates as risk levels increased. For girls, considerable discrimination was found in recidivism rates recorded between low and moderate risk, but high-risk girls had a lower recidivism rate than moderate-risk girls. The cohort of high-risk girls, however, was small: only 10% of all girls assessed were rated high risk. These results could therefore be an artifact of the small number of high-risk girls in the study cohort. The AUC for all cases in the sample was 0.68. DIFR scores ranged from 0.57 to 0.74 for girls and boys respectively (see Appendix A for details).

High-risk girls also had a lower rate of recidivism than moderate-risk boys, indicating overlap by gender. The instrument developers had already modified cut points for girls in Virginia, but based on these data, this adjustment did not fully correct the issues discussed. Current cut points are presented in Table 21.

Table 21 Virginia YASI Pre-Screen Overall Risk Level Cut Points		
Risk Level	Girls	Boys
None	0	0
Low	1–25	1–15
Moderate	26–52	16–38
High	53+	39+

Missing data limited the examination of results by race/ethnicity. Sufficient data were present to disaggregate only for Whites and Black/African Americans. These data are presented in Table 22.

Table 22				
Virginia YASI New Adjudications by Risk Level				
Risk Level	Whites		Black/African Americans	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	39%	8.2%	26%	17.5%
Moderate	41%	23.9%	48%	31.9%
High	20%	34.6%	26%	50.8%
Base Rate	19.9%		33.0%	
Sample Size	1,150		701	

Moderate discrimination was found for both racial groups, although the recidivism rate for high-risk Whites was only 2.7% higher than that found for moderate-risk Black/African Americans. The difference in overall base rates for Whites and Black/African Americans (19.9% versus 33.0%) was more pronounced in Virginia than in most other jurisdictions in the study.

Using the YASI data, simple actuarial instruments were constructed for boys and girls. The boys' instrument resulted in much-improved separation of recidivism rates by risk level without substantially altering the distribution of cases across risk levels. Both the AUC and DIFR values improved as well to 0.71 and 0.80, respectively. Results for the total sample, for Black/African Americans, and for Whites are presented in Table 23.

Table 23						
Virginia Revised Boys' Risk Assessment Instrument						
Risk Level	All Boys (n=1,106)		Black/African Americans (n=451)		Whites (n=618)	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	28%	10.7%	23%	13.3%	31%	9.4%
Moderate	48%	30.3%	49%	35.6%	47%	25.8%
High	24%	51.1%	27%	56.5%	22%	45.9%
Base Rate		29.9%		36.1%		25.1%

The risk instrument developed for girls also produced substantially better separation in risk levels than the YASI pre-screen. Results are presented in Table 24.

Table 24						
Virginia Revised Girls' Risk Assessment Instrument						
Risk Level	All Girls (n=333)		Black/African Americans (n=124)		Whites (n=191)	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	36%	5.9%	34%	9.5%	36%	2.9%
Moderate	42%	16.3%	43%	17.0%	41%	15.2%
High	22%	38.4%	23%	51.7%	23%	30.2%
Base Rate		17.4%		22.6%		14.1%

The DIFR score for the revised assessment for the entire sample was 0.89; it was 0.90 for Black/African Americans and 1.12 for White youth. The AUC score for the revised girls' instrument was 0.74. These were some of the highest values attained in the current study.

a. *Summary of Findings*

Overall, the YASI produced substantial separation of re-adjudication rates by risk level. Furthermore, cases were well-distributed across risk levels. The instrument appears to work better for boys than girls: moderate-risk girls had higher rates of recidivism than high-risk girls. This result could be an anomaly attributable to the limited sample size.

Development of a 10-item risk instrument significantly improved the level of discrimination attained and produced a balanced distribution of cases across low, moderate, and high levels of risk.

5. Nebraska and Arkansas YLS/CMI

The YLS/CMI was developed in the 1990s by Robert D. Hoge and D. A. Andrews at Carleton University. It is a modified version of the LSI-R, a risk assessment instrument designed in the 1980s to evaluate adult offenders. The YLS/CMI is used by numerous juvenile justice departments in the United States. The YLS/CMI scores 42 risk items in eight major domains: prior and current offenses/dispositions, family circumstances/parenting, education/employment, peer relations, substance abuse, leisure/recreation, personality/behavior, and attitudes/orientation in order to obtain an overall risk level for the youth (low, moderate, high, or very high). Additionally, the instrument is used to indicate needs and special considerations, which may be taken into account to assist with case management.

The YLS/CMI instrument is available through Multi-Health Systems (MHS), an online service that distributes a variety of clinical, educational, and public safety-oriented assessments and tools. An online version of YLS/CMI is also available at Assessments.com.

Several studies have investigated the predictive validity and reliability of the YLS/CMI and have supported the instrument's ability to classify youth appropriately (Onifade, Davidson, Campbell, Turke, Malinowski, & Turner, 2008; Bechtel, Lowenkamp, & Latessa, 2007; Schmidt, Hoge, & Gomes,

2005; Flores et al., 2003). However, a 2004 evaluation of Nebraska's use of the YLS/CMI risk assessment in the juvenile justice system revealed concerns over the propensity of the instrument to classify too many youth as moderate risk (Kadleck, Herz, Gallagher, & Nava, 2004). Flores and colleagues (2003) concluded, "This research indicates that agencies planning to use the instrument only for initial risk assessment should consider a shorter and more economical assessment tool" (p. 47).

The YLS/CMI is used in three agencies that participated in this study: Nebraska Probation, Nebraska OJS, and Arkansas DYS. Nebraska implemented the YLS/CMI in 2002; Arkansas implemented in 2008.

Table 25 outlines the results of the validation conducted in each site. In Arkansas, both the number of cases available for analysis and the length of the follow-up period were limited; hence, these results should be viewed with considerable caution. Still, when combined with results from the two Nebraska agencies, data indicate that the YLS/CMI appears to have limited value as a classification tool, as it produced only minor separation in recidivism rates for cases at different risk levels and a lack of distribution of cases across risk categories.

More than 90% of probation cases in Nebraska were classified to two of the four possible risk levels; no case was rated very high risk and only 6% were classified to the high risk level. Recidivism rates ranged from 18% for low-risk cases to 25% for high-risk youth. This level of discrimination was well below that observed for most other instruments in the study.

Results were not better for youth placed in facilities. In both Arkansas and Nebraska, 95% of all cases were classified as moderate or high risk. No appreciable difference in recidivism rates occurred between these two classifications; in fact, moderate-risk cases had higher rates of recidivism than high-risk youth. In Nebraska, 3% of committed youth were rated low risk and 2% were rated very high risk. Despite the level of selectivity, the difference in recidivism rates between those classified as low risk and those classified as high risk was only 12.2% (Table 25).

Table 25						
YLS/CMI Results for Probation and Committed Youth in Nebraska and Arkansas						
Risk Level	Arkansas		Nebraska Probation		Nebraska OJS	
	Percent at Level	Return to a Facility	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	5%	0.0%	27%	17.9%	3%	10.0%
Moderate	76%	14.4%	67%	23.0%	32%	17.7%
High	19%	0.0%	6%	25.0%	63%	16.8%
Very High	0%	--	0%	--	2%	22.2%
Base Rate		10.9%		21.7%		16.9%
Sample Size		119		1,077		597

Changes to cut points did not improve the instrument's discrimination power. Several other modifications were tested, including selecting cut points that placed 25% (quartiles) of the sample at each risk level (i.e., the 25% of cases with the lowest scores were classified as low risk, the next 25% as moderate risk, etc.). These changes, too, did not improve the level of discrimination observed among risk levels (not shown). Nebraska administrators had, in fact, lowered the cutoff for high-risk offenders from 23 to 16, though as illustrated above, with unsatisfactory results.

These YLS/CMI results stem from low correlations between risk factors and outcomes in the three agencies using the instrument. The best results were obtained for probation cases in Nebraska. However, even for this population, no single item on the YLS/CMI had a correlation of 0.1 or above with recidivism. The highest correlated item was in the prior history domain, three or more prior convictions, which was correlated with recidivism at 0.08 (see Appendix A for site-specific results).

When classification results were delineated by race and ethnicity, the instrument worked well for White youth. The instrument did not perform well for Black/African Americans and Hispanic/Latinos. For Hispanic/Latinos, data in prior records were often unavailable, which may have influenced results for this subgroup.

As expected, given the results outlined above, both the AUCs and the DIFR scores computed for the YLS/CMI in each jurisdiction were very low. AUCs were generally below 0.6 and the highest DIFR score was 0.19 for White youth. Compared to results obtained on other risk assessment instruments in the study, the YLS/CMI provided poor discrimination of outcomes across risk levels and low AUCs and DIFR scores.

The Arkansas database was too small to support the development of an actuarial risk assessment, but actuarial instruments were constructed for both Nebraska Probation and Nebraska OJS. Results are presented in Table 26.

Table 26												
Nebraska Probation and Nebraska OJS Revised Risk Assessment Instruments Recidivism by Risk Level												
Risk Level	All Cases		Boys		Girls		Whites		Black/African Americans		Hispanic/Latinos	
	%	Recidivism Rate	%	Recidivism Rate	%	Recidivism Rate	%	Recidivism Rate	%	Recidivism Rate	%	Recidivism Rate
Nebraska Probation												
Low	26%	12.7%	25%	16.3%	27%	5.5%	28%	11.2%	20%	16.3%	25%	15.0%
Moderate	60%	21.8%	59%	25.2%	62%	15.0%	59%	19.4%	65%	29.3%	55%	19.3%
High	15%	37.2%	16%	36.4%	11%	39.5%	13%	40.0%	15%	40.6%	20%	25.8%
Sample Size	1,077		735		342		659		215		159	
Base Rate	21.7%		24.8%		15.2%		19.7%		28.4%		19.5%	
Nebraska OJS (Commitment)												
Low	16%	6.1%	15%	5.9%	22%	6.7%	16%	5.9%	16%	6.3%	17%	5.3%
Moderate	56%	14.1%	56%	13.9%	55%	14.7%	58%	15.0%	51%	13.5%	50%	15.5%
High	28%	29.1%	29%	27.6%	23%	35.5%	26%	32.1%	33%	27.3%	33%	21.1%
Sample Size	597		461		136		312		101		115	
Base Rate	16.9%		16.7%		17.6%		17.9%		16.8%		15.7%	

While these results represent a substantial improvement over the YLS/CMI, the analysis was restricted, in large part, to elements collected and categorized for the current risk assessment instrument. As a result, this instrument also works better for Whites than for other racial groups.

a. Summary of Findings

Results from three different populations included too little overall discrimination of outcomes by risk level, poor distribution of cases across risk levels, and serious equity issues. A simple actuarial risk instrument, developed using data collected by the current risk assessment instrument, produced much better results.

6. Arizona AOC Risk Assessment Instrument

The Arizona risk/needs assessment was developed for the Arizona Supreme Court AOC, Juvenile Justice Division. The first iteration of the assessment was constructed in the late 1980s in conjunction with the Juvenile Justice Classification Committee and Jim Riggs, PhD, from Research Information Specialists.

The original scale was composed of 10 variables based on their ability to identify the probability of subsequent juvenile criminal offenses and was designed to assess all youth referred to juvenile court. In the early to mid-1990s, AOC collaborated with the Arizona DJC and NCCD to examine factors related to juvenile recidivism and subsequently implemented a revalidated risk and needs assessment to classify every youth upon referral. In 1998, LeCroy & Milligan Associates revalidated and revised the assessment (LeCroy, Krysik, & Palumbo, 1998); and in 2007, another revalidation study found the instrument performing at moderate levels (AUC=0.652), though there were problems with the needs assessment (Schwalbe, 2009).

The risk assessment instrument in use today is the version revalidated in 1998 and again in 2007. It is a composite of three risk assessment “scoring streams” based on a youth’s prior offense

history. One stream is used for youth referred for their first offense, a second stream is used for youth referred for their second offense, and the third stream is used for youth referred three or more times.

Risk scores for each youth are calculated from a set of 10 risk factors, three of which are shared across the instruments. Additional risk items are scored depending on which risk stream is used. These items consist of questions related to type of offense, school-related information, behavioral problems, and peer relationships. Item weights reflect regression coefficients, and risk level cut points vary by scoring stream used. The risk assessment is completed for all youth referred to juvenile court.

In 2011, AOC implemented a new needs assessment, which is a derivative of the needs assessment in the Ohio Youth Assessment System (OYAS). The OYAS is a product of the University of Cincinnati. The Arizona AOC is the only jurisdiction in the United States that uses the risk assessment. The needs assessment was not examined in the current study.

The validity of each of the three scoring streams of the instrument was examined independently. Table 27 presents the combined results of all three scoring streams as well as results for each stream.

Table 27								
New Adjudications by Risk Levels								
Three Versions of the Arizona AOC Risk Assessment Instrument								
Risk Level	Combined Results		Version 1		Version 2		Version 3	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	21%	12.7%	72%	11.2%	22%	18.7%	0%	0.0%
Moderate	25%	22.4%	25%	18.4%	52%	19.8%	17%	27.5%
High	54%	29.0%	35%	22.6%	26%	24.5%	83%	29.6%
Sample Size	7,589		1,788		1,430		4,371	
Base Rate	23.9%		13.4%		20.8%		29.3%	

Over half of the probation population in Arizona was classified as high risk. This is in stark contrast to results in other states, where risk assessment systems tend to place the majority of probationers in the lower risk categories. Distribution issues in Arizona exist in part because the agency assesses all cases referred to juvenile court and diverts low-risk cases. Nonetheless, the low level of separation attained between moderate- and high-risk cases reflects minimal capacity to differentiate between cases at the highest risk levels.

Over half (57.6%) of the cases in the study entry cohort were classified using Version 3 of the system (Table 28). This version classified 83.1% of all cases to the high risk level, placed no case at the low risk level, and showed little separation of moderate- and high-risk cases. In contrast, Version 1 placed nearly 72% of all cases at low risk and only 3.5% at high risk (see Appendix B, page B19).

Table 28		
Arizona AOC Risk Assessment Instrument New Adjudication Rates by Version of Instrument Used		
Instrument	Percent of Cases	Percent New Adjudication
Version 1 (n=1,788)	23.5%	13.4%
Version 2 (n=1,430)	18.8%	20.8%
Version 3 (n=4,371)	57.6%	29.3%

More girls are classified by this risk assessment instrument as high risk than boys. However, high-risk girls recidivate at about the same rate as moderate-risk boys. Table 29 breaks down combined results by gender; racial/ethnic breakdowns are presented in Table 30.

Table 29 Arizona AOC Risk Assessment Instrument New Adjudications by Risk Level by Gender				
Risk Level	Boys		Girls	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	22%	13.4%	18%	9.7%
Moderate	25%	23.6%	26%	18.2%
High	53%	30.6%	56%	23.7%
Sample Size	5,922		1,667	
Base Rate	25.1%		19.8%	

Table 30 Arizona AOC Risk Assessment Instrument New Adjudications by Risk Level by Race/Ethnicity								
Risk Level	Whites		Black/African Americans		Hispanic/Latinos		Native Americans	
	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication	Percent at Level	Percent New Adjudication
Low	22%	12.1%	22%	14.1%	20%	13.0%	22%	12.9%
Moderate	25%	20.3%	24%	29.3%	26%	23.2%	25%	22.6%
High	53%	27.7%	54%	29.7%	54%	30.5%	54%	27.8%
Base Rate	22.4%		26.2%		25.0%		23.3%	
Sample Size	3,062		625		3,388		433	

A strength of the Arizona risk assessment instrument is that it places a similar population of cases, regardless of race or ethnicity, at each risk level. However, some “overlap” is evident: moderate-risk Black/African Americans had higher recidivism rates than high-risk Native Americans and Whites. In addition, the rates of subsequent adjudication were nearly identical for moderate-risk and high-risk Black/African American subgroups.

A revised instrument greatly reduced the proportion of cases classified as high risk, increased the degree of separation of outcomes between risk levels, and maintained a fair degree of equity

across racial/ethnic groups represented in the Arizona probation system.¹⁰ Results are presented in Table 31.

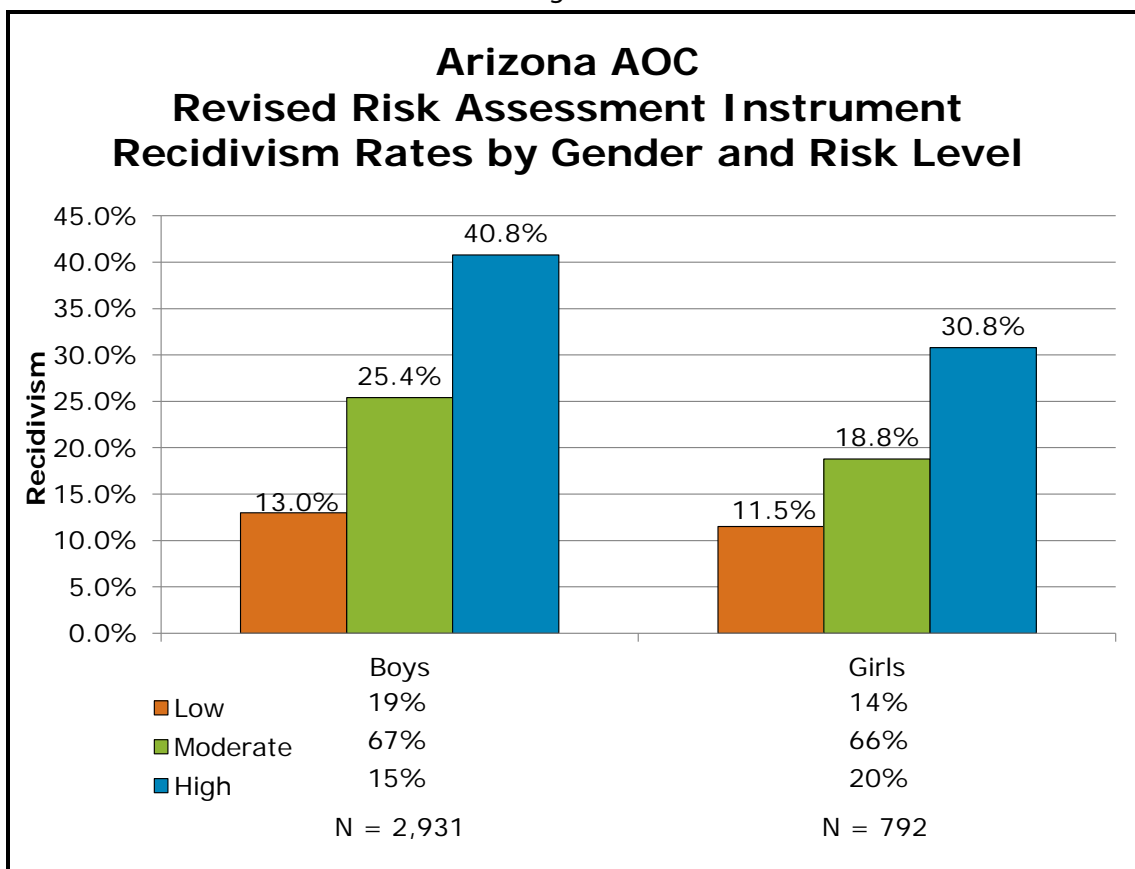
Table 31 Arizona AOC Revised Risk Assessment Instrument New Adjudications by Risk Level										
Risk Level	All Cases		Hispanic/Latinos		Whites		Black/African Americans		Native Americans	
	% at Level	% New Adjudication	% at Level	% New Adjudication	% at Level	% New Adjudication	% at Level	% New Adjudication	% at Level	% New Adjudication
Low	18%	12.8%	16%	12.5%	20%	13.1%	18%	10.3%	15%	16.1%
Moderate	67%	24.1%	68%	24.3%	66%	23.6%	63%	25.0%	67%	23.7%
High	16%	38.1%	16%	38.2%	15%	36.6%	19%	37.7%	17%	45.7%
Base Rate	24.3%		24.7%		23.5%		24.8%		26.4%	
Sample Size	3,723		1,678		1,484		323		2,014	

Note: Reflects validation sample.

The revised instrument also worked similarly for both boys and girls. In particular, this approach significantly reduced over-classification of both boys and girls to the high risk level. Results by gender are presented in Figure 2.

¹⁰ These analyses are presented only as an example of the degree to which risk assessment could be improved in the agency. In Arizona, additional analysis is recommended, especially the testing of additional potential risk factors that could further improve results. Ideally, more cases should be “pushed” into the low- and high-risk groups. Better distribution would significantly increase the potency of the classification system.

Figure 2



a. Summary of Findings

Though based on an actuarial design, the Arizona AOC risk instrument failed to provide substantial discrimination between risk levels. Equity issues were also found. Results may be attributable to two issues. First, the system uses regression coefficients as item weights, which complicates scoring; secondly, it appears that little, if anything, is gained from the three-tiered system of risk assessment. A single actuarial scale with the best combination of risk factors produced better separation and eliminated the equity problems found with the existing instrument.

7. Arizona DJC DRI

The DRI was developed in 2007 by the Arizona DJC in conjunction with LeCroy & Milligan Associates. Arizona DJC is the only jurisdiction in the nation that uses the DRI. The DRI consists of 18 items embedded in a broader assessment of youth functioning, the Criminogenic and Protective Factors Assessment (CAPFA). The CAPFA consists of more than 180 items in 12 domains and is conducted for all juveniles committed to the DJC. The DRI consists primarily of dynamic factors to assess a youth's likelihood of recidivism. According to the DJC, the dynamic components allow for the worker to track a juvenile offender's treatment progress over the duration of his/her system involvement, and these components provide a more comprehensive general picture of the youth. The scoring system employed for the DRI is based on item weights that reflect coefficients computed to the thousandth. For example, a five-point scale for "manipulation" is multiplied by -4.740.

The DRI was validated in 2008 by the Research and Development division of DJC. At the time of validation, the instrument was found to classify youth better than chance according to their likelihood of recidivating as evidenced by an AUC value of 0.64. However, in practice, very little distinction was shown in recidivism rates between medium and high risk classifications. This result was determined by the study's authors to be a product of the sample size and the small number of recidivists in the sample (Chengalath, 2008).

Analysis for the Arizona DJC site was limited by data availability. Data on new arrests, petitions, and adjudications were not available from the agency's information system. Tests of validity were thus limited to a single outcome measure: recommitment within 12 months of release.

The sample comprised a total of 1,265 youth released in 2007 or 2008. The recommitment rate for all cases in the sample was 37.9%. Overall results of the validation study are presented in Table 32, which also provides breakdowns by gender.

Table 32 Arizona DJC Dynamic Risk Instrument Recommitment Rates by Gender						
Risk Level	All Cases		Boys		Girls*	
	Percent at Level	Percent Recombitment	Percent at Level	Percent Recombitment	Percent at Level	Percent Recombitment
Low	55%	31.1%	52%	30.2%	70%	35.3%
Moderate	20%	45.0%	20%	45.4%	16%	41.7%
High	25%	47.3%	22%	46.7%	9%	61.5%
Base Rate	37.9%		37.9%		38.6%	
Sample Size	1,265		1,112		153	

*The girls' sample was too small to support breakdowns by race/ethnicity in subsequent tables.

As shown in Table 32, girls had a higher base rate of recidivism than boys. Second, despite higher recommitment rates, more girls were rated low risk. High-risk girls had a much higher recommitment rate than high-risk boys, but this may be an artifact of the small number of girls classified as high risk. Only 13 girls were rated high risk and eight of these were recommitted during the 12-month follow-up.

For boys, there was moderate discrimination in recommitment rates between low and moderate risk, but no significant difference in rates reported for moderate- and high-risk groups. Classification was skewed toward low risk for both genders; in total, 55% of the release cohort was classified low risk. The AUC for the total sample was 0.59; the DIFR score was 0.32, both relatively low values.

The DRI is primarily composed of items described in the literature as dynamic (factors that can change over time, or more specifically, factors that can improve as a result of services provided or maturation). However, many of the factors in the DRI have little statistical correlation with recommitment; in addition, several correlations were not in the expected direction. See Appendix B for details.

Table 33 breaks down results of the validation study by race and ethnicity. Within each population subgroup are two classification levels where the difference in recommitment rates between levels is less than 4%.

Table 33						
Arizona DJC Dynamic Risk Instrument Recommitment Rates by Risk Level						
Risk Level	Hispanic/Latinos		Whites		Black/African Americans	
	Percent at Level	Percent Recommitted	Percent at Level	Percent Recommitted	Percent at Level	Percent Recommitted
Low	54%	29.7%	57%	29.2%	56%	37.3%
Moderate	22%	44.9%	17%	43.9%	17%	40.9%
High	24%	47.7%	27%	46.2%	27%	55.6%
Base Rate	37.3%		36.2%		42.9%	
Sample Size	659		398		133	

Results of a revised risk assessment instrument based on available data are presented in Table 34. The revised risk assessment instrument can be found in Appendix B.

Table 34				
Arizona DJC Department of Juvenile Corrections Recommitment by Revised Risk Level				
Risk Level	N	%	Recommitment Rate	
			N	%
Low	343	27.1%	63	18.4%
Medium	638	50.4%	245	38.4%
High	284	22.5%	172	60.6%
TOTAL SAMPLE	1,265	100.0%	480	37.9%

The revised instrument produced a more balanced distribution of cases and a high level of discrimination on recommitment rates. Further, as Table 35 illustrates, the instrument worked equitably across all major population subgroups.

Table 35				
Arizona DJC				
Recommitment by Revised Risk Assessment Level and Youth Race/Ethnicity				
Risk Level	N	%	Recommitment Rate	
			N	%
TOTAL SAMPLE	1,265	100.0%	480	37.9%
Hispanic/Mexican National				
Low	186	28.2%	32	17.2%
Medium	347	52.7%	137	39.5%
High	126	19.1%	77	61.1%
Subgroup Total	659	100.0%	246	37.3%
Black/African American				
Low	29	21.8%	7	24.1%
Medium	64	48.1%	24	37.5%
High	40	30.1%	26	65.0%
Subgroup Total	133	100.0%	57	42.9%
White				
Low	109	27.4%	19	17.4%
Medium	193	48.5%	69	35.8%
High	96	24.1%	56	58.3%
Subgroup Total	398	100.0%	144	36.2%

Gender breakdowns are outlined in Table 36. Results indicated that the revised risk assessment instrument was effective in distributing cases across the risk continuum and separating cases into classifications with substantially different rates of recidivism.

Table 36				
Arizona DJC				
Recommitment by Revised Risk Level and Youth Gender				
Risk Level	N	%	Recommitment Rate	
			N	%
TOTAL SAMPLE	1,265	100.0%	480	37.9%
Girls				
Low	32	20.9%	5	15.6%
Medium	76	49.7%	29	38.2%
High	45	29.4%	25	55.6%
Subgroup Total	153	100.0%	59	38.6%
Boys				
Low	311	28.0%	58	18.6%
Medium	562	50.5%	216	38.4%
High	239	21.5%	147	61.5%
Subgroup Total	1,112	100.0%	421	37.9%

a. Summary of Findings

This risk assessment instrument relies heavily on dynamic factors scored using a complex formula and produces moderate levels of discrimination. Substantial improvements in both distribution and the level of discrimination were attained when the instrument was revised. The revised instrument worked equally well across race/ethnicities and gender. Use of a new risk assessment instrument might substantially improve risk classification in this agency.

8. Oregon JCP Assessment

The JCP risk assessment instrument was developed in the late 1990s by the Oregon Juvenile Department Directors Association. The instrument was used to target high-risk youth and link them to crime prevention services. Oregon is the only jurisdiction in the United States that uses the JCP. The assessment is now used in all 36 counties and nine federally recognized American Indian tribes in Oregon.

Since its inception, the JCP has routinely been evaluated for effectiveness by NPC Research, a social science research organization in Portland, Oregon (see, for example, Finigan, Mackin, Seljan, & Tarte, 2003; Tarte, Mackin, Cox, & Furrer, 2007). In the early 2000s, NPC conducted a risk validation study on the JCP assessment, which resulted in changes to the assessment that were implemented in 2006.

The JCP includes 30 risk factors organized into seven domains: school/academic issues, peers, behavioral issues (this domain captures information on school behavior, criminal history, runaway history, violence/aggressive behavior, and prior use of weapons), family dynamics, substance use, attitudes, and mental health. Risk factors are anchored with explicit definitions and scoring instructions.

The risk assessment instrument is embedded in an automated data collection system that also identifies factors that need to be addressed in case plans and collects data on “test items” (i.e., factors that could be used to improve risk assessments in the future depending on their statistical relationship to outcomes).

The base rates for both new referrals and new adjudications were exceptionally low in Oregon. Despite these limitations, the risk assessment instrument produced substantial discrimination in outcomes across risk levels. The analysis also found that the instrument worked equally well across different racial and ethnic groups. Overall results are presented in Table 37.

Table 37														
Oregon JCP New Adjudication by Risk Level														
Risk Level	All Cases		Boys		Girls		Whites		Black/African Americans		Hispanic/Latinos		Native Americans	
	Percent at Level	New Adjudication	Percent at Level	New Adjudication	Percent at Level	New Adjudication	Percent at Level	New Adjudication	Percent at Level	New Adjudication	Percent at Level	New Adjudication	Percent at Level	New Adjudication
Low	47%	4.9%	45%	5.9%	49%	2.7%	48%	4.6%	40%	5.4%	48%	5.8%	36%	5.1%
Moderate	38%	14.1%	39%	15.3%	36%	11.2%	37%	13.3%	40%	17.2%	37%	16.5%	43%	11.5%
High	16%	22.6%	16%	23.7%	16%	20.1	15%	22.0%	21%	28.5%	17%	22.4%	21%	29.0%
Base Rate			12.3%		8.5%		10.6%		14.9%		13.0%		12.9%	
Sample Size			8,678		3,692		8,305		658		2,440		326	

The AUC and DIFR values for the JCP were 0.70 and 0.71, respectively. The unusually low rate of recidivism observed in Oregon makes results from the JCP difficult to compare with other instruments evaluated in this study. Relative to the overall rate of recidivism observed in Oregon, the JCP achieved substantial separation in new adjudication rates. However, differences in absolute terms were only in the 8% to 10% range.

The low base rate, combined with the relative strength of the existing instrument, made improvements through revision difficult to attain. Efforts to do so resulted in slightly better discrimination, but these were offset by a more skewed distribution of cases across risk levels. Results of this analysis are presented in Appendix B.

a. Summary of Findings

Despite the low rate of subsequent adjudications reported during the 12-month follow-up period, the Oregon JCP produced a high degree of relative separation in recidivism rates recorded for low-, moderate-, and high-risk youth. The low base rate for re-adjudication observed in Oregon made attempts to improve on the current instrument difficult.

B. Comparison of Results Across Jurisdictions and Assessments

The following description of general findings is organized around four major areas of inquiry: reliability, validity (the level of discrimination attained, the distribution of cases across risk levels), equity, and cost.

1. Reliability

In nearly every site, the average percent agreement among workers was at least 75%, the minimum threshold for acceptability. Percent agreement, however, was 75% or higher between

workers and the expert in only five of the 10 study jurisdictions. Risk assessment instruments that exhibited the highest degree of agreement with expert scores were the Solano County JSC; the Georgia CRN (though the CRN levels may reflect an inflated percent agreement due to the way risk levels were calculated for the reliability test),¹¹ the Virginia YASI, and the Arizona AOC instrument. Instruments with lower reliability levels included the YLS/CMI, the PACT, the Oregon JCP, and the Arizona DRI (Table 38).

Table 38						
Inter-Rater Reliability Results Summary by Site						
Site and Assessment	Number of Raters	Percent Agreement		ICC		Kappa
		Among Workers	With Expert	Risk Level	Risk Score	Risk Level
Arizona AOC Risk Assessment Instrument	45	81.9%	79.0%	0.72	0.85	0.56*
Arizona DJC DRI	5	75.6%	55.6%	0.66	0.75	0.44
YLS/CMI						
Arkansas	15	75.2%	68.3%	0.54	0.67	0.33*
Nebraska Probation	26	79.2%	38.8%	0.62	0.80	0.42*
Nebraska Commitment	45	77.4%	73.4%	0.61	0.73	0.40*
Florida PACT	51	76.6%	68.4%	0.83	N/A	0.50*
Georgia CRN	50	92.0%	92.0%	0.88	0.93	0.80*
Oregon JCP	51	77.1%	62.1%	0.68	0.77	0.46*
Solano County						
Boys (JSC)	27	92.0%	92.0%	0.90	0.92	0.78*
Girls (Girls Link)	27	84.3%	83.3%	0.74	0.89	0.65*
Virginia YASI	69	84.7%	79.4%	0.77	0.89	0.61*

*Significant at $p < 0.05$

Note: ICC and kappa calculations include only cases in which workers completed all 10 case vignettes. PACT relies on a matrix of criminal and social history scores; therefore, an ICC could not be computed.

¹¹ The CRN risk levels consist of scores for age at first adjudication, number of prior adjudications, and a general delinquency score. Only general delinquency items were tested. To calculate risk levels for the study, age at first adjudication and number of prior adjudications were automatically scored based on study cases. Therefore, the likelihood of workers reaching the same risk level was enhanced.

Inter-rater reliability findings for individual items on the risk assessments varied by jurisdiction and type of item. Staff agreed at least 75% of the time for the majority of items; two exceptions were Arizona's DRI and Georgia's CRN. More than half of the DRI items (13 of 18) had inter-rater agreement lower than 75%, which is likely related to question content on the DRI that was not addressed or could not be answered based on the video vignette, or, because in practice, multiple specialists complete the DRI. More importantly, the number of raters able to participate was very low (n=5). Across the five raters testing the DRI, the items with lowest agreement were empathy, resistance to peer influence, and respect for authority.

Like the DRI, the majority (63%) of CRN items did not achieve the 75% agreement threshold, though results may be more related to item design than a lack of information in the case vignettes. More possible responses can lead to less consistency. Most items on the CRN allow up to five possible responses, which may have hindered inter-rater reliability. The items that did not reach the threshold included items about promiscuity, substance use, and youth and family functioning. In comparison, the YLS/CMI and the JCP limit item responses to yes or no; the lowest percent agreement obtained on those risk assessments was 65.3% and 56.5%, respectively (Table 39). For detailed inter-rater reliability results, see Appendix C.

Table 39				
Inter-Rater Reliability Summary				
Risk Items				
Risk Assessment Instrument	Number of Items	Minimum Percent Agreement Attained, Any Item	Maximum Percent Agreement Attained, Any Item	Proportion (#) of Items With < 75% Agreement
Arizona AOC Risk Assessment Instrument	13	55.8%	94.6%	23% (3)
Arizona DJC DRI	18	42.2%	100.0%	72% (13)
YLS/CMI				
Arkansas	42	67.6%	97.9%	24% (10)
Nebraska Probation	42	71.2%	98.1%	7% (3)

Table 39 Inter-Rater Reliability Summary Risk Items				
Risk Assessment Instrument	Number of Items	Minimum Percent Agreement Attained, Any Item	Maximum Percent Agreement Attained, Any Item	Proportion (#) of Items With < 75% Agreement
Nebraska Commitment	42	65.3%	96.2%	19% (8)
Florida PACT	44	52.8%	99.4%	4% (5)
Georgia CRN	56	32.1%	97.8%	63%(35)
Oregon JCP	30	56.5%	96.1%	30% (10)
Solano County				
Boys (JSC)	10	56.2%	99.4%	20% (2)
Girls (Girls Link)	8	83.3%	100.0%	-- (0)
Virginia YASI	87	46.3%	100.0%	10% (9)

Various factors related to item design might explain the low inter-rater reliability. More response options (i.e., categories) might have a negative impact on reliability, as may the absence of definitions and clear thresholds. For example, at least two risk assessment instruments (PACT and CRN) ask workers to select one of five responses. One CRN item about how often a youth goes out with friends or is alone after school is measured as never, <1 time per week, 1 to 3 times per week, 4 to 7 times per week, or unknown. Five items related to remorse or guilt include possible responses of definitely no, suspect no, unknown/no opinion, suspect yes, and definitely yes.

In general, items requiring varying degrees of subjectivity were found to be less reliable than clearly objective items. For example, the items with the lowest percent agreement across the assessments were related to disruptive behavior at school, positive friendships, harming or injuring animals, consequential thinking skills, and parental supervision. The most reliable items tended to be related to prior offense history and youth age at first contact with the juvenile justice system. However, these observations cannot tell us exactly why inter-rater reliability is low for a particular item. For additional information on inter-rater reliability results, see Appendix C.

2. Validity

The JSC and YASI instruments achieved the greatest separation in recidivism rates for cases assigned to high, moderate, and low risk levels. A moderate degree of separation between low and high risk was produced by both the PACT and the CRN, but analysis showed little difference in recidivism rates for moderate- and high-risk cases for each instrument. The YLS/CMI produced little separation in any of the three agencies that use this risk assessment instrument, although limitations in both the sample available and follow-up period in Arkansas limit the value of results from that state.

Measures of AUC were highest for the Oregon JCP and the Girls Link and JSC risk assessments that are used in Solano County. Table 40 compares results from all participating sites. This summary of results should be viewed in conjunction with data on case distribution presented in Table 41.

Table 40						
Validity Results by Risk Assessment Instrument						
Risk Assessment Instrument	Recidivism Within 12 Months				AUC	DIFR*
	Recidivism Rate	Recidivism by Risk Level				
		Low	Medium	High		
Arizona AOC Risk Assessment Instrument	24%	13%	22%	29%	0.62	0.40
Arizona DJC DRI	38%	31%	45%	47%	0.59	0.32
YLS/CMI						
Arkansas	11%	0%	14%	0%	0.40	Could not calculate
Nebraska Probation	22%	18%	23%	25%	0.55	0.15
Nebraska Commitment	17%	10%	18%	17%	0.54	0.12
Florida PACT						
Commitment	44%	29%	40%	47%	0.58/0.52**	0.28
Probation	36%	30%	44%	52%	0.59/0.63**	0.37
Georgia CRN						
Probation	29%	25%	52%	58%	0.64	0.40
Commitment	31%	24%	43%	46%		
Oregon JCP	11%	5%	14%	23%	0.70	0.71

Table 40						
Validity Results by Risk Assessment Instrument						
Risk Assessment Instrument	Recidivism Within 12 Months				AUC	DIFR*
	Recidivism Rate	Recidivism by Risk Level				
		Low	Medium	High		
Solano County						
Boys (JSC)	51%	19%	48%	64%	0.68	0.68
Girls (Girls Link)	35%	24%	29%	58%	0.68	0.34
YASI	25%	11%	27%	42%	0.68	0.68

Notes: Recidivism is measured by new adjudication except for Arizona DJC and Nebraska commitment populations (new commitment). DIFR is not applicable when outcome rate is 0% for one or more risk levels. Four agencies assign cases to four different risk levels. In two of those agencies, no one scored at the highest level. In the other two agencies, 5% or fewer were classified at the highest level. For this comparison, the two highest risk levels were combined.

*DIFRs reflect original classifications: four levels for YLS/CMI and PACT; three levels for the other risk assessment instruments.

**Criminal history/social history

Table 41			
Current Risk Assessment Distribution by Site			
Risk Assessment Instrument	Risk Level	Distribution	
		N	%
Arizona AOC Risk Assessment Instrument	Low	1,596	21%
	Medium	1,930	25%
	High	4,063	54%
	Total	7,589	100%
Arizona DJC DRI	Low	695	55%
	Medium	251	20%
	High	319	25%
	Total	1,265	100%
Nebraska Probation YLS/CMI	Low	291	27%
	Moderate	718	67%
	High	68	6%
	Very High	0	0%
	Total	1,077	100%

Table 41			
Current Risk Assessment Distribution by Site			
Risk Assessment Instrument	Risk Level	Distribution	
		N	%
Nebraska Commitment YLS/CMI	Low	20	3%
	Moderate	192	32%
	High	376	63%
	Very High	9	2%
	Total	597	100%
Florida PACT Probation	Low	18,350	67%
	Moderate	4,839	18%
	Moderate-High	2,741	10%
	High	1,439	5%
	Total	27,369	100%
Florida PACT Commitment	Low	1,410	13%
	Moderate	1,830	16%
	Moderate-High	3,636	33%
	High	4,278	38%
	Total	11,154	100%
Georgia CRN	Low	5,692	77%
	Medium	1,395	19%
	High	325	4%
	Total	7,412	100%
Oregon JCP	Low	5,774	47%
	Medium	4,678	38%
	High	1,918	16%
	Total	12,370	100%
Solano County Boys (JSC)	Low	128	15%
	Medium	376	43%
	High	376	43%
	Total	880	100%

Table 41			
Current Risk Assessment Distribution by Site			
Risk Assessment Instrument	Risk Level	Distribution	
		N	%
Solano County Girls Link	Low	17	13%
	Moderate	86	66%
	High	28	21%
	Total	131	100%
Arkansas DYS YLS/CMI	Low	6	5%
	Moderate	90	76%
	High	23	19%
	Very High	0	0%
	Total	119	100%
Virginia YASI Girls Sample	Low	651	34%
	Medium	841	44%
	High	427	22%
	Total	1,919	100%

Distribution (or dispersion) was problematic for several of the instruments evaluated. Some of these distribution patterns can be traced to policy and practice, such as diversion of low-risk cases (Solano County and Arizona AOC). But the YLS/CMI, for example, placed few cases at the very high risk level, regardless of where it was implemented. The Georgia CRN, using current cut points, placed 88% of probationers in the low risk classification and less than 1% at the high risk level, a low level of discrimination.

As noted earlier, base rates for each site are critical for understanding results. In Oregon, nearly half (47%) of the assessed population was rated low risk. These cases had an exceptionally low rate of recidivism (4.9%); hence, the low risk rating can be considered accurate. Rates of subsequent adjudication for moderate- and high-risk offenders were also well under those reported in other jurisdictions. It is possible that the Oregon risk assessment instrument is less likely to adjudicate

(differences in new arrest rates were not pronounced when Oregon was compared to other jurisdictions) and/or that Oregon has effective programs and practices in place that produce higher rates of success for youth on probation.

The high rate of recidivism in Solano County probation results from the fact that most low-risk youth are screened out of probation; only low-risk youth with more serious offenses or low-risk cases that were overridden to higher levels by the officer doing the assessment are admitted to probation. As a result, only 15% of the Solano County cohort was assessed as low risk.

Despite the differences noted in Oregon and Solano County, the risk assessment instruments employed in each of these jurisdictions effectively separated cases to different risk levels, relative to the base rates observed in each jurisdiction. Because data on screened-out cases in Solano County were not available, it was not possible to include these cases in the analysis. It should be noted that this had substantial impact on the proportion of cases at each risk level and probably resulted in artificially low DIFR scores for the Solano County risk assessment instrument.

3. Equity

Equity issues were found with several of the risk assessment instruments evaluated in this study. The YLS/CMI, in particular, exhibited problems: It produced better separation for White youth than for Hispanic/Latinos or Black/African Americans. Problems with equity were also found for the PACT risk assessment instrument used in Florida, particularly for youth who were placed in state facilities. Several instruments, most notably the YASI, did not classify girls appropriately.¹²

Overall equity results are presented in Tables 42 and 43 and are limited to AUC and DIFR. Relative to the other instruments in the study, the AUCs and DIFRs for various race/ethnicities were

¹² Based on YASI system documentation provided to NCCD for this study, YASI developers attempted to resolve this issue by altering cut points to classify girls. However, as results illustrate, the system remained ineffective at classifying girls by the likelihood of recidivating.

strongest for the Oregon JCP, the JSC instrument for boys in Solano County, and the YASI. When examined by gender, the risk assessment instruments that performed well relative to the other instruments included the Oregon JCP, the YASI, and the JSC instrument for boys. Due to the complexity of presenting results from 10 agencies, only summary statistics are presented here.

Table 42							
Validity Results By Race/Ethnicity							
Risk Assessment Instrument	Recidivism Rate*	AUC			DIFR		
		White	Black	Hispanic/ Latino	White	Black	Hispanic/ Latino
Arizona AOC Risk Assessment Instrument	24%	0.62**	0.60*	0.62**	0.41	0.39	0.42
Arizona DJC DRI	38%	0.60**	0.58	0.60*	0.35	0.32	0.36
YLS/CMI							
Arkansas	11%	0.27	0.49	N/A	Could not calculate	Could not calculate	N/A
Nebraska Probation	22%	0.58**	0.48	0.51	0.18	0.02	0.09
Nebraska Commitment	17%	0.57	0.54	0.47	0.17	N/A	0.57
Florida PACT							
Probation Criminal History/Social History	36%	0.59**/0.63**	0.59**/0.63**	0.59**/0.63**	0.38	0.36	0.35
Commitment Criminal History/Social History	44%	0.58**/0.55**	0.56**/0.53**	0.56**/0.54**	0.33	0.23	0.23
Georgia CRN	31%	0.64**	0.63**	0.69**	0.35	Could not calculate	Could not calculate
Oregon JCP	11%	0.70**	0.71**	0.67**	0.73	0.81	0.67
Solano County							
Boys JSC	51%	0.70**	0.65**	0.67**	0.70	0.73	0.56
Girls Link	35%	0.56**	0.70**	0.70**	0.25	Could not calculate	0.64
Virginia YASI	25%	0.68**	0.66**	N/A	0.74	0.57	N/A

*New adjudication except in Arizona DJC and Arkansas (new commitment).

**AUC significantly different from 0.50.

Notes: Could not calculate some cells because recidivism did not increase with each successive risk level and/or recidivism value equaled zero. N/A means the cohort size was too small and/or data were unavailable.

Table 43					
Validity Results by Risk Assessment Instrument By Gender					
Risk Assessment Instrument	Recidivism Rate*	AUC		DIFR	
		Boys	Girls	Boys	Girls
Arizona AOC Risk Assessment Instrument	24%	0.62**	0.60**	0.41	0.40
Arizona DJC DRI	38%	0.60**	0.56	0.34	0.30
YLS/CMI					
Arkansas	11%	0.40	0.45	Could not calculate	Could not calculate
Nebraska Probation	22%	0.52	0.61**	0.11	0.22
Nebraska Commitment	17%	0.51	0.63**	0.23	Could not calculate
Florida PACT					
Probation Criminal History/Social History	36%	0.60**/0.62**	0.58**/0.65**	0.37	0.39
Commitment Criminal History/Social History	44%	0.58**/0.54**	0.57**/0.52	0.28	0.23
Georgia CRN	31%	0.64**	0.61**	Could not calculate	0.31
Oregon JCP	11%	0.69**	0.74**	0.65	0.95
Solano County	47%	0.68**	0.68**	0.68	0.34
Virginia YASI	25%	0.67**	0.71**	0.58	Could not calculate

*New adjudication except in Arizona DJC and Arkansas (new commitment).

**AUC significantly different from 0.50.

Note: Could not calculate some cells because outcome rate did not increase with risk level increase, or outcome rate for a group equaled zero.

4. Revised Risk Assessment Instruments Constructed in the Study

Simple, actuarial risk assessment instruments were created (or modified, for those jurisdictions already using actuarial models) for every study jurisdiction except Arkansas, using data from the existing instrument (the study cohort in Arkansas was too small to support the analyses needed to construct a new risk instrument). As noted earlier, if the cohort of cases available for analyses

exceeded 2,000 it was divided into construction and validation samples, and results from validation samples were used. While it was possible to construct actuarial instruments for probation cases in Florida, it was not possible to construct an instrument for committed youth in Florida that worked across all racial/ethnic groups because of data limitations and substantial differences in recidivism by race.

In most instances, the new instruments constructed for each agency produced markedly better results than the instrument currently in use. The two exceptions were the JSC boys' instrument used in Solano County and the JCP assessment used in Oregon. In the case of the JCP, the low base rate observed in Oregon may, in part, account for the fact that improvement was difficult. However, testing in a population with a higher rate of recidivism would provide useful information that may enable further improvement.

Table 44 presents a comparison of results from the existing instrument in each jurisdiction with those attained with a simple actuarial design. Factors available for development of an actuarial instrument were generally limited to those collected in the existing instrument. The instruments developed therefore do not necessarily provide optimal classification results, but they do demonstrate the potential for substantial improvement.

Table 44					
Comparison of Current and Revised Risk Assessment Instruments by Site					
Risk Assessment Instrument	Risk Level	Current Risk Assessment Instrument		Revised Risk Assessment Instrument (Validation Sample When Available)	
		Level %	% Re-Adjudicated*	Level %	% Re-Adjudicated*
Arizona AOC Risk Assessment Instrument	Low	21%	12.7%	18%	12.8%
	Medium	25%	22.4%	67%	24.1%
	High	54%	29.0%	16%	38.1%
	Overall	100% (n=7,589)	23.9%	100% (n=3,723)	24.3%

Table 44					
Comparison of Current and Revised Risk Assessment Instruments by Site					
Risk Assessment Instrument	Risk Level	Current Risk Assessment Instrument		Revised Risk Assessment Instrument (Validation Sample When Available)	
		Level %	% Re-Adjudicated*	Level %	% Re-Adjudicated*
Arizona DJC DRI	Low	55%	31.1%	27%	18.4%
	Medium	20%	45.0%	50%	38.4%
	High	25%	47.3%	22%	60.6%
	Overall	100% (n=1,265)	37.9%	100% (n=1,265)	37.9%
Nebraska Probation YLS/CMI	Low	27%	17.9%	26%	12.7%
	Moderate	67%	23.0%	60%	21.8%
	High	6%	25.0%	15%	37.2%
	Very High	0%	-		
	Overall	100% (n=1,077)	21.7%	100% (n=1,077)	21.7%
Nebraska Commitment YLS/CMI	Low	3%	10.0%	16%	6.1%
	Moderate	32%	17.7%	56%	14.1%
	High	63%	16.8%	28%	29.1%
	Very High	2%	22.2%		
	Overall	100% (n=597)	16.9%	100% (n=597)	16.9%
Florida PACT Probation Boys' Sample	Low	66%	31.1%	22%	19.8%
	Moderate	18%	45.2%	42%	34.2%
	Moderate-High	11%	49.8%		
	High	6%	57.4%	36%	50.7%
	Overall	100% (n=20,621)	37.0%	100% (n=10,370)	36.9%
Florida PACT Probation Girls' Sample	Low	70%	31.1%	24%	18.3%
	Moderate	17%	41.2%	55%	30.5%
	Moderate-High	9%	44.9%		
	High	4%	58.1%	21%	51.3%
	Overall	100% (n=6,748)	32.3%	100% (n=3,397)	32.0%
Georgia CRN Boys' Sample	Low	74%	28.5%	32%	17.0%
	Medium	21%	49.3%	44%	37.1%

Table 44					
Comparison of Current and Revised Risk Assessment Instruments by Site					
Risk Assessment Instrument	Risk Level	Current Risk Assessment Instrument		Revised Risk Assessment Instrument (Validation Sample When Available)	
		Level %	% Re-Adjudicated*	Level %	% Re-Adjudicated*
	High	5%	48.4%	24%	49.1%
	Overall	100% (n=5,407)	33.9%	100% (n=2,506)	33.4%
Georgia CRN Girls' Sample	Low	85%	19.3%	23	11.7%
	Medium	13%	36.1%	54	21.0%
	High	2%	36.8%	23	33.9%
	Overall	100% (n=2,005)	21.8%	100% (n=2,005)	21.8%
Solano County Girls' Sample (Girls Link)	Low	16%	23.8%	23%	13.6%
	Moderate	59%	29.0%	49%	28.3%
	High	25%	57.8%	29%	64.0%
	Overall	100% (n=261)	35.2%	100% (n=261)	35.2%
Virginia YASI Boys' Sample	Low	27%	14.4%	28%	10.7%
	Medium	46%	28.2%	48%	30.3%
	High	27%	44.5%	24%	51.1%
	Overall	100% (n=1,412)	28.8%	100% (n=1,106)	29.9%
Virginia YASI Girls' Sample	Low	53%	6.3%	36%	5.9%
	Medium	37%	24.2%	42%	16.3%
	High	10%	21.2%	22%	38.4%
	Overall	100% (n=507)	14.4%	100% (n=333)*	17.4%

Note: Outcomes reported for Arizona DJC and Arkansas are "returns to a facility." The instruments represented in this table are those where revised instruments were able to substantially improve classification results. Gender-specific instruments are included when results attained outperformed the current risk tool.

*Only records with full YASI were included in revised risk assessment.

While YASI results were further improved by using only those factors with the strongest relationship with recidivism, substantial improvement over results of the current version of the Oregon JCP or the JSC boys' risk assessment from Solano County were not attained. The evaluation of the Oregon JCP was limited to some extent by the low rate of subsequent adjudications reported.

5. Efficiency and Cost

Juvenile justice agencies consider the issues of efficiency and cost when selecting a risk assessment instrument. Over the last two decades, the objectives and, consequently, the number of issues covered in risk assessment have increased substantially. Some assessment procedures can require two or more hours to complete.

Risk/needs assessments included in this study took from approximately 30 to 90 minutes to complete. Time to complete each risk assessment is illustrated in Table 45.¹³

Table 45	
Time to Complete by Risk Assessment Instrument	
Risk Assessment Instrument	Minutes
YLS/CMI (Arkansas)	56
AZ AOC	29
DRI	83
PACT	53
CRN	54
YLS/CMI (Nebraska OJS [commitment])	82
YLS/CMI (Nebraska Probation)	67
JCP	35
JSC/Girls Link*	54
YASI	97

*The time estimates for JSC and Girls Link assessments include data collection to establish appropriate supervision strategies and require less than 20 minutes to complete.

On average, workers in the 10 sites in the study spent about 61 minutes to complete an initial assessment. Of the instruments evaluated, the YASI takes the longest to complete. Estimates from Virginia indicated that the assessment took, on average, one hour and 37 minutes to complete. Estimates for the YLS/CMI (in Nebraska) and Arizona's DRI were both over 80 minutes. The risk instrument developed for the Arizona AOC required the least amount of time. The JSC and Girls Link

¹³ Estimates are from a survey of staff who participated in the reliability study.

instruments required, on average, 54 minutes to complete, but in Solano County they are embedded in a more comprehensive system that provides supervision strategies in addition to risk and needs assessment.

Estimating costs for various systems is complicated because (1) cost estimates depend in part on the scope of responsibility of the agencies in the study; (2) different funding formulas are used by vendors of commercially available risk assessment instruments; and (3) costs incurred by agencies that developed local instruments are often indistinct from routine personnel expenditures.

One of the largest study sites, Florida, has invested more than \$1.2 million over the past seven years to implement and sustain PACT, not including internal personnel time and training expenses for more than 800 staff who routinely complete the PACT. Initial costs included \$1 million, plus ongoing annual fees of more than \$30,000 to license and maintain online access to the instrument. Since 2001, Georgia has invested \$300,000 in start-up fees, plus ongoing fees ranging from \$50,000 to \$200,000 in subsequent years.

YASI expenditures in Virginia were substantially lower, though the pricing structure used is different. Virginia paid \$50,000 plus \$100 per user to implement the YASI. Virginia employs about 500 staff who use the YASI, which means an additional \$50,000 in user costs. The department incurs an additional \$25,000 per year for ongoing maintenance fees.

Arkansas and Nebraska incurred lower costs to purchase the YLS/CMI (from no fee to about \$2,500) and are charged from \$1.50 to \$2.85 per assessment on an ongoing basis.

Oregon and the two Arizona sites developed risk assessment instruments locally. Estimated start-up costs in Oregon were about \$100,000. DJC and AOC costs could not be separated from agency personnel expenditures, though AOC estimated that automating the risk assessment cost about \$80,000, and the cost to monitor the system runs about \$70,000 per year.

Solano County uses instruments that are available at no cost in the public domain. The county, however, uses these risk assessments as part of a web-based risk and needs model that provides

supervision strategies and data for developing case plans. Solano County spent \$7,000 per year on a subscription to this web-based system.

The cost of training staff to use the risk/needs systems can be substantial, though equally as difficult to isolate. Training costs associated with the initial implementation ranged from about \$30,000 in Georgia to \$76,000 to cover staff time and materials in Arizona. Oregon allocates \$20,000 per year for training; Nebraska pays \$125 per YLS/CMI class to cover materials each month, plus \$500 per day for trainers. Arkansas sets aside \$7,000 per YLS/CMI training session, and Georgia allocates \$10,000 per year for ongoing training (\$75 per hour). The cost of ongoing training in Florida is embedded in personnel costs, like several other agencies in the study (Arizona AOC and DJC, Nebraska, and Solano County). Florida includes the training as one of staff members' routine job responsibilities. Oregon is the only participant in which individual counties are responsible for training costs; given the scope of this study, we were unable to gather cost estimates from each of the counties that uses the JCP. Virginia is charged \$200 for each person trained in YASI. See Table 46.

Table 46					
Cost Estimates					
Site	Costs to Implement Risk Assessment Tool	Maintenance Costs	Internal Costs	Training Costs	Trainer Costs
Arizona AOC Risk Assessment	Risk: Developed in-house Needs: No cost for tool (public domain) Automation: \$80,000	Risk: None Needs: \$18,000/year for maintenance,** no other fees	Risk: None Needs: \$70,000/year for four part-time staff to monitor system	Risk: None Needs: \$76,000 (\$24,000 for initial "Train the Trainers;" \$52,000 for travel and materials)***	Risk and needs: Ongoing training provided by staff; included in job responsibilities
Arizona DJC DRI	No cost, developed by staff	None	None	No specific training curriculum	Ongoing training provided by staff; included in job responsibilities
Arkansas YLS/CMI	No implementation costs	\$2.85/per assessment; use 350–400/year	Not known	\$7,000 for two-day training (held as needed)	Trainer costs included in \$7,000

Table 46					
Cost Estimates					
Site	Costs to Implement Risk Assessment Tool	Maintenance Costs	Internal Costs	Training Costs	Trainer Costs
Florida PACT	\$1,000,000	\$34,500 annual fee for license and maintenance	Built into job responsibilities	25 probation officers trained as trainers; two-day training	Ongoing training provided by staff; included in job responsibilities
Georgia CRN	First year: \$300,000 for automation and programming	Second and third years: \$200,000 for validation and tracking	2001: \$200,000 2002/2003: \$150,000 2004/2005: \$50,000; 2006: \$100,000;* 2007–2012: \$50,000	2001: \$31,200 for four-person training team; 2002–2012: \$10,000/year	\$75/hour
Nebraska (OJS and Probation) YLS/CMI	\$450 initial software purchase; \$2,000 licensing agreement	No annual fee; \$1.50/assessment	None	\$125/class (held every four weeks) for materials and travel	Ongoing training provided by staff; included in job responsibilities (staff are certified trainers); training for assessment tool is provided in new worker training; \$500/day for contract trainers
Oregon JCP	\$100,000	Automation costs unknown	\$150,000 contract to vendor for evaluation work	Training paid for by each county; specific costs unknown	\$20,000 allocated for travel and consultant to coordinate trainers
Solano County Risk Assessments	No cost for risk assessment; public domain (costs for web-based system)	\$7,000/year	Unknown	N/A – no new hires since initial training	Ongoing training provided by staff; included in job responsibilities
Virginia YASI	\$50,000 to customize software; \$100/user for initial software purchase	\$25,000/year to vendor	\$90,000	Two-day training; material costs: \$125 for Part 1, \$35 for Part 2	\$200/person

Note: All costs are approximate and based on interviews with site administrators.

*Increased costs due to tool revisions.

**Related to training and implementation of new needs assessment, 2011.

Comparisons suggest that cost savings might be realized over time by developing a risk assessment instrument locally or validating an imported instrument on the local juvenile population.

Overall costs also include costs specific to local revalidations, which may occur every three to five years. An informal survey of several jurisdictions suggested that periodic revalidations can cost between \$45,000 and \$75,000.

IV. DISCUSSION

The use of risk assessment has become commonplace in the field of juvenile justice. Jurisdictions are using risk levels to guide placement decisions, assigning high-risk youth to specialized caseloads that have intensive services and more frequent contact with probation officers. These methods are effective with high-risk youth, but can increase recidivism for low-risk youth (Lowenkamp & Latessa, 2004). For this reason, it is imperative that instruments accurately differentiate between youth and accurately assign them to high, moderate, and low risk levels.

The analyses outlined in this report demonstrate that some risk instruments work well; others provide some level of discrimination between high-, moderate-, and low-risk youth but could be improved; and the validity of others is not at the level required to support decision making.

The following discussion focuses on two questions: (1) What separates highly successful risk models from those that do not provide the same degree of discrimination; and (2) Could attributes found in successful models guide juvenile justice agencies in selecting risk assessment instruments? The discussion is based on results of this study and more than 40 years of experience in the juvenile justice field.

We focus first on risk models developed for general use across jurisdictions, followed by a comparison of instruments developed for a specific jurisdiction. The comparison is presented to help agencies that undertake original research avoid problems encountered in other development efforts. The final section discusses issues that, over time, have contributed to less-than-optimal risk classification in many agencies throughout the country.

A. Instruments Developed for General Use

1. Overall Results

Instruments developed for general use evaluated in this study included the YLS/CMI, PACT, YASI, the CRN (COMPAS Youth), and the JSC and Girls Link instruments. Of these, the JSC and Girls Link instruments are the only public-domain instruments; the others were developed and distributed by private organizations.

The JSC, used in Solano County, California, proved to be the most successful risk instrument evaluated in this study. This assessment produced the highest absolute level of discrimination attained between high-, moderate-, and low-risk youth as well as high AUC and DIFR scores. It was also the most reliable in identifying risk levels and worked very well across all major ethnic groups in Solano County (Table 47).¹⁴

Table 47						
Summary of Validity and Reliability Results for Risk Assessment Instruments Developed for Use Across Jurisdictions						
Risk Instrument	Re-Adjudication Rate			AUC	DIFR	Reliability Percent Agreement on Risk Level
	Low	Moderate	High			
JSC (Solano County)	18.8%	47.5%	64.4%	0.68	0.68	92%
YASI	11.1%	27.3%	41.7%	0.68	0.68	85%
PACT*	30.0%	44.4%	51.8%	0.58/0.52	0.28	77%
CRN (COMPAS Youth)	25.7%	46.9%	47.1%	0.64	0.40	92%
YLS/CMI**	17.9%	23.0%	25.0%	0.55	0.15	79%

Note: Four levels of risk are used in Nebraska and Florida. In Nebraska, no case scored at the highest risk level. In Florida, only 5.3% of all probationers scored at the highest risk level. To obtain three risk levels for comparison purposes, the moderate/high and high risk categories were combined. The combined group accounted for 15.3% of Florida probationers.

*The PACT utilizes two scales. The first AUC value reflects legal history scores; the second AUC is for the social history score.

**To make results as comparable as possible, cases represented in these analyses are probation cases only. The YLS/CMI results are from the analysis conducted on probation cases in Nebraska. The Arkansas sample was too small to produce stable estimates of validity and, if included, would have further diminished the relationship between YLS/CMI scores and outcomes.

¹⁴ The CRN produced the same level of overall reliability (percent agreement) as the JSC, but this was largely due to the fact that raters were not required to rate the two factors that drive the classification process. These two factors—age at first adjudication and prior adjudications—are auto-scored and were not tested in the study. The reliability of the remaining 50 or so items used in the scoring algorithm varied from 21% to 95%.

2. Development Methods

Methods of development varied greatly, ranging from selecting risk factors based on theory to a general assessment of prior research results that identified selected factors proven valid in development studies in a variety of agencies. YASI and PACT fall between these two approaches: Both are based on research conducted in the State of Washington, but each evolved somewhat differently over time.

The risk instrument used in Solano County was developed for the Graduated Sanctions Center at the National Council of Juvenile and Family Court Judges (Wiebush, 2002). The risk scale is a compilation of results obtained in 14 different jurisdictions representing every region of the country.¹⁵ All of these jurisdictions used similar instruments; each was constructed via original research on cases in each jurisdiction. Items included on the JSC were those that (1) appeared on all or nearly all instruments or (2) were found on the majority of instruments and exhibited particularly strong relationships to recidivism. In total, the JSC contains 10 risk factors.¹⁶ The weights assigned to each item were also based on results obtained from the jurisdictions that were reviewed. The result is a simple risk instrument contained on a single page that provides raters with solid anchors for designating scores for each factor and represents a combination of offense history and social history factors. This approach to the design of an instrument for general use ensures that factors included on the instrument have a high degree of “universality” and have been tested on highly diverse populations from a wide variety of agencies.

The PACT model, in contrast, was originally developed and validated in the State of Washington and subsequently implemented in Florida, Texas, and other jurisdictions around the

¹⁵ Jurisdictions included Arizona; Cuyahoga County, Ohio; District of Columbia; Maryland; Michigan; Missouri; New Mexico; North Carolina; Oklahoma; Rhode Island; Travis County, Texas; and Virginia.

¹⁶ The 10 items are age at first referral, number of referrals, number of referrals for violent offenses, number of out-of-home placements, school discipline/attendance issues, substance abuse, peer relationships, prior abuse/neglect, parental supervision, and parent/sibling criminality.

country. The risk instrument in PACT contains approximately twice as many factors as the JSC, some of which are associated with violence rather than general recidivism. While it has proved possible to assess both the risk of violence and general recidivism with a single instrument, a number of issues with design and scale construction should be considered. First, the base rate for violence is relatively low in most jurisdictions, particularly among probationers (see, for example, Johnson, Wagner, & Matthews, 2002; NCCD, 2004). Second, some factors related to violence have little relationship (or may even be inversely related) to general recidivism (Johnson et al., 2002). Hence, including more than a few factors related to violence will often have a detrimental impact on the instrument's ability to accurately classify youth based on general recidivism rates.

In addition, basing item selection on results from a single jurisdiction may not be the most effective development strategy for a general-use instrument, given the variation in juvenile justice practice across the United States. Evidence for this can be found in widely disparate rates of detention, placement, and adjudication. When a large number of factors are included on a risk instrument, potential exists for supplanting factors universally related to recidivism with factors that reflect practices specific to the jurisdiction where the scale was developed. Such instruments may not transfer well to other jurisdictions where different legislation, different policies, and different practices are in place.

In addition, states adopting risk assessment models developed elsewhere should carefully assess the data sources used to develop these models. We have found that while instruments developed for probation cases may transfer reasonably well to populations of youth placed in state facilities (often with substantial changes in cut points and, in some instances, changes to weights assigned to individual risk factors), it is less likely that models developed for aftercare populations will work well for probation (Wagner, Ehrlich, & Baird, 1997). Youth on aftercare often have more extensive delinquent histories and are, in many instances, far more likely to have been adjudicated for assaultive offenses. In fact, we have found that separate instruments are needed for probationers and

committed youth in some jurisdictions in order to maximize the effectiveness of risk classification (Wagner et al., 1997). PACT, however, is applied to both populations. As a result, the model contains many factors with minimal relationships to recidivism for probationers in Florida (Table 48).

Table 48	
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.

Many of the above factors (e.g., escapes, prior weapon offenses, prior offenses against persons, commitment orders) will be observed more frequently for populations of committed youth than for probationers. While it can be argued that these factors so seldom apply to probationers that they have little impact on risk scores, issues of face validity and efficiency remain—both of which can undermine the overall effectiveness of the risk instrument.

The possibility also exists (although an issue less frequently encountered) that risk factors identified for probationers will not work as well for committed populations. For example, while several social history factors exhibited relatively strong relationships to recidivism for probation cases in Florida, not a single social history factor was correlated with recidivism at the 0.1 level or above for the committed population. It appears that the combined effect of all these issues limits the capacity of the PACT risk instrument to optimally classify cases in Florida.

Both PACT and YASI were derived from research conducted in the state of Washington, raising the question of why the results from Virginia, a YASI site, were so much better than those observed in

Florida. The answer perhaps lies in the timing of this study. Virginia was the only participating jurisdiction where the implementation of the risk model in use was not completed prior to data collection. At the time of the study, YASI results were available for less than 20% of all cases. Selected counties were using YASI, but most jurisdictions in Virginia were still in the planning and implementation stages of the project. Training on YASI in the study counties was thus a relatively recent event, so the results obtained may reflect a “halo,” or Hawthorne, effect that frequently occurs when change is introduced. Certainly the results observed in Virginia surpass those found in New York (Orbis Partners, 2007). It is possible that Virginia results will, over time, more closely approximate those found in Florida and New York. It will be important for Virginia to continue to monitor results as more counties begin to use the YASI.

3. Other Design Issues

Other design issues may also impact reliability and validity of risk models developed for general use. The YLS/CMI, for example, purportedly based item selection on theory and prior risk assessment research (Andrews & Bonta, 2003). While the YLS “domains” seem appropriate, some of the actual factors used do not. For example, under the “Leisure/Recreation” domain, youth receive a score for “could make better use of time” (Figure 3). There are several problems with this item: (1) it seems doubtful that juvenile justice agencies, prior to development of the YLS/CMI, collected data that would demonstrate that such a factor had any relationship to recidivism; (2) it is a subjective item that is difficult to score reliably; and yet (3) it is given the same weight as several prior delinquency factors. The relationship between prior criminal history and recidivism is well-established; the relationship between “could make better use of leisure time” and recidivism is not. Selecting appropriate domains is only a first step in scale construction; the actual factors used to obtain scores for each domain are critical to scale validity.

Figure 3

YLS/CMI Risk Item	
6. Leisure/Recreation:	
a. Limited organized activities	<input type="checkbox"/>
b. Could make better use of time	<input type="checkbox"/>
c. No personal interests	<input type="checkbox"/>

Other instruments reviewed in this study also contain factors with questionable rationale for inclusion. The Georgia CRN contains items rating how frequently a youth attends movies or “hangs out” at a shopping mall. While these factors seem to be aimed at rating a youth’s use of leisure time, it is again doubtful that much research support exists for such items. We found no justification for their inclusion in the extensive literature review undertaken for this study, and these factors exhibited minimal relationships to outcomes in Georgia. All of this suggests that much greater care is required in the selection of factors used to rate domains.

In addition, it appears that reliability (and therefore validity) is also harmed by simple inefficiencies in the design of some instruments. For example, minimal definitions are provided on the actual YLS risk form to guide workers in scoring youth on each domain. It is our experience that

reliance on prior training and/or instructions provided in a manual are insufficient given staff turnover rates and the day-to-day pressures encountered in supervising delinquent youth. Including clear instructions and definitions of each potential answer on the form can significantly enhance reliability.

The following domain presents a classic example of problems that arise from a lack of readily available definitions and/or instruction (Figure 4).

Figure 4

YLS/CMI Substance Abuse Item	
5. Substance Abuse:	
a. Occasional drug use	<input type="checkbox"/>
b. Chronic drug use	<input type="checkbox"/>
c. Chronic alcohol use	<input type="checkbox"/>
d. Substance abuse interferes with life	<input type="checkbox"/>
e. Substance use linked to offense(s)	<input type="checkbox"/>

The first two options on the YLS/CMI substance abuse item appear to be mutually exclusive given the definitions of “occasional” and “chronic”; that is, to check both would be counterintuitive. However, workers are in fact instructed to also check “occasional abuse” when “chronic abuse” is checked. (In effect, chronic abuse receives a score of 2, while occasional abuse receives a score of 1.) In

automated versions of the YLS/CMI, this scoring rule is automatically enforced. But when the system is not automated, errors can and do occur. For Nebraska commitment cases, workers neglected to comply with this rule in 12.3% of the YLS/CMI instruments completed. In Arkansas the error rate was 50%. A minor change in the design of the risk model would rectify this problem. The items could be treated as mutually exclusive choices if item weights were added to the assessment form (i.e., 2 for chronic abuse, 1 for occasional abuse).

Researchers, practitioners, and purveyors of the YLS/CMI cite the need for quality training as well as the need to focus on implementation issues to ensure system fidelity. While both are important, it is also critical to design risk instruments to avoid potential problems like those noted above. The reality is that training resources are limited, and both training and implementation requirements are integrally linked to the design of the system. Simple, well-designed approaches present fewer implementation issues than their more complex counterparts.

4. Would Simpler Models Transfer Better Among Agencies?

Development of simple actuarial instruments for each participating agency demonstrates that equal or better classification results can be obtained by reducing the number of factors considered and by using only variables that, in combination, create the greatest level of discrimination between high-, moderate- and low-risk groups. In three sites (Florida, Georgia, and Arizona AOC), we used construction and validation samples to test the validity of the newly created risk instruments. (Using validation samples allows a better estimate of how an instrument will perform over time.)¹⁷ These instruments, however, were developed using cases from each site. These analyses do not address the question of what type of risk model would best transfer to other jurisdictions.

¹⁷ Optimal results are usually observed for the sample of cases used to construct the risk instrument. Testing the results on a separate cohort of cases provides a better test of scale validity. The decline in results observed between construction and validation samples is commonly known as “shrinkage.” In this study, when large construction and validation samples were available, the amount of shrinkage observed was minimal (for example, see Florida and Arizona AOC results).

Two sites, Florida and Georgia, were selected to test the idea that simple actuarial instruments might transfer better than more complex instruments. Large cohorts of probationers were available in both of these sites and databases in both states contained sufficient information to produce a close approximation of JSC scores. The Florida database allowed for the closest approximation of the JSC assessment (in other words, the Florida database contained the same or similar items as those used to score the JSC).¹⁸

JSC scores were computed for all males on probation and compared to PACT results for the same population. Because PACT identifies four levels of risk, the current study used the cut-off points suggested by Wiebush (2002), which also identify four risk levels. Results are presented in Table 49.

Table 49				
Comparison of PACT and JSC Classification Results for Male Probationers in Florida				
Risk Level	PACT		JSC	
	Percent at Level	Re-Adjudication Rate	Percent at Level	Re-Adjudication Rate
Low	66%	31.9%	13%	20.7%
Moderate	18%	45.2%	56%	33.9%
Moderate/High	11%	49.8%	26%	48.1%
High	6%	57.4%	5%	58.4%

As shown above, the JSC accurately divided the very large PACT low-risk group into low and moderate risk categories. Both the low- and moderate-risk JSC groups had lower rates of recidivism than PACT low- and moderate-risk cases. Further, the JSC risk assessment identified a moderate/high-risk group with 2.5 times the number of cases assigned to this level by PACT, yet these cases recidivated at nearly the same rate as the PACT cases. Both instruments identified about the same

¹⁸ The following PACT items approximated items on the JSC: age at first offense, misdemeanor referrals, felony referrals, misdemeanor against person, felony against person, weapons referrals, history of court-ordered or voluntary placement, school enrollment, school conduct, school attendance, academic achievement, alcohol and drug use, current friends/companions, history of violent/physical abuse, history of neglect, parental control and authority, and history of household member incarceration.

number of high-risk youth, with very similar rates of re-adjudication. The overall difference in recidivism rates moving from low to high risk was 25.5% for the PACT and 37.7% for the JSC risk assessment.

These improvements were produced using a risk assessment instrument with half the number of risk factors contained on the PACT instrument. These results also produced higher AUC and DIFR scores (Table 50). As the findings demonstrate, in Florida, the JSC risk assessment produced slightly higher AUC and DIFR scores for boys on probation than PACT.

Table 50		
Florida PACT* and JSC Comparison		
Risk Assessment	AUC	DIFR
Florida PACT	0.60/0.62**	0.37
JSC	0.63	0.44

*Male probationers.

**The PACT includes two risk scores; the AUC for the criminal history score is 0.60 and the AUC for the social history score is 0.62. Note that there is only one overall PACT risk level, and therefore only one DIFR score.

In Georgia, the simulation of JSC scores was nearly as robust as that produced in Florida. JSC scores were computed for both boys and girls on probation. Because Georgia classifies cases to three levels of risk, new cut points were selected for the JSC. CRN cut points used in this comparison were the revised thresholds recommended in this report, rather than those used during the study period. Results are presented in Table 51.

Table 51				
Comparison of CRN and JSC Classification Results for Youth Probationers* in Georgia				
Risk Level	CRN		JSC	
	Percent at Level	Re-Adjudication Rate	Percent at Level	Re-Adjudication Rate
Low	62%	22.5%	45%	18.4%
Moderate	19%	36.3%	40%	37.2%
High	20%	47.5%	15%	46.2%

The JSC risk assessment instrument provided better separation of low- and moderate-risk group cases even after CRN cut points were modified. More cases were classified as moderate risk without substantially altering the rate of recidivism observed for moderate-risk cases. The CRN, after adjustments to cut points, was more efficient in identifying high-risk youth. Overall, the range in recidivism rates, from low to high risk, was 27.8% for the JSC and 25% for the CRN. AUC and DIFR scores were similar for both assessments (Table 52). The CRN score produced an AUC of 0.642 and the DIFR was 0.47; the AUC and DIFR scores for the JSC were slightly higher at 0.658 and 0.48.

Although the differences between the CRN and JSC were not as pronounced as those found between the PACT and the JSC, the JSC performed as well or better than the much more complex CRN.

Table 52		
Georgia CRN and JSC Comparison		
Risk Assessment	AUC	DIFR
CRN (revised cut scores)	0.64	0.47
JSC	0.66	0.48

In sum, the JSC risk assessment worked well in both Florida and Georgia. These findings, combined with the results obtained from efforts to develop simple actuarial risk instruments for each participating site, provide strong evidence that restricting the goal of risk assessment to optimal identification of high-, moderate-, and low-risk groups—and using only those factors that optimize the separation of these groups—improves classification. Other important issues with scale development are discussed below.

5. Are Complex Scoring Algorithms or Classification Methods Needed or Beneficial?

In the introduction, we discussed how risk models have changed over time as more objectives have been added to the assessment process. In addition, risk instruments are further complicated by unnecessarily complex algorithms for computing scores. Although such complexities may appear to add a level of sophistication to risk assessment, the current study findings suggest they do not improve classification results.

For decades, researchers have compared results obtained from the simplest of development methods (Burgess scoring based on bivariate relationships) to those obtained using the most sophisticated statistical techniques available, and found little difference in the validity of scales produced by advanced statistical methods (Gottfredson & Gottfredson, 1980). Prior research has demonstrated, and the current study confirms, that simple additive scales produce valid results, are easy to use, and are easily understood by staff and key decision makers.

The CRN used in Georgia employs a complicated scoring system where scores from dozens of domains are combined to create a social motivation score, a family vulnerability score, and a normative deviance score. These scores are added together to form a general-delinquency score, which is converted to a Z-value, then a T-value. Cut points are applied to T-values and added to age-level values and adjudication-level values to calculate a risk score. Finally, cut points are applied to the risk score, resulting in an overall risk classification level.

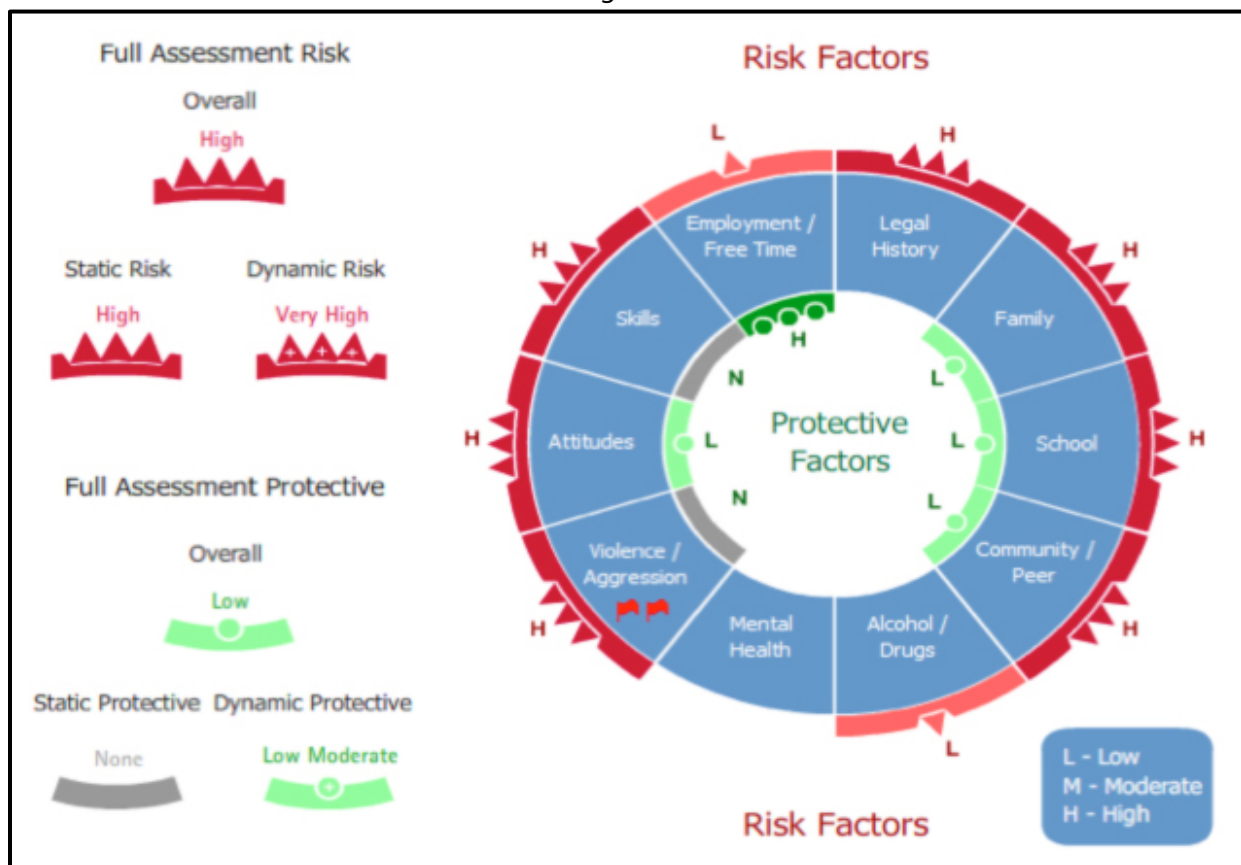
This type of scoring mechanism masks the fact that risk scores are driven in large part by two simple factors: age at first adjudication and prior delinquency. Similar issues have been raised by other researchers who evaluated the adult COMPAS model; one group asked why so many factors were needed, demonstrating that equal or better results could be obtained with a few risk factors (Zhang, Roberts, & Farabee, 2011).

Even for systems developed for specific agencies, this study confirmed that complex scoring systems tended to diminish results over time. Using precise weighting systems, such as regression coefficients, can tie results too tightly to a construction sample, and instruments with these scoring systems may not prove as robust over time.

Finally, the expansion of the objectives of risk assessment over the last two decades has led to an increase in “outputs” produced by various risk assessment models. The value of some of the outputs provided by these systems is questionable. The full YASI model, for example, produces ratings for static risk; dynamic risk; overall risk; static protective factors; dynamic protective factors; overall protective factors; and high, moderate, or low risk ratings for nine separate domains, all contained on the “YASI wheel” (Figure 5). Gottfredson and Moriarty (2006) questioned the concept and value of separating static and dynamic risk factors and we share their concerns. Breaking down risk into so many categories seems to stretch the concepts of risk and needs assessment to a level that is difficult to comprehend. Even though the YASI was relatively new to Virginia at the time the current study was conducted, nearly 40% of all staff surveyed did not think YASI provided assistance with case planning.¹⁹ Complexity also may add to the amount of training required to ensure that staff understand and use the risk assessment model correctly.

¹⁹ This information was collected prior to Virginia staff’s effort to implement assessment-based case planning in probation and parole beginning in 2012.

Figure 5



B. Risk Instruments Developed for a Specific Agency

Of the risk assessment instruments evaluated in this report, three were developed for and used by a single agency. These agencies were the Arizona AOC, the Arizona DJC, and Oregon. Our belief entering this study was that systems developed specifically for use in a single jurisdiction would outperform risk assessment models developed for general use. However, only the Oregon system performed at the level expected. Although construction of a revised instrument indicated that modest improvement was possible in Oregon, it was not at a level sufficient to recommend any changes to the current Oregon JCP model.

As in the discussion of general-use instruments, we focus on differences in the models in an attempt to provide the juvenile justice community with guidance on scale construction and

usefulness of the instruments in practice, identifying attributes that distinguish more successful risk models from those that provide lower levels of discrimination.

Both Arizona systems were developed using actuarial methods. Both models displayed a modest level of discriminatory power when tested in this study (the AOC risk assessment producing somewhat better results than the DJC instrument), but efforts to revise these instruments indicated both could be improved. When results from this study were shared with these agencies, they were reviewed and acted upon immediately. Changes have already been introduced. The ability to respond quickly to new information and make changes is one advantage of systems developed for and maintained by a specific justice agency. A summary of results from all three “homegrown” assessments is presented in Table 53.

Table 53						
Summary of Validity and Reliability Results for Risk Assessment Instruments Used in Arizona and Oregon						
Risk Model	Re-Adjudication Rate			AUC	DIFR	Reliability Percent Agreement on Risk Level
	Low	Moderate	High			
Oregon JCP	17.5%	34.0%	47.7%	0.67	0.56	77%
Arizona AOC	20.7%	32.8%	44.1%	0.63	0.44	82%
Arizona DJC	31.1%	45.0%	47.3%	0.60	0.32	76%

Note: Recidivism rates reported in this table are for new referrals/complaints for Oregon and Arizona AOC and recommitments (the only outcome available) for Arizona DJC. The rate of re-adjudication observed in Oregon was much lower than rates observed in other study sites. Hence, data on new referrals/complaints were used to present a useful comparison of results for probationers in Arizona and Oregon.

As noted earlier, unneeded complexities reduced the effectiveness of both Arizona risk instruments. These are discussed in detail in this report. That discussion centers on the use of regression coefficients that likely tie results too closely to the construction sample. This level of precision can prove detrimental: more robust instruments commonly round weights to generally reflect statistical relationships observed. This mitigates, to a degree, the impact of minor changes in

law, policy, practice, and offender population that occur over time. In effect, it adds to the “generalizability” of the risk instrument.

In addition, the DRI used by the Arizona DJC focuses primarily on changeable (or dynamic) risk factors. This may well stem from recent claims in risk assessment literature that dynamic factors are better predictors of recidivism than static factors (Gendreau, Little, & Goggin, 1996). However, the research conducted here and by NCCD in other jurisdictions finds that delinquent history factors are some of the strongest available predictors of recidivism.²⁰ Including more delinquent history factors on the DRI would, in all probability, lead to better results.

The Oregon JCP is embedded in a larger assessment system, but the risk assessment instrument is relatively brief (it could be represented on one or two pages, definitions included). The model contains 30 items organized into six domains. Risk factors are well-defined; scoring is simple and well-documented so workers are more likely to understand precisely how risk is assessed. Although we found that slightly better results were attainable with fewer factors, the improvement was not sufficient to recommend any changes to the model. A few counties in Oregon have added a “very high” risk category and assign cases to four risk levels. However, we found use of a very high risk level counterproductive: Cases at that level did not have higher rates of recidivism than high-risk cases.

Nearly all factors used in Oregon JCP can be found in other actuarial models. Other items, such as needs, protective factors, etc., are clearly labeled, and issues that should be addressed in case plans are clearly identified. The simplicity; use of well-developed definitions; and clear, unambiguous instructions are all strengths of the system.

Risk assessment instruments constructed and validated in the agencies where they are used should provide the best results. However, this study demonstrates that other factors matter as well.

²⁰ In the last 20 years, NCCD has conducted juvenile justice risk assessment validations in several jurisdictions including New Mexico; Missouri; Arizona; Indiana; Nebraska; Travis County, Texas; Maryland; North Carolina; and Virginia.

Steps taken in Oregon to simplify scoring, to provide clear definitions and instructions, and to separate risk factors from other items that have a role in case planning and supervision but not risk assessment serve as a guide to other jurisdictions undertaking the construction of a comprehensive approach to assessment.

C. Comments From Advisory Board Members and Authors' Responses

1. Best-Practice Implications of the Study Findings: Comments on the Validity, Reliability, and Equity of Commonly Used Juvenile Risk Instruments, by James Howell, PhD, and Aron Shlonsky, PhD

The successful operation of juvenile justice systems is dependent on valid offender risk and treatment assessments and evidence-based service matching. Unfortunately, the state of the art for these vital juvenile justice system operations is at a crossroads, having been complicated by misguided efforts. First, many states are now struggling with inappropriate instruments for assessing risk of recidivism. Second, tools proffered in many states for assessing treatment needs are ill-suited for this purpose. Third, the process of combining risk and treatment needs in the same instrument has complicated matters. Fourth, instruments designed for use on adult clients are sometimes applied to juveniles. As a result of these developments, many state juvenile justice systems are overwhelmed with unmanageable risk-need instruments that are misused, require enormous amounts of staff time and expense to complete, and often produce a large mound of data that program administrators and supervisors do not have the staff capacity to analyze nor put to good use in everyday practice. Risk assessment science and forward-looking offender management systems promise more objective, equitable, and effective juvenile justice systems.

We suspect that some of the well-intentioned instrument purveyors may not be intimately familiar with juvenile justice system operations—in particular, the dual purpose of juvenile justice systems: public protection and rehabilitation. These are statutory mandates in every state. To meet

these mandates, there is enormous utility in grouping offenders into distinctive risk levels to protect the public with accompanying levels of supervision and, if necessary, loss of freedom to commit crimes. Juvenile courts and correctional agencies then match offenders' treatment needs to services that reduce recidivism. Related to this point, some instrument purveyors seem unaware that risk and need assessments may be performed on numerous occasions on the same offenders as they move through the system. Hence, tools are needed that systematically move juvenile offenders across a continuum of services and sanctions, governed by a disposition matrix within which a continuum of services is available. This set of tools has worked very effectively in North Carolina. Admissions to secure correctional facilities were reduced by two thirds within a decade.

For more than 30 years in American juvenile justice history, actuarial instruments have been effectively used to assess risk of recidivism. Wiebush and colleagues' validations of 14 actuarial instruments provided the basis for the model JSC risk instrument adopted by the National Council of Juvenile and Family Court Judges. From the beginning, separate risk and needs instruments were developed for juvenile justice systems, mainly because of the distinctly different functions associated with them. In addition, combining risk and needs assessments into a single instrument can cause confusion because risk should be reassessed only when recidivism occurs, whereas needs must be reassessed regularly to chart treatment progress (at least every 30 days). In other words, risk assessment and needs assessment serve uniquely different purposes.

Moreover, assessment for treatment needs should be a two-step process. In the first step, a general or global assessment—often called a pre-screen (a shortened version of the full assessment instrument)—is made after collecting information that is readily available from agency records and a short, structured interview with the offender. In the second step, specific needs assessment instruments investigate a particular aspect of the youth more deeply. It is important that risk assessment instruments and needs assessment instruments are in sync with the developmental stages of offender careers. First, they must cover each of the developmental domains (family, school, peers,

individual problems). Second, these instruments must be capable of prioritizing treatment needs in each of these developmental domains and as these change with age and criminal involvement. Used in tandem, risk assessment instruments help determine placements and levels of supervision, and needs assessment instruments facilitate matching services to treatment needs at each level of advancement in criminal careers and juvenile justice system involvement. These conditions require regular reassessments of risk and needs; hence, the entire process must be parsimonious and possess high practical utility for smooth and effective operations.

Some of the lengthy instruments currently available have led to confusion with respect to their uses. Some of these tools measure psychological constructs to estimate recidivism likelihood. This is inappropriate because an offense is an actual event, not a construct. Recidivism is not a thought pattern; it is an overt behavior, an event that is observed by parents, other authorities, and victims and recorded in official records. Auto insurance agencies do not use psychological instruments to estimate accident risk; rather, they use accident reports to create age-graded insurance rate charts. An inside joke among neuroscientists is that car rental companies surely have in their employ the best neuroscientists because they refuse to allow a person to rent a car under age 25. But the real reason is clear to any actuary: Each year, about 4,000 teenagers are killed in motor vehicle accidents and as many as 300,000 are injured.

This same procedure is followed in the design of actuarial risk assessment instruments in juvenile justice systems—looking backward at offender characteristics that strongly correlate with (predict) recidivism. Early and persistent delinquency involvement is the best predictor of future delinquency, thus actuarial risk instruments must prominently rely on *static factors* (e.g., age of first arrest or conviction, number of previous arrests, convictions, or incarcerations, runaway episodes etc.) and also *dynamic factors* (current offender circumstances) that can strengthen predictions.

Another source of confusion is the assumption that immutable treatment needs exist—for example, thinking distortions that can be rectified with cognitive behavioral therapy. Unfortunately,

the treatment enterprise is not that simple for juvenile offenders with multiple problem behaviors. The sources of these problems typically span the major developmental domains: family, school, peers, individual problems, and environmental conditions. Hence, multiple services are required that address a full array of problems that may change with time.

European scholars and practitioners led by Van Domburgh have drawn attention to several important best-practice issues in the risk/needs assessment enterprise in Europe that parallel experiences in the United States.

- Lengthy instruments are time-consuming for staff and may place an unnecessary burden on parents and youth.
- Assessment is seldom based on multi-phase or longitudinal screening techniques.
- Attempts are sometimes inappropriately made to adapt instruments used for older age groups for use with children.
- There often is a lack of cooperation and sharing of results across agencies.
- In some cases, screening and assessment results differ, creating confusion.
- Duplication of assessments creates confusion for the parents and children, particularly in not knowing what can be expected from the various agencies.

In addition, lengthy risk instruments are not well-suited for everyday practice in juvenile justice systems with large volumes of cases—10,000 or more annually in many states. Scales containing as many as 25 variables and 100 items introduce significant distortions, create potential problems with reliability, and impose enormous administrative costs. Certainly, the use of dynamic factors to assess risk can be done and should be further explored. However, the exchange of dynamic factors for more predictive static factors in a risk framework is ill-advised. While some analysts note that this approach has promise, the results seem to indicate that the current slate of tools have not lived up to this promise and either more work needs to be done or an entirely different approach should be pursued. We support the latter position. It is unrealistic to assume that “criminogenic

factors” can be specified with sufficient precision at the individual level to create an aggregate tool that accurately measures a client's progress. The main reason these tools are so lengthy is that, in this enterprise, they attempt to cover such a wide range of behaviors and attitudes that none is covered sufficiently. Our view is that the treatment enterprise should focus less on risk and more on services; that is, quickly obtain a measure of risk, use this information to set priorities where necessary, and focus on behavior change that is measurable and specific to the individual. If there is a specific behavior problem, define that problem and employ tools (or create them) that can measure its frequency and severity; find an effective service or program that specifically addresses that particular problem; then treat and monitor using validated measures specific to that problem where possible. Many youths will have similar sets of problems depending on location and history, and good administrative data will indicate what these are so that access to effective services can be ensured. The use of dynamic assessments, while admirable in theory, simply tries to capture too much for everyone and does not focus adequately on the individual’s presenting problem.

- Keep it simple. Short instruments—either a stand-alone risk assessment instrument or a pre-screen of the better instruments—are most easily implemented and have good inter-rater reliability.
- In no case should jurisdictions adopt a lengthy risk assessment instrument that does not contain a pre-screen group of static and dynamic factors.
- Rely on selected factors that best separate offenders at least into in high-, medium-, and low-risk groups to facilitate program placement and service matching.
- Actuarial instruments work best because they are developed or validated for the population to which they apply.
- Include static and dynamic factors.
- When new offenses occur, re-administer to assess current risk.
- Provide extensive staff training.
- Revalidate instruments periodically (every few years).

This research report reveals the shortcomings of risk assessment instruments that do not incorporate a preponderance of static factors along with dynamic ones. Analysts can easily identify a parsimonious set of factors that increase the validity of unwieldy and unreliable instruments, as demonstrated in this report. Juvenile justice and allied fields are enormously indebted to Chris Baird and his colleagues for this courageous and highly scientific study that reveals important limitations of several risk instruments that are widely promoted today. This research report is based on science at its best: several instruments tested in multiple sites simultaneously, using common study methods and analysis procedures. Hence the findings from this report provide an urgently needed foundation for taking stock of risk assessment instruments in play and moving forward only with those that are actuarial, that is, based on risk of future offending and are capable of grouping offenders according to risk levels. In sum, this is a landmark study that promises to advance the state of the art in supporting juvenile justice system operations with valid, reliable, and practical risk management tools.

2. Youth Risk Assessment Approaches: Lessons Learned and Questions Raised by Baird et al.'s Study (2013), by Jennifer Skeem and (in alphabetical order) Robert Barnoski, Edward Latessa, David Robinson, and Claus Tjaden

a. Overview

i. Context and Purpose

In juvenile justice agencies across the United States, it has become common to apply structured tools to assess a youth's risk of re-offending and/or to inform efforts to reduce that risk. For good or for ill, an industry has grown up around "risk/needs" assessment, and states increasingly are developing their own "risk assessments." Many risk assessment tools are now available. Although most tools stem from the same root, they vary in their degree of complexity, structure, and independent research support. These tools, in turn, are being implemented in agencies that differ in their levels of

organizational commitment to both the value(s) of risk assessment and the necessity of ensuring that staff have adequate training, skills, and motivation to score the tools correctly.

Given this diversity of tools and implementation efforts, the time is ripe for a snapshot of the reliability and utility of risk assessment in juvenile justice agencies. That snapshot has just been provided for several agencies in the form of a study by Baird et al. (2013).

We are delighted that Baird et al. (2013) conducted this study. We believe that their data provide a valuable picture that can be used to advance “real-world” risk assessment. We are concerned, however, that their presentation of these data will promote mistaken conclusions. The field should not abandon an entire, relatively new approach to risk assessment because some tools have some problems in some jurisdictions—that would amount to throwing the baby out with the bathwater.

Before beginning, it is important to note who we are. This comment was written by four of the five advisory board members who participated in the final meetings held in Baltimore, where Baird et al.’s (2013) report was discussed at length, along with an additional member who could not attend those meetings. Like Baird (who helped create the Solano County instruments, the JSC and Girls Link), three of us have a conflict of interest because we are directly attached to a tool/approach evaluated in this study. Some of these tools performed well, as implemented in this study; others did not. The primary author of this rebuttal (Skeem) and the final coauthor (Latessa) are professors with no such conflict of interest.

This comment focuses on “big-picture” issues most relevant to policymakers and practitioners. We leave aside specific methodological problems with Baird et al.’s (2013) report that may have affected the results.²¹

²¹ For example, the CRN was developed with one scoring method for adjudicated and probated youth, but the authors disaggregate the two samples; the YASI has a pre-screen, but the authors develop their new scale using items from the

ii. *Summary of Key Points*

In this commentary, we articulate four conclusions that can be drawn from this study. We then present the fundamental question that this study cannot address. The key points follow.

- **Conclusion 1: There is room for improvement in both risk assessment tools AND the quality with which they are implemented.** Although Baird et al. (2013) tend to attribute their findings solely to tools, their study cannot disaggregate the quality of a tool from the quality with which it was implemented. At the broadest level, their results indicate that a variety of tools, *as implemented in a variety of sites*, have room for improvement in their reliability and predictive utility.
- **Conclusion 2: Inter-scorer reliability is not self-evident.** In almost half of the sites studied, staff were unable to score the tool in a manner that was consistent with that of an expert. When staff score a tool incorrectly, the tool's ability to inform accurate decisions about youth is limited. Inter-rater reliability cannot be ignored during processes of development or implementation.
- **Conclusion 3: Risk classifications must be cross-validated and/or customized.** Above all, this study provides a compelling reminder that agencies must check and "customize" risk classifications (e.g., low, medium, high) based on local sample characteristics. Based on differences in youth populations and recidivism rates, one agency's high-risk case may be another agency's low- to moderate-risk case. When classifications are not fit to an agency, the predictive utility of an otherwise accurate tool will be forsaken in everyday practice.
- **Conclusion 4: Short tools can predict as well as (not better than) longer ones.** Most of Baird et al.'s (2013) report seems allocated to the argument that "shorter is better" and that the "Solano JSC is best." The data do not support these conclusions. The tools with the greatest predictive utility, as implemented in this study, were the Oregon JCP (31 items), Virginia YASI (32 items), and Solano JSC and Girls Link (nine items). Like past studies, this study indicates that short tools sometimes predict as well as longer ones. Similar levels of predictive utility can be achieved by (a) statistically selecting and combining a few highly predictive risk factors and (b) sampling risk domains more broadly and including risk factors that can inform risk reduction efforts.
- **Open question: What value is added by risk reduction-oriented approaches?** Contemporary risk assessment approaches are oriented toward the *prediction of recidivism*, the *reduction of recidivism*, or both. Tools oriented solely toward prediction tend to be simpler than those oriented toward reduction. Baird et al.'s (2013) study raises a question that it cannot address: What evidence is there that reduction-oriented risk assessment tools add value to those that are prediction-oriented? For reduction-oriented tools, it is not enough merely to demonstrate that adding variables

full instrument; and the PACT combines two subscales into a single risk assessment, but the authors present AUCs for two subscales as if they are independent (where a single AUC for the sum of subscales better represents the PACT).

“does no harm” to predictive utility. Precious juvenile justice resources should not be spent on pointless assessment exercises. Instead, these tools must demonstrate that the variables they add actually bring something of value to the risk-reduction enterprise. Several potential avenues exist for doing so. It is time for the field to get serious about addressing this important and challenging question.

b. *Conclusions Supported by Data*

There is room for improvement in risk assessment tools and/or their implementation. At the broadest level, the results of this study indicate that a variety of risk assessment tools, *as implemented* in a variety of sites, have room for improvement in their reliability and predictive utility.

Baird et al.’s (2013) opinion aside, the AUC is the most appropriate statistic for comparing the predictive utility of tools across sites. In part, this is because unlike the DIFR, its size is not affected by base rates of recidivism, which range from 11% to 51% across sites in this study (see Table 40).

Only one tool at one site—Oregon’s JCP—achieved an AUC of .70, the minimum level of predictive accuracy “considered acceptable for clinical application purposes” (Zhang, Roberts, & Farabee (2011), p. 5). As shown in Baird et al.’s (2013) Table 40, five tools/sites manifested a “medium” effect in predicting readjudication (i.e., $AUC \geq .649$), four manifested a “small” effect (i.e., $AUC \geq .556$), and four essentially had no effect. None of the tools/sites achieved a large effect size ($AUC \geq .712$).²³

This study cannot pull apart the quality of a risk assessment tool from the quality with which it is implemented. Although Baird et al. (2013) tend to attribute their findings solely to instruments, each finding also reflects implementation quality.²⁴ To identify high-quality tools for the field (on one hand)

²³ As shown by Rice and Harris’ analyses (1995), minimum AUCs of .556, .639, and .712 correspond to “small,” “medium,” and “large” effect sizes, respectively.

²⁴ For example, based on results for two YLS/CMI sites included in the present study, Baird et al. (2013, p. 51) conclude that “the YLS/CMI appears to have limited value as a classification tool.” Nevertheless, a large body of peer-reviewed research provides more favorable results for the predictive utility of the YLSI/CMI. The discrepancy between Baird et al.’s (2013) findings and past research are consistent with the well-validated correctional principle that implementation quality matters.

and guidelines for implementing them (on the other), future work should attempt to differentiate between these two issues. This would allow researchers and practitioners to develop guidelines for (a) demonstrating that a tool is well-validated before it is disseminated and (b) adequately implementing well-validated risk assessment tools.

Inter-scorer reliability is not self-evident. This study examines a critical, but routinely ignored issue: inter-scorer reliability. When staff score a risk assessment tool in an inconsistent or incorrect manner, that tool cannot inform accurate decisions about youth. Reliability is a necessary (but not sufficient) condition for a tool to accurately predict recidivism. It is, therefore, a key element of evidence-based practice in risk assessment.

Baird et al. (2013) found that staff provided with exactly the same information about a youth were able to attain “good” scoring agreement with other staff in nine of 10 study sites,²⁵ but attained adequate scoring agreement with an expert in only five of the 11 study sites.²⁶ In other words, staffs’ scores are often consistent with one another, but not necessarily “correct.”²⁷

Reliability problems typically reflect poorly defined items and/or inadequately trained staff. Both causes seem to be culprits here. First, across tools (from the Solano JSC to the Virginia YASI), items that were abstract and/or poorly defined tended to be less reliable. This suggests that tool developers must define items carefully and empirically demonstrate that they can be scored reliably. Second, staff at different sites scored the same tool with different levels of reliability.²⁸ This suggests that the quality of training and implementation matters—in keeping with a large body of correctional

²⁵ See Table 38, Column 6. “Good” is defined as an ICC > .75, following guidelines by Parkerson, Broadhead, & Tse (1993). (Because it is not appropriate to compute ICCs for ordinal data, the ICCs reported for “risk levels” in Table 38, Column 5 are questionable.)

²⁶ See Table 38, Column 4, which depicts the average proportion of staff scores that exactly match expert scores across items. “Inadequate” is defined as < 75%.

²⁷ This possibility could be tested by using consensus scores generated by an expert panel of scorers as the criterion for staff, rather than scores provided by a single individual.

²⁸ See Table 38, where the YLS/CMI attains an ICC of .80 in Nebraska, but only .67 in Arkansas.

treatment research. Agencies should train their staff until they attain a specified level of reliability, and then periodically reassess whether staff are scoring the tool correctly.²⁹

Risk classifications must be cross-validated and/or customized. Above all else, the results of this study provide a compelling reminder that agencies must check and “customize” risk classifications, based on local sample characteristics (see Andrews & Bonta, 2003). Risk classifications involve nothing more—and nothing less—than chopping up a continuous score on a risk assessment tool to create a number of ordinal categories (e.g., “low,” “medium,” “high”). Tool developers often use a particular sample of youth to optimize risk classifications, i.e., identify cut scores that create reasonably sized groups of youth with recidivism rates that are as different as possible. Using the language of the DIFR statistic, one goal is to maximize “base rate dispersion” (Silver, Smith, & Banks, 2000).

The problem is that risk classifications that are optimized in one sample can degrade when they are applied to a new sample—particularly when the new sample has a much different risk score distribution, base rate of recidivism, or both. Based on differences in their youth populations and recidivism rates, one agency’s high-risk case may be another’s average bear.

This underscores the necessity of locally assessing and validating the predictive utility of risk assessment scores and classifications. In some cases, risk classifications will not be meaningful unless they are customized. One sign that this is the case is when the predictive utility of scores (as indexed by the AUC) outstrips the discrimination ability of classifications (as indexed by the DIFR). Based on the sites studied by Baird et al. (2013; see Table 40), this “outstripping” happens often enough to be concerning. Specifically, risk assessment scores moderately predicted new adjudications in five sites

²⁹ These two factors probably interact. Even though research indicates that they robustly predict criminal behavior, abstract risk factors like criminal attitudes or poor parental supervision are harder to measure than concrete risk factors like criminal history. Tools probably vary in how well they measure those abstract risk factors. Sites vary in how well they train and monitor staff. When an abstract risk factor manifests poor predictive utility on a tool within a site, is that a fault of the tool, a problem with its implementation in that site, or both? Without additional information, it will be impossible to tell.

(i.e., $AUC \geq .639$).³² Although risk classifications also performed well in three of these five sites (i.e., high DIFR for Oregon JCP, Solano JSC, Virginia YASI), they performed poorly in the remaining two (i.e., low DIFR for Girls Link Solano, CRN Georgia).³³

For example, in Georgia there is a 64% probability that a (randomly selected) adjudicated youth will obtain a higher CRN score than a (randomly selected) non-adjudicated youth ($AUC=.64$). Therefore, CRN scores do a moderately good job of distinguishing between youth with—and without—a new adjudication. However, CRN classifications performed relatively poorly ($DIFR = .40$). Specifically, there wasn't much difference between "moderate" and "high" groups in their adjudication rates. This is a sign that the agency needs to customize cut scores to their sample. If the agency uses risk classifications that do not fit their sample to inform decision making about youth, then they are forsaking the predictive utility of scores on that tool. The "high"-risk youth isn't meaningfully different from the "moderate"-risk youth.

Assuming that *scores* on the tool are predictive in the new agency, the good news is that risk classifications can be modified to fit the new agency's population. Ideally, an agency would modify risk classifications not only to maximize their base rate dispersion, but also to fit the decision(s) that they want classifications to inform. For example, if the goal is to identify low-risk cases to divert from detention, then (a) only two risk classifications are needed ("low" and "not low"), and (b) the cut score can be adjusted (within limits) to be lower or higher, to reflect that agency's weighting of public safety, youth rights, and resource concerns.

In short, this finding is an important call to the field to get serious about cross-validating and (if necessary) customizing risk classifications to their setting. As Baird et al. (2013) note, agencies tend to use classifications more than scores to inform their decision making about youth. These results

³² See footnote 21 above for AUC interpretation guidelines.

³³ No interpretation guidelines (e.g., "small," "medium," "large") are available for the DIFR. Users must be cautious in applying the DIFR because its size is affected by recidivism rates.

suggest that researchers and policy makers should articulate guidelines for cross-validating and customizing risk classifications. Ensuring that risk classifications are valid is essential when implementing any risk assessment tool.

Short tools can predict recidivism as well as (not better than) longer ones. Some of the tools included in this study are relatively short and simple (i.e., the Solano JSC with nine items and Arizona AOC with nine items); most others are relatively long and/or complex (like the Virginia YASI, 32 items). Loosely, these tools represent an evolution in risk assessment over time, from prediction-oriented approaches (which were designed solely to achieve efficient prediction) to reduction-oriented approaches (which also emphasize variable risk factors that theoretically can be changed to reduce risk).³⁴

Most of Baird et al.'s (2013) report seems allocated to the argument that "shorter is better" and that the "Solano JSC is best."³⁵

In their introduction, the authors caution, "If changes to risk assessment instruments have resulted in diminished capacity to accurately discriminate among high-, moderate-, and low-risk youth, then decision making in juvenile justice has been adversely affected." In addition to planned analyses that test the reliability and predictive utility of each instrument at each site, the authors perform extensive post hoc analyses in an attempt to (a) create shorter and (ideally) more predictive versions of relatively long tools and (b) create a JSC proxy that (ideally) predicts better than rival tools. The authors conclude that "the JSC, used in Solano County, proved to be the most successful risk instrument evaluated in this study;" that their Solano JSC proxy "transferred better" than rival tools;

³⁴ For a review of this evolution and the confusion it has created, see Monahan and Skeem (2013).

³⁵ This argument is apparent in the authors' review of past research, which is highly selective. For example, the authors select only the least favorable finding (from 20 largely positive comparisons) when referencing results from a New York YASI sample (Orbis Partners, 2007).

that “complex scoring systems ... diminish results;” and that most shorter instruments they created “produced *markedly better* results than the instrument currently in use.”

The study’s results do not support these conclusions. First, although the results of planned analyses indicate that predictive utility varies across sites (see Table 40), there is no evidence that this variability is a simple function of a tool’s length or complexity.³⁶ For example, scores on both the Virginia YASI and Solano JSC—this study’s prototypes of “long/complex” and “short/simple”—manifested good inter-rater reliability (ICCs = .89-.92; see Table 38) and equivalent predictive utility (AUC = .68 for both; see Table 40). Indeed, the tools with the most predictive scores and classifications were a locally created tool (the Oregon JCP, 31 items), a simple public domain tool (the Solano JSC, nine items), and a “later generation” commercial tool (the Virginia YASI, 32 items).

Second, at best, the results of the authors’ post hoc analyses demonstrate that short tools can predict re-adjudication as well as longer ones. The authors created new tools for 10 sites to maximize prediction within each dataset by (a) using statistical criteria to select and combine variables and then (b) customizing risk classifications (see above).³⁷ As a rule, tools constructed in this way capitalize on chance associations between variables in a particular sample and will “shrink” in predictive power when applied to new samples. So, tools must be cross-validated with an independent sample.³⁸

³⁶ Baird et al. (2012) could directly test the relationship between tool length (and/or complexity) and predictive utility by performing a small meta-analysis with their data. We did not do so because they do not operationalize either variable in their report (i.e., item number, item/scoring complexity) ... and there are too few exemplars of short/simple tools (i.e., two) to support an adequate test.

³⁷ Unfortunately, the authors conflated the development of new tools with the customization of classifications. Their analyses would have been much more informative if they had customized risk classifications based on original scores to assess the degree of improvement this yielded before developing new scores and risk classifications that were tightly fitted to a particular dataset. (This is another reason the AUC is a more comparable indicator of performance across tools than the DIFR, particularly in this study.)

³⁸ The table below compares the predictive utility of the original tools and of Baird et al.’s new tools that were not cross-validated. Estimates for the new tools are likely inflated because the same sample was used to optimize and “test” the new tool. Still, the pattern of results suggests that tools with moderate predictive utility were difficult to improve, regardless of their length. The unvalidated new tools generally did not predict recidivism better than the original Oregon JCP, Solano

Baird et al. cross-validated three of the 10 tools they created. The authors did not test whether their new tools predicted re-adjudication significantly better than the original tools (by testing differences in AUCs or any other statistic). In fact, for some instruments, they did not even provide estimates of predictive utility that could be directly compared (e.g., inconsistently separating estimates by gender). Nevertheless, as shown in Table 54 below, there generally is little difference in the predictive utility of the original (“longer”) tools and cross-validated new (“shorter”) tools. The average AUC difference is .02. The average DIFR difference was also a modest 0.13—and this difference may be based more on customization of risk classifications than on any substantive change to the tool (see footnote 34). The only direct comparison that can be made is for the Arizona AOC, where performance is essentially equivalent. Incidentally, the new scale had *more* items than the original Arizona AOC scale.

JSC, Girls Link Solano, or Virginia YASI (average AUC difference = .02). There was more room for improvement among scales with weaker utility (Arizona DJC and YLS/CMI Nebraska, average AUC difference = .09). The degree of improvement appears unrelated to the degree of shortening.

Predictive Utility of Original Tools and Non-Cross-Validated New Tools				
Assessment	Number of Items, Original vs. New	Original Tool AUC	New Tool, Construction AUC	AUC Difference
Arizona DJC DRI	18 vs. 15	.59	.69	.10
YLS/CMI Nebraska Probation	42 vs. 16	.55	.61	.06
YLS/CMI Nebraska Commit.	42 vs. 11	.54	.66	.12
Oregon JCP	31 vs. 12	.70	.70	.00
JSC Solano County	10 vs. 9	.68	.70	.02
Girls Link Solano County	10 vs. 9	.68	.73	.05
YASI Virginia	32 vs. 15 (boys) and 11 (girls)	.68	.71 (boys) and .74 (girls)	non-nested

Table 54			
Similar Predictive Utility for Original Tools and Baird et al.'s (2013) New Tools			
Assessment	Number of Items, Original vs. New Tool(s)	Original Tool AUC (boys and girls combined)	New Tool, Cross- Validation sample AUC
Arizona AOC Risk Assessment Instrument	9 vs. 12	.62	.63
PACT Florida Probation	22 vs. 12 (boys) and 11 (girls)	.59 "Criminal"; .63 "Social"	.66 boys only; .66 girls only
CRN Georgia	59 vs. 9	.64	.67 boys only

A similar case is apparent when the performance of original instruments is compared with the Solano JSC proxies. As shown in the table below, the "longer" tools perform about as well as the "shorter" tools. It is highly unlikely that AUC differences of .02 or .03 are statistically significant.

Table 55			
Similar Predictive Utility for Original Tools and Solano JSC Proxy			
Assessment	Number of Items, Original vs. New Tool	Original Tool AUC	Solano JSC Proxy AUC
PACT Florida Probation	22 vs. 12	.60 "Criminal"; .62 "Social"	.63
CRN Georgia	59 vs. 9	.64	.66

Fundamentally, this study provides evidence that tools that differ in their length, format, and foci can achieve similar levels of predictive utility. This finding is consistent with research on the relative predictive utility of alternative risk assessment tools that, as a group, are much better validated than those studied here.³⁹ Despite heated debate about which type of tool predicts best ("actuarial" vs. "clinical," simple vs. complex; etc.), research is making it increasingly clear that there is no winner in this horse race. For example, in a meta-analysis of 28 separate studies, Yang, Wong, and Coid (2010) found that the predictive efficiencies of nine validated risk assessment

³⁹ See Campbell, French, & Gendreau (2009); Kroner, Mills, & Reddon (2005); and Yang, Wong, & Coid (2010).

instruments were essentially “interchangeable,” with estimates of accuracy falling within a narrow band (AUC = .65 to .71). The tools examined included a short actuarial device that emphasizes simple risk markers (the Violence Risk Appraisal Guide), a more clinically oriented tool that emphasizes variable risk factors (the Historical Clinical Risk Management-20)—and virtually everything in between (like the LSI-R).

Two factors may help explain the similar predictive performance of well-validated instruments. First, these tools seem to tap “common factors” or shared dimensions of risk, despite their varied items and formats.⁴⁰ Second, these tools seem to reach a “glass ceiling” of predictive utility beyond which they cannot improve. If a limiting process makes recidivism impossible to predict beyond a certain level of accuracy, each tool can reach that limit quickly with a few maximally predictive items before reaching a sharp point of diminishing returns. Baird et al.’s (2013) post hoc results are consistent with this possibility and echo the results of other studies. For example, based on a sample of over 1,000 released prisoners, Coid et al. (2011) found that most individual items included in risk assessment tools did not significantly predict violence. When these items were removed, the resulting reduced scales predict violence as well as (but usually not better than) the original full scale. For example, a five-item version of a prediction-oriented scale (the VRAG) performed as well as the full 12-item version (AUCs = .70, .71, respectively). It is important to recognize that if there is a glass ceiling, it can be reached via alternative routes. If measured validly, some variable risk factors (e.g., attitudes supportive of crime) predict recidivism as strongly as common risk markers (e.g., early or “pre-adult” antisocial behavior; Gendreau et al., 1996).

⁴⁰ In an innovative demonstration, Kroner, Mills, and Reddon (2005) printed the items of four well-validated instruments (e.g., LSI-R, VRAG) on strips of paper, placed the strips in a coffee can, shook the can, and then randomly selected items to create four new tools. The authors found that the “coffee can instruments” predicted violent and nonviolent offenses as well as the original instruments did. Factor analyses suggested that the instruments tap four overlapping dimensions: criminal history, an irresponsible lifestyle, psychopathy and criminal attitudes, and substance-abuse-related problems. Each of these dimensions were predictive of recidivism.

In short, similar levels of predictive utility can be achieved by (a) statistically selecting and combining a few highly predictive risk factors and (b) sampling risk domains more broadly and including risk factors that can inform risk-reduction efforts. For these reasons, Skeem and Monahan (2011) concluded:

“Given a pool of instruments that are **well-validated** for the groups to which an individual belongs, our view is that the choice among them should be driven by the ultimate purpose of the evaluation. If the ultimate purpose is to characterize an individual’s likelihood of future [criminal behavior] relative to other people, then choose the most efficient instrument available. This is appropriate for a single event decision in which there is no real opportunity to modify the risk estimate based on future behavior. If the ultimate purpose is to manage or reduce an individual’s risk, then value may be added by choosing an instrument that includes treatment-relevant risk factors ... This choice is appropriate for ongoing decisions in which the risk estimate can be modified to reflect ebbs and flows in an individual’s risk over time.”

c. *Open Question: Does Reduction-Oriented Risk Assessment Add Value?*

At its core, the study by Baird et al. (2013) raises a fundamental question that it cannot address: What evidence is there that reduction-oriented risk assessment tools add value to those that are prediction-oriented? It is time for the field to get serious about addressing this important and challenging question.

At the risk of oversimplification, Baird et al. (2012) mistakenly assume that the only purpose of risk assessment is classification; and the only real measure of a tool’s performance in meeting that

purpose is predictive utility (i.e., base rate dispersion). Their yardstick of success is defined by parsimony and predictive utility. Period.⁴¹

This yardstick is both sensible and sufficient when the ultimate purpose of risk assessment is merely to characterize a youth's likelihood of recidivism compared to other youth. In this case, what the tool assesses is irrelevant because there is no interest in explaining or reducing risk. For example, if a tool that efficiently assesses accuracy in playing street dice strongly predicts recidivism (see Nunnally, 1978), then the tool is valid for characterizing risk. As summarized by Gottfredson and Moriarty, "if a variable can be measured reliably, and if it is predictive, then of course it should be used—absent legal or ethical challenge."

When the ultimate purpose of risk assessment is to reduce a youth's risk of recidivism, predictive utility is a necessary—but not sufficient—measure of success. Contemporary thinking and "later generation" risk assessment tools have been infused with the concepts of risk management and risk reduction. Theoretically, these tools add value to simple tools by assessing variable risk factors (e.g., antisocial attitudes; poor parental supervision)⁴² that may help explain the process that leads to recidivism. The goal is to inform risk reduction efforts by (a) specifying risk factors to target in treatment and (b) capturing any changes in risk over time to inform ongoing decisions about supervision and treatment.

⁴¹ The evolution of correctional risk assessment tools has created a largely artificial distinction between "risk" and "needs" assessment (see Monahan & Skeem, 2013). "Risk" assessment tends to be reduced to an actuarial formula that heavily weighs risk markers. Sometimes the items that comprise this formula are explicitly separated from other items (e.g., Baird's JAIS/JSC), and sometimes they are embedded among other items (e.g., YASI, PACT, CRN). "Needs" assessment tends to be whatever content remains on the tool once the predictive items have been removed. Baird et al.'s (2013) evaluation criteria imply that the "risk" part of these tools is subject to scientific scrutiny, but "anything goes" for needs assessment. If the field follows this suggestion, few gains will be made in understanding and reducing risk among youth. Instead, we believe that the field can and should evaluate whether these tools—in their entirety—are capable of fulfilling their intended purposes.

⁴² Variable risk factors are variables that have been shown to predict recidivism and to be changeable (see Monahan & Skeem, 2013 for clear definitions of fixed markers, variable markers, variable risk factors, and causal risk factors). Sometimes variable risk factors are called "dynamic risk factors" or "criminogenic needs."

Baird et al.'s (2013) yardstick is not sufficient for measuring the success of these tools. All risk assessment tools must manifest adequate predictive utility ... but this only gets "later generation" tools to first base. For these tools, it is not enough merely to demonstrate that adding variables "does no harm" to predictive utility. Precious juvenile justice resources should not be spent on pointless assessment exercises. Instead, these tools must demonstrate that the variables they add actually bring something of value to the risk reduction enterprise. There are several potential avenues for doing so. For example, one could test a tool's construct validity to determine whether it actually measures the variable risk factors that it says it measures (for an example, see Kennealy, Hernandez, & Skeem, 2013). Or test whether variable risk factors assessed by a tool change over time and whether those changes predict recidivism (for an example, see Howard & Dixon, 2013). Or test in a well-controlled study whether youth are significantly less likely to recidivate when professionals use a reduction-oriented rather than prediction-oriented assessment approach. The most rigorous (and treatment-relevant) test would be a randomized controlled trial in which a targeted intervention was shown to be effective in changing a variable risk factor(s) on a tool, and the resulting changes were shown to reduce the likelihood of post-treatment recidivism (see Monahan & Skeem, 2013).

Practice has far outpaced research at this intersection between risk assessment and risk reduction. An absence of evidence that these tools add value to risk reduction efforts, however, is not the same as counter evidence. We strongly recommend that researchers and policy makers work together to articulate concrete measures for testing the value added by reduction-oriented risk assessment tools. The time could not be better to take on this challenge, given the current level of interest in using science to inform real problem solving in the juvenile justice system.

3. Authors' Responses to Comments

We begin by responding to specific concerns raised by advisory board members, then close with general observations. First, however, it is important to note that in drawing their conclusions,

Skeem et al. relied exclusively on the AUC to compare results from systems tested as well as those developed as examples of how actuarial instruments could improve classification. They base the exclusive use of the AUC on two publications. Although use of the AUC has accelerated in recent years, particularly in justice research, the AUC has its detractors. Some analysts have cautioned that the AUC does not account for distribution problems or base rates (Rice & Harris, 1995). Other distinguished researchers rely completely on measures other than the AUC in their studies of risk instruments (for example, Gottfredson & Snyder, 2005; Altman & Royston, 2000). The view of this group of advisory board members appears to be that tools with similar AUCs should produce approximately equal classification results. It is only a matter of selecting the proper cut-off points. There is little evidence, however, that this is true. In our study, there was only one instance (that of the CRN in Georgia, a very unusual circumstance where there was a narrow range of scores and two factors accounted for virtually all of the discrimination attained, rendering all other risk factors irrelevant) where a change in cut-off points significantly improved classification results. In the sites where we were able to construct and validate actuarial risk instruments, these instruments produced significantly better classification results despite producing only marginal improvements in the AUC.

If similar AUCs equal similar classification capability, it is hard to understand why better results were not obtained with the instruments used in these organizations. The Florida PACT experience is instructive. Several validations have been completed over the years; given the similarity in AUC scores obtained for the PACT and the risk tool created in this study, these validations should have produced a classification scheme with results that mirrored those of the newly created actuarial instrument or the simulated JSC. In fact, classification results from both of the latter instruments represent significant improvements over those produced by PACT.

It is our view that the AUC represents one measure of validity and certainly not the best measure. Its reported advantages are also weaknesses. It does not reflect either the base rates or

distribution of cases across risk levels. Hence a risk tool with little practical utility could attain a high AUC score.

We have long maintained that classification systems must be judged on four criteria: validity, reliability, equity, and utility. This group's response ignores both equity and utility, despite the fact that equity issues were found with several instruments.

Debates about appropriate measures of validity can go on forever without providing much guidance to the field. The most important point is this: All of the tools evaluated in this study assign cases to different risk levels. Either implicitly or explicitly, the risk level plays a role in case decision making, ranging from assigning a supervision level in the community to helping determine if a youth should be incarcerated. Given the importance of the risk level assigned, agencies need systems that optimize differences in outcomes observed for cases at different risk levels. The grant proposal clearly stated that this would be the primary measure of validity. Classification results cannot simply be ignored.

Only three models produced a satisfactory level of discrimination: the Oregon JCP, the JSC (Solano County) and the YASI (Virginia). The Oregon model is an elegant system designed for and, to our knowledge, used only in Oregon. It may transfer well, but no data supports its use in other jurisdictions as of yet. The YASI results are based on limited implementation in Virginia and far exceed results produced elsewhere. We do not feel "Solano is best." We believe, based on the fact that it represents a compilation of actuarial research conducted in 14 jurisdictions (Wiebush, 2002), that it is a simple, easily implemented instrument and that the results of this study found not only a high degree of validity and reliability, but equity as well. It is more likely to work across organizations than other general use instruments tested in this study.

Following are responses to additional specific points raised by Skeem et al.

Authors' Note # 1: The fourth paragraph by Skeem et al. discusses potential conflicts of interest (p. 108). This attempt to raise the issue of a conflict of interest blurs reality. First, Latessa was asked to join the advisory board because he and colleagues at the University of Cincinnati have been long-time supporters of the LSI family of risk assessment instruments. While we do not maintain this represents a conflict of interest—Latessa also actively promotes a model he developed in Ohio—it is important to note that his views may be influenced by past associations. Second, the introduction as written leaves the impression that Latessa attended the meeting in Baltimore. He did not. This group obviously reached out to him, but did not make a similar attempt to solicit input from two other independent researchers, James Howell and Aron Shlonsky.

NCCD is a nonprofit research organization. While NCCD developed the JSC instrument, it is public domain and is not an instrument we actively promote. It is used as a “placeholder” in our Juvenile Assessment and Intervention System™ (JAIS), but it is replaced with any risk model that has been sufficiently validated on the population implementing JAIS™. We also use the database to validate and revise, if appropriate, any instrument including the JSC embedded in JAIS. Our approach has always been to recommend developing risk instruments in the jurisdiction that will use the tool. Over the last three decades, we have developed dozens of such risk instruments. The JSC is a compilation of work conducted in 14 agencies, but we have not marketed or generally promoted its use.

Had this study included an evaluation of JAIS, there would be conflict of interest. JAIS, however, is principally a method for developing supervision strategies; any valid actuarial risk instrument can be embedded in the JAIS model.

Authors' Note # 2: In their comments, Skeem et al. suggest methodological concerns that they explain in footnote 21 (p. 108). We fail to see how any of the issues cited in footnote 21 represent methodological problems. We disaggregated the population in Georgia (and other jurisdictions) so

that direct comparisons could be made between sites—some of which only supervise probationers, and others only supervise aftercare or parole. It is also important to understand how the CRN works for each population within Georgia. Complete data on each subsample including sample size is reported, which allows any reader to combine results. However, combining the results has no impact on overall findings.

Regarding development of the actuarial scale in Virginia, there was no reason—other than possibly to maximize the sample size—to constrain the analysis to factors on the YASI pre-screen. Each analysis was already constrained by factors collected by each system. It is also important to note that we did not recommend any of the instruments developed in these limited analyses. The purpose of these analyses was simply to determine if a focus on factors with strong relationships to recidivism could produce better discrimination.

Finally, the footnote states that the criminal history and social history risk instruments are combined to establish a risk level in Florida, but how they are combined is not transparent. In correspondence, this group recommended that totals from the two scales be summed and the AUC be computed on summed score. In practice, they are not summed, but cross-referenced in a matrix. Thus, if summed, the same score can result in as many as three different risk levels, depending on the individual risk factors that were checked. Summing scores from the two scales also results in a situation where a score of 14 could be rated low risk, while a score of 12 could result in a moderately high risk rating. If the youth scoring 12 was a true negative, the youth scoring 14 was a true positive, and these cases were randomly selected and compared, the results would help increase the AUC score but clearly would not support the validity of the system.

This circumstance would help to validate a scoring system that does not exist. The methods used to validate the Florida PACT model were appropriate. Although beside the point and of little relevance, summing the totals from the two scales makes little difference in the AUC scores attained for the PACT system.

Authors' Note #3: In the summary of key points, first conclusion, Skeem et al. allege that the authors attributed findings solely to tools without adequately considering implementation (p. 109). We agree that quality training and implementation are important to system fidelity and that both can impact reliability and validity results. While we made attempts to gather information on both issues, it was ultimately impossible to gauge their impact on outcomes in any site.

That said, we did attempt to select sites where model developers thought training and implementation were handled well. We, in fact, changed our initial PACT site from a California agency to Florida, based on recommendations from Assessment.com. Furthermore, based on data collected and analyzed in this study, there is little indication that implementation and training efforts were below standard in nine of the 10 participating agencies. Most either involved system developers or followed their recommendations. In one agency, Arkansas, too little data was available to reach conclusions regarding any aspect of the system.

While it is not accurate to say that we attributed differences solely to the risk tools used in each jurisdiction, we do believe that design issues can and do create problems with both training and implementation. It is difficult to argue that the longer, more complex systems do not require more training, complicate implementation, and have greater propensity for error. (Examples of design issues that impact reliability were identified and discussed earlier in this section of the report.)

In essence, just as validity and reliability cannot be completely separated from the quality of training and implementation support provided, training and implementation issues cannot be entirely separated from the complexity and design of the tool itself.

Authors' Note #4: In their second conclusion under summary of key points, the group emphasized the importance of reliability (p. 109). We agree that reliability testing is crucial. The problem is that several of the instruments evaluated in this study were marketed before sufficient reliability testing was conducted.

Authors' Note #5: The third conclusion under summary of key points argues for the customization of risk assessments (p. 109). We are very pleased to see that this group concurs with a long-standing NCCD position on the need to customize. This perspective is particularly important because two members of this group served in an advisory capacity to the Models for Change (MacArthur Foundation) effort to develop an implementation guide that states “local validation is not required” if an instrument has been validated in three or more sites or the “agency has evidence of multiple validations in similar settings.” Clearly, the results of this study challenge that view.

Authors' Note #6: The fourth conclusion under summary of key points suggests that the authors argue that shorter risk assessments are better (p. 109). This discussion misses our point entirely. The primary issue we identify is not the number of items that comprise a risk scale, but how risk items are selected, what they represent, and their relationship to recidivism. In fact, given the differences in design, it is difficult to compare the number of items in different tools. For example, in the YLS/CMI “count,” two possible ratings of substance use are counted as separate items. In other instruments, several ratings of substance use are combined as “values” within a single factor. Clearly, such “counts” do not mean much. What may appear to be a long list of factors in Oregon, for example, in reality comprises a fairly concise tool. Our point is that risk tools should only include factors that are related to recidivism and aid in the classification process.

It is true, almost by definition, that tools that focus solely on classifying youth based on recidivism rates tend to be more concise than tools that introduce additional goals and objectives (e.g., risk reduction). Hence, a “shorter-is-better” narrative emerges. The real difference between the views of this group and what we believe this study supports is that instruments that focus solely on differentiating youth based on proclivities for future offending are better classification tools. A more complete response to adding a “risk reduction” purpose to risk assessment is presented later.

Some risk tools include items for which we can find no research that establishes any relationship to recidivism (or “risk reduction,” for that matter). In other instances, factors that may have a relationship to recidivism for particular populations have remained on tools after it is demonstrated that they have little or no correlation to recidivism in agencies where they are being used. Both circumstances have potential to reduce the efficacy of the risk assessment tool.

Authors’ Note #7: In footnote 32, page 114, Skeem et al. suggest that the authors were selective in reviews of relevant literature. We were not highly selective in our review of past research. The lone example cited to support their point, however, is highly selective. We have already responded to Orbis Partners, Inc.’s concerns, clarifying why we selected this data element and demonstrating that other suggested comparisons from the New York data were not, in reality, any more positive than what was presented. Readers are directed to the footnote on p. 9 for a full explanation.

Authors’ Note #8: On page 115, Skeem et al. point out the need to cross-validate using an independent sample. We concur that instruments generally produce the best results for the sample on which they were developed (the construction sample). When samples had sufficient cases available to support the use of construction and validation samples, this was done. For two of the largest samples, results from validation samples are presented. Thus, where possible, cross-validation was conducted.

Authors’ Note #9: Skeem et al. also raise the issue of separating risk and needs in assessments (see specifically footnote 38 on p. 120). We have never taken the position that “anything goes” for needs assessment. We developed the first needs assessments used in both adult and juvenile justice. They are structured, anchored with definitions and scoring guides, and, as research has shown, quite reliable. Our point is that identifying certain needs as “criminogenic” based on group data and assuming that this relationship means anything at the individual case level is not “scientific scrutiny.” It implies a power that risk assessment cannot legitimately claim. To understand the factors that are

influencing criminal behavior in an individual offender requires a clinical evaluation or a system designed to provide clinical insight. Labeling a need as “non-criminogenic” because it has a limited correlation with recidivism conflates the appropriate roles of group and individual data and can be dangerous. Low self-esteem, to use their example, may infrequently cause a youth to commit a crime (and exhibit a low correlation with recidivism), but in certain instances it may be a major driver of offending behavior (youth who commit violent crimes in schools may be a prime example).

Our point is simply this: Some needs exhibit moderate correlations with recidivism and these should be treated as risk factors. However, the mere existence of such needs does not mean they are criminogenic for an individual offender. Other needs, with little statistical relationship to recidivism could be the most important to address to reduce the risk represented by an individual. No risk instrument, by itself, is equipped to make this judgment. Additional assessment is required. Yet, the language incorporated in marketing tools implies such capability. The YLS/CMI for example, professes to address “responsively”...matching programs to offender needs and learning styles. There is nothing in the system that would provide such insight.

Authors’ Note #10: In the same section, fifth paragraph, the group suggests that the goals of risk assessment go beyond the accurate estimation of the likelihood of future delinquency (p. 120). Their statement, “When the ultimate purpose of risk assessment is to reduce a youth’s risk of recidivism, predictive utility is a necessary—but not sufficient—measure of success” underscores the source of the problem with many “later generation” risk tools. First, it should be obvious that the purpose of any assessment system cannot be “risk reduction.” Risk reduction can only be achieved through interventions (counseling, treatment, education, etc.), maturation, or both. The purposes of assessments in juvenile justice include (1) identifying youth most likely to recidivate; (2) identifying treatment issues that need to be addressed; and (3) identification of interventions most likely to increase success of the youth (sometimes referred to as “responsivity”). It is our position that these

purposes are best addressed through a combination of assessments, each with a well-defined goal. To combine all of these issues under the rubric of risk assessment conflates the role and utility of group data with individual case factors relevant to case planning and intervention, establishes unrealistic expectations of risk instruments, and can result in measures that significantly misrepresent the power of “dynamic risk factors.”

As the respondents note, systems are needed “to identify factors to target in treatment and to track reductions in risk levels to inform ongoing decisions about supervision and treatment.” However, there is absolutely no need to address all of these issues in a single risk instrument. Reclassification schemes have been around for decades (a fact largely ignored in academic journals) that effectively cover all of these issues, yet keep initial risk assessment focused solely on optimally classifying offenders to different risk levels. Criteria are changed at reassessment to focus on the current behavior of the youth and progress made in treatment programs. Changes in both risk levels and needs can be tracked and programs can be evaluated for effectiveness.

Authors’ Note #11: The last section of the prior comment is a call for reduction-oriented risk assessments (p. 121). There is, at this point, no evidence that instruments focusing on risk reduction produce lower recidivism rates. Nor do we think there will be because systems using actuarial risk assessment that focuses solely on optimum classification of cases do not stop there. NCCD’s approach, for example, begins with risk assessment, moves on to assessment of needs, and completes the process with a clinically oriented evaluation to identify (a) needs that are driving delinquent behavior and (b) programs and supervision strategies most likely to reduce recidivism.

We think the conclusions drawn in this study are accurate. In most sites we were able to create actuarial instruments that significantly improved risk classification. Although results based solely on construction samples need to be viewed with caution, validation samples were used in two sites with little decline in the levels of discrimination attained in the construction samples. In addition, when the

Solano JSC instrument was simulated in the two largest agencies in the study, it produced better overall results despite the fact that the tools in use in both sites have been revalidated, providing the opportunity for customization this group says is required. The group's exclusive focus on AUC scores allows them to simply ignore these analyses.

Howell and Shlonsky linked use of valid risk information to the practice environment—specifically, use of a comprehensive needs assessment to identify individual youth treatment foci and the importance of a continuum of services, incentives, and sanctions. We concur with Howell and Shlonsky's statement that an accurate risk assessment must have static and dynamic risk factors, and each risk factor, static or dynamic, must have a strong relationship to re-offending. As evidenced in this study, assessments built from a risk reduction perspective do not result in more accurate and specific estimates of recidivism.

The reason for an objective risk assessment is to ensure appropriate allocation of resources to guarantee that youth are served and supervised relative to individual needs and risk of re-offending. Risk assessment needs to focus on one thing: the optimal classification of cases to different risk levels. Other objectives should reside with instruments specifically designed to address those objectives or with programs and supervision strategies designed to respond to issues identified by assessment tools.

V. LIMITATIONS

As with any study of this magnitude, several issues were encountered that suggest some caution in interpreting results. First, juvenile justice systems vary in policies and practice, the way juvenile records are collected within and across the system, and as importantly, in the way recidivism is measured. We made every effort to accommodate differences across sites; however, differences in policies and practice resulted in very different rates of recidivism among the jurisdictions that

participated in this study, and that fact alone had an impact on findings. In addition, sites were at different points in implementation; some agencies had used risk assessment for years, while others had recently completed implementing a new risk assessment instrument.

Arkansas had just begun collecting YLS/CMI results electronically at the time of the sample. We were able to match YLS/CMI data to about 42% of youth released from secure commitment. Due to small sample size issues, we were unable to employ a standardized 12-month follow-up period and instead limited the follow-up to nine months following release from secure commitment. Virginia was in the early stages of a phased implementation of the YASI at the time of the sample; therefore, the sample was limited to youth for whom a YASI had been completed, which was about 20% of all youth who started probation during the sample timeframe.

The Florida sample timeframe is earlier than that in other sites because data were provided for youth whose probation ended, rather than began, during the sample timeframe. We were able to identify probation start dates for nearly all youth, so the standardized follow-up timeframe aligns with cohorts from other sites. In Florida and Virginia, there was no way to distinguish a PACT or YASI assessment from a reassessment; therefore, a few of the assessments in the study could be reassessments. To ensure sufficient sample size in Solano County, we included assessments conducted at any time for youth in the sample cohort.

In addition, limited data availability affected the ability to construct revised risk assessments. At the time of the study, Arizona AOC was in the process of implementing a new needs assessment because of reliability issues with their existing one. This eliminated needs items as possible items to consider on the revised risk assessment and limited our capacity to construct a valid risk assessment for girls, even though results indicate that one would be beneficial. We encountered similar issues for the Nebraska, Arkansas, and Florida commitment populations. In Nebraska and Arkansas, needs data for the YLS/CMI were not recorded electronically and were therefore unavailable for consideration on revised risk assessments. Due to limited relationships between YLS/CMI items and recidivism (and the

small cohort of cases and limited follow-up period available in Arkansas), we were unable to construct a valid risk assessment for use in Arkansas or Nebraska OJS. In Florida, the relationships between PACT items and subsequent recidivism were limited, particularly across race/ethnicity; therefore, we were unable to construct a revised risk assessment for youth released from secure commitment in Florida.

Finally, some sites were able to provide limited recidivism information. Only returns to state facilities were provided for Arkansas cases and youth released from Arizona DJC facilities. This limited our capacity to examine recidivism measures in depth in these sites. In the interests of limiting the size and complexity of this report, re-adjudication (or readmission for Arkansas and youth released from Arizona DJC facilities) is the only measure of recidivism reported in the main body of the report. This worked reasonably well for most sites, but because of the low rate of re-adjudication in Oregon, re-arrest is probably a better measure for that jurisdiction. To view results based on multiple measures, see Appendix B.

VI. CONCLUSION

The proper use of valid, reliable risk assessments can clearly improve decision making. However, results of this study indicate that the power of some risk assessment instruments to accurately classify offenders by risk level may have been overestimated. Only three of the risk instruments examined demonstrated considerable capacity to accurately separate cases into low, moderate, and high risk levels with progressively higher recidivism with each risk level increase. Several risk instrument approaches emphasized over the last decade have substantial shortcomings and fail to convey what is most important to correctional administrators: the difference in outcomes between risk levels and the distribution of cases across the risk continuum.

The lack of standards for measuring validity and reliability of risk assessment instruments further complicates decision making for administrators. Greater emphasis should be placed both on

reliability testing and validation studies before and after risk assessment instruments are transferred to other jurisdictions. This is an area where national standards could be established to ensure due diligence.

Risk assessment should be a simple process that can be easily understood and articulated. This study's findings show that simple, actuarial approaches to risk assessment can produce the strongest results. Adding factors with relatively weak statistical relationships to recidivism—including dynamic factors and criminogenic needs—can result in reduced capacity to accurately identify high-, moderate-, and low-risk offenders.

REFERENCES

- Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19(4), 453–473.
- Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson.
- Andrews, D. A., & Bonta, J. (1995). *LSI-R: The Level of Service Inventory-Revised*. North Tonawanda, NY: Multi-Health Systems.
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency*, 52(1), 7.
- Andrews, J. D. (1990). Interpersonal self-confirmation and challenge in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 27(4), 485.
- Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Institute on Crime, Justice and Corrections at the George Washington University.
- Baglivio, M. T. (2009). The assessment of risk to recidivate among a juvenile offending population. *Journal of Criminal Justice*, 37(6), 596–607.
- Baglivio, M. T., & Jackowski, K. (2012). Examining the validity of a juvenile offending risk assessment instrument across gender and race/ethnicity. *Youth Violence and Juvenile Justice*, 11(1), 26–43.
- Baird, C. (2009). A question of evidence: A critique of risk assessment models used in the justice system. *Special Report*.
- Baird, C. (1991). *Validating risk assessment instruments used in community corrections*. Madison, WI: National Council on Crime and Delinquency.
- Baird, C., Heinz, R., & Bemus, B. (1979). *The Wisconsin Case Classification/Staff Deployment Project: Two-year follow-up report*. Madison, WI: Wisconsin Division of Corrections.
- Barnoski, R. (2004). *Assessing risk for re-offense: Validating the Washington state juvenile court assessment*. Retrieved October 29, 2012, from <http://www.wsipp.wa.gov/rptfiles/04-03-1201.pdf>
- Bechtel, K., Lowenkamp, C. T., & Latessa, E. J. (2007). Assessing the risk of re-offending for juvenile offenders using the youthful level of service/case management inventory. *Journal of Offender Rehabilitation*, 45(3/4), 85–108.
- 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.

- Campbell, M. A., French, S., & Gendreau, P. (2007). *Assessing the utility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders*. Ottawa, ON: Public Safety Canada.
- Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior*, 36(6), 567–590.
- Chengalath, G. (2008). Research and Development Division. *Revalidation of Dynamic Risk Instrument*. Prepared for the Arizona Department of Juvenile Corrections.
- Coohey, C., Johnson, K., Renner, L. M., & Easton, S. D. (2013). Actuarial risk assessment in child protective services: Construction methodology and performance criteria. *Children and Youth Services Review*, 35(1), 151–161.
- Dolan, M., & Doyle, M. (2000). Violence risk prediction: Clinical and actuarial measures and the role of the Psychopathy Checklist. *The British Journal of Psychiatry*, 177(4), 303–311.
- Ereth, J., & Healy, T. (1997). *Cook County juvenile female offenders risk assessment study*. Madison, WI: National Council on Crime and Delinquency.
- Farabee, D., Zhang, S., Roberts, R. E. L., & Yang, J. (2010). *COMPAS validation study: Final report*. California Department of Corrections and Rehabilitation. Retrieved from http://www.cdcr.ca.gov/adult_research_branch/Research_Documents/COMPAS_Final_Report_08-11-10.pdf
- Finigan, M. W., Mackin, J. R., Seljan, B. J., & Tarte, J. M. (2003). NPC Research. *Juvenile Crime Prevention Program Evaluation Final Report 2003*. Retrieved from <http://www.npcresearch.com/Files/JCP%20Eval%20Final%20Report%20July%202003.pdf>
- Flores, A. W., Travis, L. F., & Latessa, E. J. (2003). *Case classification for juvenile corrections: An assessment of the Youth Level of Service/Case Management Inventory (YLS/CMI)*, (98-JB-VX-0108). Washington, DC: US Department of Justice.
- Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works. *Criminology*, 34(4), 575–608.
- Gottfredson, D. M. (1987). Prediction and classification in criminal justice decision making. *Crime & Justice*, 9, 1.
- Gottfredson, D., & Snyder, H. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts*. Washington, DC: US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.
- Gottfredson, S. D., & Gottfredson, D. M. (1980). Screening for risk: A comparison of methods. *Criminal Justice and Behavior*, 7(3), 315–330.

- Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency*, 52(1), 178–200.
- Hartney, C., & Silva, F. (2007). *And justice for some: Differential treatment of youth of color in the justice system*. Oakland, CA: National Council on Crime and Delinquency.
- Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior*, 29, 380–396.
- Howard, P. D., & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism: Causal dynamic risk factors in the OASys violence predictor. *Law and Human Behavior*, 37(3), 163–174.
- Johnson, K., Wagner, D., & Matthews, T. (2002, January). *Missouri Juvenile Risk Assessment re-validation report*. Madison, WI: National Council on Crime and Delinquency.
- Kadleck, C., Herz, D., Gallagher, K. W., & Nava, J. (2004). *An evaluation of substance abuse screening and risk assessment in Nebraska's Juvenile Justice System*. Retrieved from http://www.ncc.ne.gov/pdf/juvenile_justice_materials/2004_YLSI_Report.pdf
- Kennealy, P., Hernandez, I., & Skeem, J. (2012, March). *Youth risk assessment: Inter-rater reliability of state corrections staff*. Paper presented at the annual meeting of the APLS Conference, San Juan, Puerto Rico.
- Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, 28(4), 360–374.
- Landis, R. J., & Koch, G. G. (1977a). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- LeCroy, C. W., Krysik, J., & Palumbo, D. (1998). *Empirical validation of the Arizona Risk/Needs Instrument and assessment process* (p. 161). Arizona Supreme Court, Administrative Office of the Courts, Juvenile Justice Services Division.
- Liu, H., Li, G., Cumberland, W. G., & Wu, T. (2005). Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. *Journal of Data Science*, 3(3), 257–278.
- Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections*, 2004, 3–8.
- Monahan, J., & Skeem, J. (2013). *Risk redux: The resurgence of risk assessment in criminal sanctioning*. Presented at the National Association of Sentencing Commissions Annual Conference, University of Minnesota Law School.
- National Council of Juvenile and Family Court Judges. (2002, November). *Graduated sanctions for juvenile offenders: A program model and planning guide*, Ch. 5.

- National Council on Crime and Delinquency. (2004, July). *New Mexico Children, Youth and Families Department prospective validation of the Juvenile Justice Risk Assessment*. Madison, WI: Author.
- National Council on Crime and Delinquency. (2011). *Juvenile detention in Cook County: Future directions*. Madison, WI: Author.
- National Institute of Corrections. (1981). *Model probation and parole management project*. Washington, DC: National Institute of Corrections.
- O'Keefe, M. L., Klebe, K., & Hromas, S. (1998). *Validation of the level of supervision inventory (LSI) for community based offenders in Colorado: Phase II*. Colorado Springs, CO: Colorado Department of Corrections.
- Onifade, E., Davidson, W., Campbell, C., Turke, G., Malinowski, J., & Turner, K. (2008). Predicting recidivism in probationers with the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior*, 35(4), 474–484.
- Orbis Partners. (2007). *Long-term validation of the Youth Assessment and Screening Instrument (YASI) in New York State Juvenile Probation*. Ottawa, Ontario:
<http://criminaljustice.state.ny.us/opca/pdfs/nyltyasifullreport20feb08.pdf>
- Parkerson, G. R., Eugene Broadhead, W., & Tse, C.-K., J. (1993). The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and comorbidity. *Journal of Clinical Epidemiology*, 46(4), 379–393.
- Pope, C. E., Lovell, R., & Hsia, H. M. (2002). *Disproportionate minority confinement: A review of the research literature from 1989 through 2001*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention. Last accessed October 24, 2006, from
http://ojjdp.ncjrs.org/dmc/pdf/dmc89_01.pdf
- Rice, M. E., & Harris, G. T. (1995) Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, 53, 737–748.
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, 29(5), 615–620.
- Rodella, S. (1996). *Exploring reliability in epidemiology and clinical research*. McGill University, Montreal, Canada.
- Royston, P., Moons, K. G., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ*, 338.
- Schmidt, F., Hoge, R. D., & Gomes, L. (2005). Reliability and validity analyses of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior*, 32(3), 329–344.
- Schwalbe, C. S. (2007). Risk assessment for juvenile justice. *Law and Human Behavior*, 31(5), 449–462.

- Schwalbe, C. S. (2008). Strengthening the integration of actuarial risk assessment with clinical judgment in an evidence based practice framework. *Children and Youth Services Review*, 30, 1458–1464.
- Schwalbe, C. S. (2009). *Risk assessment stability: A revalidation study of the Arizona Risk/Needs Assessment Instrument*. Research on Social Work Practice 19.2 (2009): 205. ProQuest Research Library. Web. 9 Oct. 2012.
- Shepherd, S. M., Luebbers, S., & Dolan, M. (2013). Gender and ethnicity in juvenile risk assessment. *Criminal Justice and Behavior*, 40(4), 388–408.
- Short, J., & Sharp, C. (2005). *Disproportionate minority contact in the juvenile justice system*. Child Welfare League of America: Washington, DC. Retrieved from <http://www.cwla.org/programs/juvenilejustice/disproportionate.pdf>
- Silver, E., & Banks, S. (1998). *Calibrating the potency of violence risk classification models: The dispersion index for risk (DIFR)*. Paper presented at the American Society of Criminology conference, Washington, DC.
- Skeem, J., & Eno Loudon, J. (2007). *Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*. Unpublished report prepared for the California Department of Corrections and Rehabilitation. Available at www.cdcr.ca.gov/.../COMPAS_Skeem_EnoLouden_Dec_2007.pdf
- Skeem, J., & Monahan, J. (2011). *Current directions in violence risk assessment*. Virginia Public Law and Legal Theory Research Paper No. 2011-13. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1793193#
- Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. L. (2013a). Risk factors and crime. In F.T. Cullen & P. Wilcox (Eds.). *The Oxford handbook of criminological theory* (pp. 89-111). New York, NY: Oxford University Press.
- Tape, T. G. (n.d.). *Interpreting diagnostic tests: The area under an ROC curve*. Retrieved on June 4, 2012 from <http://gim.unmc.edu/dxtests/ROC3.htm>
- Tarte, J. M., Mackin, J. R., Cox, A., & Furrer, C. J. (2007). NPC Research. *Juvenile Crime Prevention Program 2005–2007, Evaluation Report*. Retrieved from http://www.npcresearch.com/Files/JCP_2005-07_Final_1107.pdf
- Tjaden, C. D. (2006). *Preliminary CRN risk scale validation study*. Prepared for the Georgia Department of Juvenile Justice.
- Uebbersax, J. S. (1987). Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin*, 101(1), 140.
- Van Der Put, C. E., Dekovic, M., Geert, J. J., Stams, M., Van Der Laan, P. H., Hoeve, M., & Van Amelsfort, L. V. (2011). Changes in risk factors during adolescence: Implications for risk assessment. *Criminal Justice and Behavior*, 38, 248–262.

- Van Domburgh, L., Vermeiren, R., & Doreleijers, T. (2008). Screening and assessments. In R. Loeber, H. M. Koot, N. W. Slot, P. H. Van der Laan, & M. Hoeve (Eds.) *Tomorrow's Criminals: The Development of Child Delinquency and Effective Interventions* (pp. 165–178). Hampshire, England: Ashgate.
- Van Voorhis, P., Salisbury, E., Wright, E., & Bauman, A. (2008). *Achieving accurate pictures of risk and identifying gender responsive needs: Two new assessments for women offenders*. University of Cincinnati Center for Criminal Justice Research, National Institute of Corrections, Washington DC.
- Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation*, 72, 22.
- Wagner, D., Ehrlich, J., & Baird, C. (1997). *Wisconsin juvenile offender classification study: Juvenile parole risk assessment validation report*. Madison, WI: National Council on Crime and Delinquency.
- Wiebush, R. G., Baird, C., Krisberg, B., & Onek, D. (1995). Risk assessment and classification for serious, violent, and chronic juvenile offenders. In J. C. Howell, B. Krisberg, J. D. Hawkins, & J. J. Wilson (Eds.), *Sourcebook on Serious, Violent and Chronic Juvenile Offenders* (pp. 171–212). Thousand Oaks, CA: Sage Publications.
- Wiebush, R. Ed. (2002). *Graduated sanctions for juvenile offenders: A program model and planning guide*. Reno, NV: Juvenile Sanctions Center, National Council of Juvenile and Family Court Judges.
- Winokur-Early, K., Hand, G. A., & Blankenship, J. L. (2012). *Validity and reliability of the Florida Positive Achievement Change Tool (PACT) risk and needs assessment instrument: A three-phase evaluation (Validation study, factor analysis, inter-rater reliability)*. Tallahassee, FL: Justice Research Center.
- Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, 136(5), 740.
- Zhang, S. X., Roberts, R. E. L., & Farabee, D. (2011). An analysis of prisoner reentry and parole risk using COMPAS and traditional criminal history measures. *Crime & Delinquency*. doi: 10.1177/0011128711426544
- Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561–577.