Psychological Assessment
2017, Vol. 29, No. 6, 664–678

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000445

Are Risk Assessments Racially Biased?: Field Study of the SAVRY and YLS/CMI in Probation

Rachael T. Perrault and Gina M. Vincent
University of Massachusetts Medical School

Laura S. Guy
Simon Fraser University

Risk assessment instruments are widely used by juvenile probation officers (JPOs) to make case management decisions; however, few studies have investigated whether these instruments maintain their predictive validity when completed by JPOs in the field. Moreover, the validity of these instruments for use with minority groups has been called into question. This field study examined the predictive validity of both the Structured Assessment of Violence Risk in Youth (SAVRY; $n = 383$) and the Youth Level of Service/Case Management Inventory (YLS/CMI; $n = 359$) for reoffending when completed by JPOs. The study also compared Black and White youth to examine the presence of test bias. The SAVRY and YLS/CMI significantly predicted reoffending at the test level, with most of the variance in reoffending accounted for by dynamic risk scales not static scales. The instruments did not differentially predict reoffending as a function of race but Black youth scored higher than White youth on the YLS/CMI scale related to official juvenile history. The implications for use of risk assessments in the field are discussed.

Public Significance Statement
The present study found that two of the most widely used youth risk assessment instruments (SAVRY and YLS/CMI) significantly predicted reoffending when completed by trained juvenile probation officers in the field. While the instruments did not differentially predict reoffending as a function of race there were some significant differences by race on a few items (e.g., community disorganization, substance abuse, history of maltreatment, etc.).

Keywords: risk assessment, SAVRY, YLS/CMI, field validity, racial bias

The use of valid risk assessment instruments to guide programming for youth involved in the juvenile justice (JJ) system has been widely accepted as best practice (National Resource Council, 2013; Singh et al., 2014; Vincent, Guy, & Grisso, 2012) and strongly recommended in U.S. legislation in particular (Juvenile Justice & Delinquency Prevention Act, 2002). For case management planning designed to reduce risk, the type of risk assessment tool used must include factors that are changeable through intervention, broadly referred to as dynamic risk factors (Borum, 2000; Heilbrun, 2010; Hoge, 2002; Vincent, Guy, & Grisso, 2012). Most U.S. state juvenile probation agencies have adopted one of these comprehensive risk assessment instruments to be completed by juvenile probation officers (JPOs) for case planning purposes (Wachter, 2015). Although use of risk instruments to guide JJ decision-making is a very promising strategy, a few concerns have been raised about this approach.

First, adequacy of the field reliability and validity of more comprehensive risk instruments when conducted by JPOs or other JJ staff has been challenged (National Council on Crime and Delinquency [NCCD], 2014), particularly with respect to their ability to rate dynamic risk factors (e.g., parental monitoring, antisocial attitudes) reliably. Second, there has been concern that use of these risk instruments in general may introduce more racial bias into the system. Former Attorney General Holder stated that risk assessment tools used in sentencing decisions ". . . may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society" (Holder, 2014). He noted this was particularly the case for tools that rely on static factors such as level of education, socioeconomic background, or neighborhood. Moreover, scholars have postulated that marginalized groups may score higher on risk assessment tools as a result of their increased exposure to risk and social inequality, rather than a higher propensity for perpetrating crime (Hannah-Moffat & Maurutto, 2010; Tonry, 2014). Clearly, it is important for research to examine not only the field validity of risk instruments when used by JPOs in general, but also the field validity for different racial minority groups specifically (Shepherd, Luebbers, & Dolan, 2013). In light of these issues, the aims of the

Rachael T. Perrault and Gina M. Vincent, Department of Psychiatry, University of Massachusetts Medical School; Laura S. Guy, Department of Psychology Simon Fraser University.

This research was funded by the John D. and Catherine T. MacArthur Foundation as part of the Models for Change Initiative Research Network (PI: Vincent). Gina M. Vincent is a NIDA young investigator (K01 DA026502-01A1).

Correspondence concerning this article should be addressed to Rachael T. Perrault, Department of Psychiatry, University of Massachusetts Medical School, 222 Maple Ave, Shrewsbury, MA 01545. E-mail: rachael.perrault@umassmed.edu

current study were to examine (a) the predictive validity of the two most widely validated risk tools when completed in the field by JPOs, and (b) the possible presence of racial bias for Black youth, a population that is overrepresented in all stages of the U.S. JJ system (Puzzanchera, Adams, & Hockenberry, 2012). The relative magnitude of the predictive validity of static and dynamic risk factors also was investigated overall, as well as a function of race.

Risk Assessment Instruments

The degree to which risk factors are changeable is on a continuum, with some factors highly static, such as history of violence, and some factors highly variable or dynamic, such as substance use. Variable factors are relevant to risk reduction when changes in the factor are associated with changes in the likelihood of reoffending, either through the passage of time (i.e., variable risk marker) or as a result of intervention (i.e., variable risk factor; see Monahan & Skeem, 2014). Risk assessment tools vary considerably regarding the types of factors included, with comprehensive tools including static, dynamic (variable), and/or protective factors and brief tools generally containing only static ones. Protective factors are often described as moderators or "buffers" to risk factors (Jessor, Van Den Bos, Vanderryn, Costa, & Turbin, 1995).

Among researchers who view the chief purpose of risk tools as being to provide an estimate of risk absent any consideration about how to mitigate risk, the inclusion of dynamic risk factors has been claimed to weaken the accuracy of the risk estimate (Baird, 2009). Conversely, much empirical work demonstrates the importance of including dynamic risk factors. For example, in his 2007 meta-analysis of 28 adolescent risk assessment instruments, Schwalbe (2007) noted that brief risk tools had smaller effect sizes than other types of risk instruments. Moreover, several recent studies of adolescent risk tools have demonstrated that dynamic risk factors have incremental predictive validity for reoffending over static risk factors, even when used in the field (Haqanee, Peterson-Badali, & Skilling, 2015; Thompson & McGrath, 2012; Vincent, Chapman, & Cook, 2011).

Identification of Racial Bias

Objections to the practice of using risk assessment instruments on the basis of racial discrimination tend to emphasize concerns about the use of static factors, such as educational and employment history, where individuals of color are likely to score worse than Whites (Holder, 2014). Even more notable are expressed concerns about reliance on factors related to criminal background in light of arguments that "prior criminal history is merely a proxy for skin color" (Harcourt, 2010, p. 4). Dynamic risk factors are not entirely without scrutiny, however, as some postulate these can be experienced differently and have different effects as a function of gender or race (Hannah-Moffat, 2012).

As noted by Skeem and Lowenkamp (2016), examination of the presence of racial discrimination in a risk assessment tool should refer to the standards for identification of test bias, which have been summarized in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement Education, 2014). According to these guidelines, test bias is present when scores have a different meaning (function differently) for different groups of individuals. In this case, the function of a risk assessment is to identify who is most likely to reoffend, so scores should statistically relate to or "predict" reoffending. Mean differences in test or item scores between groups are not indicative of bias if the scores reflect true group differences in reoffending.

In their investigation of racial bias on the Post Conviction Risk Assessment (PCRA) used by the federal system, Skeem and Lowenkamp (2016) conceptualized the issues as an examination of *test bias* (racial differences in predictive validity) versus *disparate impact* (racial differences in mean scores that could result in differential treatment). Following multiple analyses, they determined there was not a significant difference between White and Black adult offenders in the PCRA's predictive validity, but Black offenders did score significantly higher on the tool, largely as a function of the criminal history items.

Two of the most well validated tools for assessing reoffending among adolescents are the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2006) and the Structured Assessment of Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2006). These instruments use different approaches for determinations of one's overall risk level. The SAVRY is a structured professional judgment (SPJ) tool where evaluators make an overall risk rating that incorporates both structured ratings of prescribed risk factors and their judgment, and the YLS/CMI is an actuarial tool. A small body of research has accumulated examining these tools in field settings, some of which has investigated racial differences.

Three evaluations of the SAVRY's predictive validity when completed by JJ personnel have been conducted. In two, good to excellent interrater agreement was observed for most items and the risk ratings (Hilterman, Nicholls, & van Nieuwenhuizen, 2014; Vincent, Guy, Fusco, & Gershenson, 2012). One of these also examined predictive validity and determined the SPJ risk ratings significantly predicted both general (AUC = .73) and violent (AUC = .68) reoffending (Hilterman et al., 2014). In a third study, Chapman, Desai, Falzer, and Borum (2006) examined racial differences and found the predictive validity of a total risk score (calculated for research purposes) based on ratings made by detention staff did not differ for White, Black, and Hispanic youth for violent rearrests over 5 years. Further, there were no significant racial differences in staff members' SPJ risk ratings and the only item on which Blacks scored significantly higher than others related to community disorganization (Chapman et al., 2006).

There have been many field studies of the YLS/CMI's interrater reliability and predictive validity (e.g., Olver, Stockdale, & Wormith, 2009; Schmidt, Hoge, & Gomes, 2005; Schmidt, Sinclair, & Thomasdottir, 2011). Several have demonstrated that the YLS/CMI significantly predicts recidivism for minority youth as well as for White youth (Olver, Stockdale, & Wormith, 2014) with some indication that the minority youth may have significantly higher scores (Bechtel, Lowenkamp, & Latessa, 2007). However, the vast majority of studies have compared Canadian or Australian Indigenous to non-Indigenous youth (Jung & Rawana, 1999; Schmidt, Campbell, & Houdling, 2010; Shepherd, Luebbers, Ogloff, Fullam, & Dolan, 2014) or Whites to non-Whites (Barnes et al., 2016). One of the only field studies reporting predictive validity for Black youth indicated YLS/CMI scores were not significantly correlated with their reoffending (Onifade et al., 2008).

More broadly, meta-analyses of the predictive validity of adolescent risk assessment tools indicate that although higher risk scores have been observed among minority groups, the tools are in fact valid predictors of reoffending among minority groups (Gutierrez, Wilson, Rugge, & Bonta, 2013; Olver et al., 2014). The empirical base is small, however, in that few primary studies have directly compared risk scores among Whites and Blacks specifically.

The Current Study

Efforts to understand the degree to which racial bias may be present in risk assessment measures should give more weight to data from the field, as opposed to laboratory-like settings, given concerns about disparate impact and bias. In the current study, we examined field validity and racial differences using the SAVRY (Borum et al., 2006) and the YLS/CMI (Hoge & Andrews, 2006). The goals were to conduct a comprehensive examination of the instruments' overall predictive validity when completed by JPOs and the presence of racial bias for Black compared with White youth. Consistent with Skeem and Lowenkamp (2016), we examined racial differences with respect to both disparate impact (mean item, score, risk level differences) and test bias, defining test bias as the presence of a significant interaction between scores and race in the prediction of reoffending. We used data from the Risk/Needs Assessment in Juvenile Probation: Implementation Study (RNAJP), a multisite (six probation offices in two states) prospective study in which validated risk tools were implemented in juvenile probation (Vincent, Paiva-Salisbury, Cook, Guy, & Perrault, 2012). One state implemented the SAVRY and one state implemented the YLS/CMI following comprehensive and standardized JPO training.

We had two primary hypotheses. First, we expected the SAVRY and YLS/CMI would have significant predictive validity for both violent and nonviolent reoffending at the total score and risk rating levels. Second, we expected the assessments would not have differential predictive validity (test bias) for Black versus White youth but may have mean differences in overall risk scores (disparate impact). Lastly, we examined the relative contribution of dynamic risk factors anticipating these would significantly predict reoffending above and beyond static ones. We also anticipated disparate impact would be most prominent for static risk factors.

Method

Sample

The three offices that implemented the YLS/CMI were located in Pennsylvania (PA) and the offices that implemented the SAVRY were located in Louisiana (LA). In LA, the SAVRY was administered postadjudication to all youth who entered the system in 2009, whereas in PA the YLS/CMI was administered preadjudication to most youth referred to court (and postadjudication for about one third) from mid-2009 to mid-2010.

The initial SAVRY sample comprised 452 youth. Fifty-two youth were excluded because they were not administered the SAVRY, 12 were excluded because they were in a placement the entire follow-up period, and five were excluded because they reoffended prior to their first SAVRY administration. The final SAVRY sample (n = 383) was primarily male (72.6%) and Black (78.9%) with a mean age of 15.20 years (SD = 1.48). Of the 383 youth, six were excluded (they did not identify as White or Black), leaving the group samples at 302 Black (80.1%) and 75 White (19.9%) youth for comparative race analyses.

The initial YLS sample comprised 406 youth. Three youth were excluded because they were in a placement for the entire follow-up period, 16 were excluded because they reoffended prior to their first YLS/CMI administration, and 28 were excluded because their cases were expunged, making rearrest data unavailable. The final YLS/CMI sample (n = 359) was primarily male (74.1%) and White (64.6%) with a mean age of 15.52 years (SD = 1.60). Forty-five were excluded because they identified as biracial or other; leaving 232 (73.9%) White and 82 (26.1%) Black youth for race analyses. There were no significant differences in gender or age between White and Black youth in either the YLS/CMI or the SAVRY samples.

Institutional approval for this research project was obtained from the University of Massachusetts Medical School, following a memorandum of agreement with the appropriate authority from every probation office.

Measures

Structured Assessment of Violence Risk in Youth (SAVRY). The SAVRY (Borum et al., 2006) uses structured professional judgment (SPJ) whereby a global summary risk rating (SRR) is based on the evaluator's appraisal of the relevance of existing risk and protective factors to the individual and consideration of any additional idiosyncratic factors (Borum, 2000). The risk items are divided into three domains. The historical domain includes 10 static items that are rated based on the youth's past behavior and experiences. The six social/contextual domain items are dynamic and relate to interpersonal relationships and the youth's environment. The individual/clinical domain includes eight dynamic risk factors that focus on the youth's attitudes and psychological and behavioral functioning. The 24 risk items are rated as low, moderate, or high and the six protective factors are rated as absent or present based on the descriptions in the SAVRY manual. Meta-analyses have reported good predictive validity for the total risk score, mean r_w = .32 to .30 and median AUC = 0.71 (Olver et al., 2009; Singh, Grann, & Fazel, 2011, respectively) and SRR (mean weighted AUC values of 0.71, k = 4; Guy, 2008).

Analyses were conducted for both the SRR and a "total risk score" (used only for research purposes), which was created by assigning numerical values to item ratings (0 = low, 1 = moderate, 2 = high) and summing to yield a score with a possible range of 0 to 48. Protective factor items were assigned values (0 = absent and 1 = present) and summed into a "total protective score." Interrater reliability (IRR) for this sample was evaluated using independent ratings of JPOs and trained research assistants (RAs) at each site. The RAs observed JPOs' interviews and reviewed the same file information. Based on 80 random cases of adjudicated youths across the three sites, one way random, single measurement, absolute agreement type intraclass correlation coefficients, $ICC_{(A,\ 1)}$, (McGraw & Wong, 1996) were 0.71 for the

SRR, 0.86 for the total risk score, and ranged from 0.67 to 0.86 for the domain scores (Vincent, Guy, Fusco, et al., 2012).

Youth Level of Service/Case Management Inventory (YLS/CMI). The YLS/CMI (Hoge & Andrews, 2006) uses an actuarial approach in which item scores are summed to yield a total score that corresponds to risk level classifications. This tool also permits the evaluator to override the score based risk level using his or her professional judgment. Its 42 dynamic and static risk factors are divided into eight domains that have been identified as the most predictive of reoffending among youth: prior and current offenses/dispositions (four items), family circumstances/parenting (six items), education/employment (seven items), peer relations (four items), substance abuse (four items), leisure/recreation (three items), personality/behavior (seven items), and attitudes/orientation (five items). Each item on the YLS/CMI is coded as present or absent and summed for a total score ranging from 0 to 42. The overall risk rating is score-based: low (0–8), moderate (9–22), high (23–34), and very high (35–42). Each domain also has cut scores to indicate a risk level within that domain. Professional judgment ratings were not used in this study because this aspect of the tool's administration was not implemented into practice until much later. Good levels of predictive validity for violent and nonviolent reoffending found for the SAVRY also have been reported for the YLS/CMI (Olver et al., 2009).

Similar to the SAVRY, IRR for the YLS/CMI in this sample was evaluated using independent ratings of JPOs and trained RAs at each site. The RAs observed the interviews and reviewed the same file information as the JPO before making their own ratings. IRR for the YLS/CMI in this sample was evaluated using 61 independent ratings of randomly selected cases. IRR was calculated using a one way random, single measurement, absolute agreement ICC model. The $ICC_{(A,\ 1)}$ was 0.84 for the YLS/CMI total score, 0.71 for overall risk ratings, and ranged from 0.55 (personality/behavior) to .88 (substance abuse) for subscales (Guy & Vincent, 2011).

Data Collection Procedures

All offices followed a standardized model of implementation of the SAVRY and YLS/CMI that was completed with the assistance of the authors. JPOs underwent a lengthy training process in which they attended a 2-day SAVRY or YLS/CMI training workshop and completed three additional posttraining practice cases over a 2-month period. JPOs received feedback after each posttraining practice case to improve their proficiency. They were instructed to rate the tool after reviewing file information, conducting semi-structured interviews (using scripts provided by the authors) with the youth, parents(s) or guardians(s), and gathering information from collateral contacts. JPOs in both states were trained to incorporate the assessments into their disposition recommendations and into the services selected for case planning. There were cross-site differences in the decision point where the risk tools were administered, such that the YLS/CMI was primarily administered pre-adjudication at probation intake and the SAVRY was always administered postadjudication. Both tools were incorporated into electronic case management systems at each site. For a more detailed description about the implementation procedures used in the current study see (Vincent, Paiva-Salisbury, et al., 2012).

Recidivism was defined as any new petition (i.e., formal filing of charges) after the date of the assessment. Researchers obtained recidivism data from juvenile and adult court records for an average follow-up of 18.29 months (SD = 3.09 months; range 9.13 to 25.43 months) for the SAVRY sample and 16.51 months (SD = 3.50 months; range 3.73 to 25.43 months) for the YLS/CMI sample. There were no significant differences in follow-up time by race for either the SAVRY or YLS/CMI samples. Petitions were categorized as: *violent* (offenses related to actual or threatened harm to persons, including sex offenses), *nonviolent* (all other offenses except violations and status offenses), and *any* (violent and nonviolent petitions). Sex offenses were not examined separately because base rates were extremely low. Youth who reoffended both violently and nonviolently would be represented in all three categories. Time at-risk was calculated separately for each offense category, after accounting for time spent in any facilities, using the follow-up date as the end point for youth who did not reoffend.

Data Analysis Procedures

We examined the overall predictive validity of SAVRY and YLS/CMI total scores and risk levels using multiple statistical approaches. First, we used separate 3 (risk level) × 2 (recidivism yes/no) chi-square statistics to investigate the association between the SAVRY and YLS/CMI risk levels and the three forms of reoffending using a Bonferroni adjusted alpha level of .017 to account for multiple tests. A limitation with chi-squares is that they do not take individuals' time at-risk into account, which was variable for youth in this study. Thus, we also conducted Cox regression, a semiparametric survival analysis that accounts for variable time to reoffending by examining the proportion of cases that are surviving the time to a specific event (reoffending), based upon the values of given covariates. Time at-risk was defined as the number of days between the risk assessment administration and end of follow-up data collection at the respective site for those who did not reoffend.

As another method for quantifying the strength of the prediction of risk instruments for reoffending, we also conducted receiver operating characteristic (ROC) curves. ROC analysis generates area under the curve (AUC) values. AUCs of 0 indicate perfect negative prediction, 0.5 indicate chance, and 1 indicates perfect positive prediction. AUC is the probability that a randomly selected recidivist will have a higher score than a randomly selected nonrecidivist. In general, AUC values of .56 are described as small, .64 as moderate, and .71 as large (Rice & Harris, 2005). ROC analysis is recommended because it is not tied to base rates (Mossman & Somoza, 1991; Rice & Harris, 2005).

Next, we used two approaches to examine the presence of test bias as a function of race. We performed hierarchical Cox regressions to test for differential predictive validity by including interaction terms between race and categorical risk ratings (or total scores) on the SAVRY or YLS/CMI. Similar to procedures used by Skeem and Lowenkamp (2016), each hierarchical model contained the instrument's risk rating (or total score) in the first block, race in the second block, and an interaction term between the risk rating (or total score) and race in the third block. Separate models were conducted for each type of reoffending. A significant interaction term would indicate that scores on the instrument had a

different association with the outcome for reoffending among White youth than among Black youth. As a secondary indicator of differential predictive validity, we also compared AUCs for Black versus White youth using Stata version 14.1 (StataCorp, 2015) and a Bonferroni adjusted alpha level of .0055.

Next, we used hierarchical Cox regressions to explore the relative contribution of dynamic domain scores on both instruments over their respective static scales. The static scale of each instrument (SAVRY: Historical Scale; YLS/CMI: Prior/Current Offenses) was entered at the first block, and the dynamic scales (two for the SAVRY and seven for the YLS/CMI) were entered at the second block using the backward elimination procedure. This procedure retained only the dynamic scales that were significant predictors of reoffending while also indicating whether the scales had incremental predictive validity over the static scale. Bonferroni adjusted alpha levels of .017 were used to account for the multiple analyses within each instrument. We investigated the presence of test bias on static scales and any of the significant dynamic scales by performing Cox regressions for only these scales and including the respective interaction terms.

Finally, we examined mean differences in scores by race (disparate impact) at both the item and test levels. First, we conducted one-way multivariate analyses of variance (MANOVA) to examine racial differences in item ratings on the SAVRY and in scale scores on the YLS/CMI. MANOVA was used to conduct multiple comparisons within one analysis. The YLS/CMI was examined at the scale score level rather than the item level. Because of the large number of items (42) and their dichotomous nature, we anticipated differences would be more apparent at the scale-level. In order to quantify the magnitude of differences categorically, we also report the odds of Black versus White youth scoring moderate to high (as opposed to low) on the YLS/CMI domains and SAVRY item ratings.

Results

Table 1 provides descriptive statistics for the SAVRY and YLS/CMI total and domain scores. The percentage of youths classified at each SAVRY risk level was: 40.6% (n = 155) low risk, 46.6% (n = 178) moderate risk, and 12.8% (n = 49) high risk. There were no significant differences in risk ratings, $\chi^2(2, 376) = 3.25$, $p = .197$, Cramer's V = .09, or in total scores, $t(374) = .51$, $p = .609$, CI [−1.49, 2.54], $d = 0.06$ between White ($M = 13.33$, $SD = 8.74$; and Black youth ($M = 12.81$, $SD = 7.75$). For the YLS/CMI, the percentage of youth at each risk level was: 47.9% (n = 172) low risk, 46.0% (n = 165) moderate risk, and 6.1% (n = 22) high risk. No youth scored in the very high risk range and there also were no significant differences in risk ratings, $\chi^2(2, 314) = 2.57$, $p = .276$, Cramer's V = .09) or in total scores, $t(312) = .16$, $p = .872$, CI [−1.57, 1.85], $d = 0.02$ between White ($M = 10.24$, $SD = 6.95$) and Black youth ($M = 10.10$, $SD = 6.18$).

Overall, 37.9% (n = 145) of youth in the SAVRY sample and 21.4% (n = 77) of youth in the YLS/CMI sample received a new petition. Due to the nature of the sample (primarily preadjudication), base rates for violent recidivism in particular were particularly low for the YLS/CMI sample at 5.8% (n = 21) compared with 14.9% (n = 57) for the SAVRY sample. As a result of the low base rates, the only analyses conducted on violent recidivism for the YLS/CMI sample were the ROCs. There were no significant differences in base rates for any form of reoffending between White and Black youth in either the SAVRY or YLS/CMI samples.

Table 1
Descriptive Statistics for the SAVRY and YLS/CMI Total Scores and Domains

Domains	*M* (*SD*)	Range
SAVRY (n = 383)		
Total score (0 to 48)	12.87 (7.90)	0–38
Historical (0 to 20)	4.98 (3.31)	0–19
Social/contextual (0 to 12)	3.24 (2.23)	0–10
Individual/clinical (0 to 12)	4.67 (3.51)	0–15
Protective (0 to 6)	3.64 (2.17)	0–6
YLS/CMI (n = 359)		
Total score (0 to 42)	10.07 (6.76)	0–30
Prior and current offenses/dispositions (0 to 5)	0.28 (0.67)	0–4
Family circumstances/parenting (0 to 6)	1.67 (1.57)	0–6
Education/employment (0 to 7)	1.89 (1.58)	0–7
Peer relations (0 to 4)	1.62 (1.33)	0–4
Substance abuse (0 to 5)	1.20 (1.47)	0–5
Leisure/recreation (0 to 3)	1.23 (1.03)	0–3
Personality/behavior (0 to 7)	1.56 (1.66)	0–7
Attitudes/orientation (0 to 5)	0.61 (1.04)	0–5

Note. SAVRY = Structured Assessment of Violence Risk in Youth; YLS/CMI = Youth Level of Service/Case Management Inventory; M = mean; SD = standard deviation.

Overall Predictive Validity of SAVRY and YLS/CMI

Chi-square results indicated, as expected, there was a statistically significant association between risk level and any or nonviolent recidivism for both the SAVRY, $\chi^2(2, 382) = 13.25$, $p = .001$; $\chi^2(2, 382) = 9.17$, $p = .010$, respectively; and YLS/CMI, $\chi^2(2, 359) = 17.50$, $p < .001$; $\chi^2(2, 359) = 16.53$, $p < .001$, respectively; such that higher risk youth were significantly more likely to reoffend than lower risk youth. For example, 32.9% (n = 51) of low risk youth (n = 155) on the SAVRY, were petitioned for any new offense, compared to 61.2% (n = 30) of high risk youth (n = 49). Higher risk youth also were significantly more likely to reoffend violently than lower risk youth on the SAVRY, $\chi^2(2, 382) = 11.29$, $p = .004$.

Cox regressions were conducted to examine the association between total scores and time to any and nonviolent recidivism for the YLS/CMI and for all three recidivism types for the SAVRY. Total scores on both instruments significantly predicted any new petitions (YLS/CMI - Exp(B) = 1.06, CI [1.03, 1.09], $p < .001$; SAVRY - Exp(B) = 1.05, CI [1.03, 1.07], $p < .001$) and nonviolent petitions (YLS/CMI - Exp(B) = 1.06, CI [1.03, 1.10], $p < .001$; SAVRY - Exp(B) = 1.04, CI [1.02, 1.06], $p < .001$). SAVRY total scores also predicted violent petitions, Exp(B) = 1.08, CI [1.05, 1.11], $p < .001$. Risk ratings on both instruments also significantly predicted any new (YLS - Exp(B) = 1.90, CI [1.33, 2.67], $p < .001$; SAVRY Exp(B) = 1.55, CI [1.22, 1.93], $p < .001$) and nonviolent petitions (YLS - Exp(B) = 2.09, CI [1.42, 3.05], $p < .001$; SAVRY - Exp(B) = 1.45, CI [1.13, 1.87], $p = .004$). The SAVRY risk rating also significantly predicted violent petitions, Exp(B) = 1.94, CI [1.34, 2.81], $p < .001$).

As a secondary measure of the strength of association between instruments and reoffending, Table 2 provides the AUCs for each type of recidivism for both instruments.[1] Similar to Cox regression results, AUC values for the SAVRY and YLS/CMI total scores were significant for all recidivism types, except for violence for the YLS/CMI. The risk ratings for the SAVRY and YLS/CMI showed the same pattern of results, but the SAVRY SRR had lower AUCs than the total risk scores. AUCs for the SAVRY Protective Factor Score ranged from .43 to .45 for each type of recidivism outcome and were not statistically significant.

In order to explore the relatively low AUCs for the SAVRY risk rating, we conducted post hoc tests to examine the presence of site-level effects in JPOs' ability to make the risk rating using a Bonferroni adjusted alpha level of .0055. The SRR's AUCs for all types of reoffending were poor in one site (AUCs = .48–.50) and better in the other two sites (AUCs = .59–.67), but only achieved significance in one site, suggesting the presence of a site effect.[2]

Racial Differences in Overall Predictive Validity of SAVRY and YLS/CMI (Test Bias)

Results of the hierarchical Cox regression models including interaction terms for race are provided in Table 3 for each type of recidivism. The SAVRY risk rating was a significant predictor of all three forms of recidivism even after race was added to the model, and race was not significant. After inclusion of the interaction term, the SAVRY risk rating was still significant for only any reoffending. The interaction term was not significant for any of the regression models. Repeating this set of analyses for SAVRY total risk scores instead of the SRR resulted in the same pattern of findings.

The set of Cox regressions for the YLS/CMI also indicated that the risk ratings alone significantly predicted any and nonviolent reoffending. Contrary to the SAVRY results, when race was added to the model, both the risk ratings and race were predictive of any and nonviolent reoffending (see Table 3). However, the interaction terms were not significant in either model. Repeating this set of analyses for YLS/CMI total risk scores resulted in the same findings except that race was no longer significant for nonviolent reoffending. Comparison of the AUCs for total scores on the SAVRY or YLS/CMI between White versus Black youth (previously shown in Table 2) also indicated that there were no significant racial differences in the predictive validity of these instruments. However, AUCs could not be compared for violent reoffending because the base rates for White youth and for Black youth were too low.

Incremental Predictive Validity of the SAVRY and YLS/CMI Domains

Starting with the SAVRY, the historical scale alone was a significant predictor of all three forms of reoffending; however, once the dynamic scales were entered it was no longer significant (see Table 4). Thus, the individual/clinical scale (the only dynamic scale retained), accounted for all the variance in reoffending that was explained by the historical scale and increased the predictive accuracy of the SAVRY. Odd ratios (Exp(B)), ranged from 1.10 for nonviolent reoffending to 1.16 for violent reoffending, indicating a 1-point increase in the individual/clinical scale score resulted in an increased likelihood of recidivism by 10% to 16%.

To examine the presence of racial differences on these scales, we conducted separate Cox regressions with the scale (historical or individual/clinical) entered at the first block, race at the second block, and the interaction term in the third block. The models yielded similar results, indicating no interaction between race or the individual/clinical and historical scales for any, nonviolent, or violent reoffending.

Repeating these procedures with the YLS/CMI scales indicated that the prior/current offenses scale alone was not a significant predictor for any or nonviolent reoffending (see Table 5). Entering the dynamic scales at the second block indicated the dynamic scales retained in the model (those with the strongest predictive validity) differed as a function of the type of reoffending. As shown in Table 5, only the peer relations and substance abuse scales were significant predictors of nonviolent reoffending; whereas the peer relations, substance abuse, and personality/behavior were significant predictors of any reoffending. Hierarchical models for each of these four scales (includes the prior/current offenses scale) indicated none of these YLS/CMI domains had significant interactions with race with respect to predicting any or nonviolent reoffending.

Mean Differences in Risk Scores on the SAVRY and YLS/CMI by Race (Disparate Impact)

As previously noted, there were no significant racial differences in scores or overall risk ratings at the test level for either instrument. Here we conducted additional tests to examine differences at the domain and item levels. First, the MANOVA examining racial differences on YLS/CMI domain scores was significant, $F(8, 305) = 4.30$, $p < .001$, Wilk's $\Lambda = 0.90$, $\eta_p^2 = .10$. Examination of the univariate main effects indicated Blacks scored significantly higher than Whites on the prior/current offense scale, $F(1, 314) = 12.06$, $p = .001$, $d = -0.40$; and Whites scored significantly higher on the substance abuse scale, $F(1, 314) = 5.85$, $p = .016$, $d = 0.34$.

For the SAVRY, the MANOVA comparing race groups on all risk item ratings (low, moderate, or high) also was significant; $F(24, 310) = 4.26$, $p < .001$, Wilk's $\Lambda = 0.75$, $\eta_p^2 = .25$. Examination of the univariate main effects indicated four items significantly differed by race. White youth were significantly more likely than Black youth to have a history of suicide attempts, $F(1, 335) = 8.10$, $p = .005$, $d = 0.33$, be exposed to violence in the home, $F(1, 335) = 10.68$, $p = .001$, $d = 0.42$; and to have substance abuse difficulties, $F(1, 335) = 14.06$, $p < .001$, $d = 0.48$. Conversely, Black youth were significantly more likely to live in a disorganized community, $F(1, 335) = 43.17$, $p < .001$, $d = 0.98$. A separate MANOVA conducted to examine differences on protective factors (rated absent or present) also was significant, $F(6, 359) = 2.75$, $p = .013$, Wilk's $\Lambda = 0.96$, $\eta_p^2 = .04$. One

[1] Because youths had different lengths of time at which they had opportunity to reoffend, a fixed time period variable of 12 months was created. This variable continued to account for time in placement. ROC analyses with the fixed time variable observed fairly similar results to the variable time period in that the *p*-values changed very little.

[2] Due to the presence of a site effect, we reran all of the ROC analyses excluding the poor site. Because this did not lead to considerably large differences in AUC values, all three probation sites are included in the SAVRY analyses.

Table 2
Predictive Validity of the SAVRY and YLS/CMI Total Risk Score and Risk Rating by Recidivism Type and Race

Recidivism type	Total risk score						Risk rating					
	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
SAVRY	Total sample (n = 383)		White (n = 75)		Black (n = 302)		Total sample (n = 383)		White (n = 75)		Black (n = 302)	
Nonviolent	.60** (.03)	[.55, .66]	.60 (.07)	[.46, .74]	.60** (.03)	[.53, .67]	.56* (.03)	[.50, .62]	.56 (.07)	[.43, .69]	.55 (.03)	[.49, .62]
Violent	.69*** (.04)	[.62, .76]	—[a]	—	—	—	.63** (.04)	[.55, .70]	—	—	—	—
Any	.62*** (.03)	[.57, .68]	.65* (.07)	[.52, .78]	.61** (.03)	[.55, .68]	.58* (.03)	[.52, .64]	.61 (.07)	[.49, .74]	.56 (.03)	[.50, .62]
YLS/CMI	Total sample (n = 359)		White (n = 232)		Black (n = 82)		Total sample (n = 359)		White (n = 232)		Black (n = 82)	
Nonviolent	.65*** (.04)	[.58, .72]	.64** (.05)	[.54, .73]	.74** (.05)	[.64, .84]	.64** (.04)	[.57, .71]	.62* (.04)	[.53, .70]	.73** (.05)	[.64, .83]
Violent	.57 (.06)	[.46, .68]	—	—	—	—	.51 (.06)	[.39, .63]	—	—	—	—
Any	.65*** (.03)	[.59, .72]	.64** (.04)	[.55, .73]	.74** (.05)	[.64, .85]	.63** (.03)	[.56, .69]	.62* (.04)	[.54, .70]	.70** (.05)	[.60, .81]

Note. SAVRY = Structured Assessment of Violence Risk in Youth; YLS/CMI = Youth Level of Service/Case Management Inventory; AUC = area under the curve; *SE* = standard error; CI = confidence interval.
[a] Analyses were not conducted for violent recidivism because the base rates for each group, by race, were too low.
* $p < .05$. ** $p < .01$. *** $p < .001$.

protective item significantly differed by race; Black youth were more likely to be rated as resilient than White youth, $F(1, 364) = 8.43$, $p = .004$, $d = 0.01$. It is important to note that previous study of the SAVRY's item reliability among these JPOs indicated the $ICC_{(A,\ 1)}$ for the four risk items with significant differences were good to excellent (0.63 to 0.84); however, reliability for the resilient personality traits item was poor, 0.49.

To report differences between race groups on the YLS/CMI domains categorically (which is more consistent with its use in practice), odds ratios were conducted comparing youth scoring low to youth scoring moderate to high within each domain according to the domain's respective cut-score. As seen in Table 6, Black youth (30.5%) were 2.64 times as likely as White youth (14.2%) to score moderate to high on the prior/current offense scale. The odds ratios on the substance abuse and other dynamic scales were not significant. Odds ratios for the SAVRY items (comparing youth rated low with those rated moderate to high) were consistent with the MANOVA except for one item, childhood history of maltreatment, where Black youth were 0.51 times as likely to have been physically abused or neglected as a child than White youth, $\chi^2(1, 375) = 6.41$, $p = .02$ (see Table 6).

Discussion

The results of this research, conducted in two different states, support and expand the existing literature on the predictive validity of the SAVRY and the YLS/CMI when administered in the field by trained JPOs to make real-world case management decisions. Despite concerns about the field validity of risk instruments that incorporate dynamic risk factors (NCCD, 2014), this study adds to the mounting evidence (Haqanee et al., 2015; Thompson & McGrath, 2012; Vincent et al., 2011) that these domains were more predictive of future offending than static risk factors in many respects. Moreover, this study should provide additional confidence in use of these tools with Black youth involved in the JJ system. Total scores for both instruments, and categorical risk ratings using structured professional judgment on the SAVRY, were not significantly higher for Black youth than White youth. Mean differences suggestive of a disparate impact were limited to items based on official offense history and community disorganization. More importantly, there was no evidence that either of these risk tools, as a whole, predicted reoffending differently as a function of race.

Do the SAVRY and YLS/CMI Maintain Predictive Validity When Used in the Field by Probation Officers?

Both the SAVRY and YLS/CMI total scores and categorical risk ratings significantly predicted reoffending, which suggests the instruments maintain predictive validity when completed by trained JPOs. The field validity question becomes paramount in light of concerns about the use of instruments that incorporate so many dynamic risk factors due to the ability of justice personnel to rate these items accurately and reliably (Baird, 2009; NCCD, 2014). However, there are a few caveats about the predictive validity of the SAVRY and YLS/CMI.

Before describing the caveats, it is important to note that it was not the intention of this study to compare the accuracy of these risk tools. The tools were completed by different JPOs in different

Table 3
Cox Regressions Examining Interaction Between Race and Risk Ratings on the SAVRY and YLS/CMI

	SAVRY				YLS/CMI			
Recidivism type	β (*SE*)	Exp(B)	95% CI	χ^2 (*df*)	β (*SE*)	Exp(B)	95% CI	χ^2 (*df*)
Any								
Block 1								
Overall risk rating	.42 (.12)	1.52**	[1.20, 1.94]	11.83 (1)**	.72 (.19)	2.06***	[1.42, 2.98]	14.97 (1)***
Block 2—Chi-square change				Δ.01 (2)**				Δ5.22 (1)*
Overall risk rating	.42 (.12)	1.52**	[1.20, 1.94]		.77 (.19)	2.18***	[1.49, 3.18]	
Race	−.02 (.21)	.98	[.65, 1.48]		.60 (.26)	1.83*	[1.11, 3.02]	
Block 3—Chi-square change				Δ.94 (1)				Δ1.74 (1)
Overall risk rating	.65 (.26)	1.91*	[1.14, 3.20]		.63 (.23)	1.87**	[1.19, 2.92]	
Race	.22 (.34)	1.25	[.64, 2.41]		.14 (.45)	1.15	[.48, 2.77]	
Overall risk rating*Race	−.29 (.30)	.75	[.42, 1.34]		.58 (.44)	1.79	[.75, 4.28]	
Nonviolent								
Block 1								
Overall risk rating	.36 (.13)	1.43**	[1.11, 1.84]	7.71 (1)**	.82 (.21)	2.27***	[1.51, 3.42]	116.01 (1)***
Block 2—Chi-square change				Δ.03 (1)				Δ4.74 (1)*
Overall risk rating	.36 (.13)	1.43**	[1.11, 1.84]		.89 (.22)	2.43***	[1.59, 3.70]	
Race	.04 (.22)	1.04	[.67, 1.61]		.64 (.28)	1.89*	[1.09, 3.28]	
Block 3—Chi-square change				Δ.002 (1)				Δ2.76 (1)
Overall risk rating	.37 (.28)	1.44	[.84, 2.48]		.67 (.25)	1.96**	[1.19, 3.22]	
Race	.05 (.34)	1.05	[.54, 2.04]		−.04 (.52)	.96	[.35, 2.64]	
Overall risk rating*Race	−.02 (.31)	.99	[.53, 1.82]		.82 (.50)	2.26	[.86, 5.97]	
Violent								
Block 1					—[a]	—	—	—
Overall risk rating	.65 (.19)	1.91**	[1.31, 2.78]	11.84 (1)**				
Block 2—Chi-square change				Δ1.86 (1)	—	—	—	—
Overall risk rating	.64 (.19)	1.90**	[1.30, 2.77]					
Race	.52 (.40)	1.68	[.76, 3.70]					
Block 3—Chi-square change				Δ.50 (1)	—	—	—	—
Overall risk rating	.97 (.51)	2.64	[.98, 7.13]					
Race	.92 (.74)	2.50	[.59, 10.58]					
Overall risk rating*Race	−.38 (.55)	.68	[.23, 2.00]					

Note. SAVRY = Structured Assessment of Violence Risk in Youth; YLS/CMI = Youth Level of Service/Case Management Inventory; *SE* = standard error; CI = confidence interval; *df* = degrees of freedom.
[a] Analyses were not conducted for violent recidivism with the YLS/CMI because the base rates were too low.
* $p < .05$. ** $p < .01$. *** $p < .001$.

states with different policies and practices and differential rates of official reoffending. Additionally, the YLS/CMI sample primarily was generated from a preadjudication population. Thus, we would expect this to be a lower risk sample with lower base rates of reoffending than traditional probation samples (like the SAVRY sample) that contain youth who have been adjudicated and sentenced. Indeed, the base rate of violence in the YLS/CMI sample was only 5.8%, making the null finding for its ability to predict violent reoffending simply inconclusive. There are far too many confounds to compare the performance of these tools in this study.

The first caveat is that the AUCs for the total scores for both instruments were below the bar of 0.71 (Rice & Harris, 2005). However, these AUCs were similar to the average predictive accuracy reported in Schwalbe's (2007) meta-analysis of risk assessment instruments (average AUC = 0.64), which included primarily field studies. On another positive note, the SAVRY's total score AUC for violent reoffending was only slightly below the bar in this study (AUC = 0.69) and was closer to the average accuracy reported for the SAVRY in other meta-analyses that did not include many field studies (Singh et al., 2011).

A second caveat is that the risk rating for the SAVRY performed particularly poorly for the reoffense categories indicative of general delinquency, even within one site where the AUC for violent reoffending suggested these risk ratings were performing well (AUC = 0.67). Contrary to findings from the SPJ literature more broadly (Guy, Douglas, & Hart, 2015), SPJ risk ratings did not outperform the total scores. There are a few possible explanations for the relatively low predictive accuracy of the risk ratings in the field. First, it is important to remember that most meta-analyses and SAVRY studies have examined the predictive accuracy of total risk scores and not the SPJ risk ratings, which is how the instrument is used in practice. An earlier study with JJ personnel found that, similar to the current investigation, the detention staff member's SRRs did not outperform SAVRY total scores (Vincent et al., 2011). Second, making the overall risk decision on the SAVRY requires considering both the presence of each risk and protective factor, as well as idiosyncratic factors, and the relevance of these factors to an individual youth's risk for delinquency or violence. It may be that the subjectivity of the SPJ approach is simply reducing validity in the field in a way that it would not do in the lab,

Table 4
Hierarchical Cox Regression Comparing Predictive Validity of SAVRY Domains

Recidivism type	β (*SE*)	Exp(B) [95% CI]	χ^2 (*df*)
Any			
Block 1 Historical	.08 (.02)	1.08 [1.03, 1.13]***	11.60 (1)*
Block 2—Chi-square change			**Δ**12.35 (1)**
Historical	.02 (.03)	1.02 [.96, 1.08]	
Individual	.11 (.03)	1.11 [1.05, 1.18]***	
Nonviolent			
Block 1 Historical	.06 (.02)	1.06 [1.01, 1.12]**	5.70 (1)*
Block 2—Chi-square change			**Δ**8.35(1)**
Historical	.01 (.03)	1.00 [.94, 1.07]	
Individual	.09 (.03)	1.10 [1.03, 1.17]**	
Violent			
Block 1 Historical	.12 (.04)	1.13 [1.05, 1.20]***	12.25 (1)**
Block 2—Chi-square change			Δ10.53 (1)**
Historical	.03 (.05)	1.03 [.93, 1.13]	
Individual	.15 (.05)	1.16 [1.06, 1.27]***	

Note. Cox regression models included the historical domain score as the only predictor at Block 1 and added the social, individual, and protective domains at Block 2 using Backward elimination. χ^2 values for the overall model are reported for Block 1 and the change in chi-square is reported at Block 2. SAVRY = Structured Assessment of Violence Risk in Youth; *SE* = standard error; CI = confidence interval; *df* = degrees of freedom. * $p < .05$. ** $p < .01$. *** $p < .001$ for the Exp[B]'s and * $p < .017$. ** $p < .003$ for χ^2 's to correct for multiple tests.

regardless of the expertise of the professional completing the assessment. Additional research clearly is needed to understand the issue, but preliminary evidence indicates that professionals operating in various roles can achieve suitable levels of rater agreement on SPJ tools in the field (e.g., de Vogel & de Ruiter, 2006; Douglas & Belfrage, 2014).

In order to replicate the positive findings for the SRR observed in the lab in the field it requires teaching individuals how to weigh the relevance of risk factors differently as a function of the risk for the type of offending being assessed. JPOs in particular may need more intensive and sustained training on how to make this risk level decision. Qualitative data from these JPOs regarding how they generate the SRR suggested that a significant portion of them either put considerable emphasis on youths' history of violence, took an average of the risk item factor ratings, or misunderstood the concept of the individual relevance of items (Guy, Nelson, Fusco-Morin, & Vincent, 2014). Very little is known about how much individual JPOs weighted protective factors in their SRR decisions, but the protective factors did not appear to be significantly related to reoffending in this study. Further justification that this may be a training issue is that there was a clear site-level disparity in the ability of JPOs to make the SRRs. This site-level disparity despite the lack of a site-level disparity in the interrater reli-

Table 5
Hierarchical Cox Regression Comparing Predictive Validity of YLS/CMI Scales

Recidivism type	β (*SE*)	Exp(B) [95% CI]	χ^2 (*df*)
Any			
Block 1 Prior/current offenses	.39 (.23)	1.48 [.95, 2.30]	2.98 (1)
Block 2—Chi-square change			**Δ**25.97 (4)**
Prior/current offenses	.23 (.24)	1.25 [.78, 2.00]	
Peer relations	.44 (.17)	1.55 [1.12, 2.15]**	
Substance abuse	.34 (.15)	1.40 [1.04, 1.90]*	
Personality/behavior	.61 (.24)	1.83 [1.16, 2.90]**	
Attitudes/orientation	−.47 (.26)	.62 [.38, 1.03]	
Nonviolent			
Block 1 Prior/current offenses	.27 (.26)	1.31 [.79, 2.18]	1.13 (1)
Block 2—Chi-square change			**Δ**16.16 (2)**
Prior/current offenses	.08 (.26)	1.08 [.64, 1.81]	
Peer relations	.42 (.18)	1.53 [1.08, 2.16]*	
Substance abuse	.38 (.16)	1.47 [1.07, 2.02]*	

Note. Cox regression models included the prior/current offenses scale score as the only predictor at Block 1 and added the seven dynamic scales at Block 2 using Backward elimination. χ^2 values for the overall model are reported for Block 1 and the change in chi-square is reported at Block 2. YLS/CMI = Youth Level of Service/Case Management Inventory; *SE* = standard error; CI = confidence interval; *df* = degrees of freedom. * $p < .05$. ** $p < .01$ for the Exp[B]'s and * $p < .017$. ** $p < .003$ for χ^2 's to correct for multiple tests.

Table 6
Percentage of Youth Scoring Moderate to High Risk on SAVRY Items and YLS/CMI Domains by Race

Item	White (n = 75) % Risk	Black (n = 302) % Risk	*OR* [95% CI]
SAVRY risk items			
History of violence	50.0	52.5	1.11 [.66, 1.84]
History of nonviolent offending	62.2	56.9	.81 [.48, 1.36]
Early initiation of violence	34.7	42.5	1.39 [.82, 2.36]
Past supervision or intervention failures	26.7	22.3	.79 [.44, 1.41]
History of self-harm or suicide attempts	20.0	10.0	.44 [.23, .88]*
Exposure to violence in the home	48.0	23.9	.34 [.20, .58]***
Childhood history of maltreatment	29.3	17.3	.51 [.28, .90]*
Parental or caregiver criminality	49.3	37.5	.62 [.37, 1.03]
Early caregiver disruption	31.1	31.0	1.00 [.58, 1.73]
Poor school achievement	76.0	80.4	1.30 [.71, 2.37]
Peer delinquency	68.0	65.5	.78 [.46, 1.34]
Peer rejection	25.3	34.2	1.53 [.87, 2.72]
Stress and poor coping	45.3	51.3	1.27 [.77, 2.12]
Poor parental management	40.0	39.5	.98 [.59, 1.64]
Lack of personal or social support	24.0	26.2	1.13 [.63, 2.03]
Community disorganization	28.0	66.6	5.12 [2.93, 8.94]***
Negative attitudes	41.3	43.7	1.10 [.66, 1.84]
Risk taking or impulsivity	66.7	56.5	.65 [.38, 1.11]
Substance use difficulties	61.3	34.9	.34 [.20, .57]***
Anger management problems	50.0	61.6	1.60 [.96, 2.67]
Low empathy or remorse	29.3	34.0	1.24 [.72, 2.15]
Attention deficit hyperactivity difficulties	40.0	34.8	.80 [.48, 1.34]
Poor compliance	40.0	38.0	.92 [.55, 1.54]
Low interest or commitment to school	58.7	63.7	1.24 [.74, 2.07]
SAVRY protective items			
Prosocial involvement	42.5	47.5	1.23 [.73, 2.06]
Strong social support	61.1	67.8	1.34 [.79, 2.28]
Strong attachments and bonds	63.0	68.7	1.29 [.75, 2.19]
Positive attitude toward intervention & authority	73.6	72.0	.92 [.52, 1.65]
Strong commitment to school	46.6	44.0	.90 [.54, 1.51]
Resilient personality traits	49.3	68.7	2.25 [1.34, 3.79]**
	White (n = 232)	Black (n = 82)	
YLS risk domains			
Prior and current offenses/dispositions	14.2	30.5	2.64 [1.46, 4.81]**
Family circumstances/parenting	29.3	23.2	.73 [.41, 1.31]
Education/employment	75.9	82.9	1.55 [.81, 2.96]
Peer relations	51.3	62.2	1.56 [.93, 2.62]
Substance abuse	55.6	53.7	.93 [.56, 1.53]
Leisure/recreation	65.9	74.4	1.50 [.85, 2.64]
Personality/behavior	66.8	67.1	1.01 [.59, 1.73]
Attitudes/orientation	36.2	26.8	.65 [.37, 1.13]

Note. SAVRY = Structured Assessment of Violence Risk in Youth; YLS/CMI = Youth Level of Service/Case Management Inventory; *OR* = odds ratio; CI = confidence interval.
* $p < .05$. ** $p < .01$. *** $p < .001$.

ability of the SRR, which fell in the good to excellent range in each site ($ICC_{(A, 1)}$ = 0.75 for Site 1, 0.68 for Site 2, and 0.70 for Site 3). Site 2 may have informally instituted some procedures for making the SRR that degraded its validity.

Do These Risk Assessment Instruments Appear to Be Racially Biased?

Consistent with Skeem and Lowenkamp (2016), we used two methods to investigate whether these assessments were racially biased: (a) examination of test bias as evidenced by differential predictive validity; and (b) examination of the potential for disparate impact, evidenced by mean score differences. With respect to test bias, there was no evidence that the YLS/CMI or the SAVRY operate differently for Black and White youth at the test level. The AUCs for total scores and risk ratings were comparable for Blacks and Whites.

Contrary to popular belief (Holder, 2014; Tonry, 2014), Black youth did not score significantly higher on either instrument at the total scores or categorical risk rating levels; indicating Black youth would not be negatively impacted or treated differently as a result of use of these risk tools. In the few instances where Black youth scored higher than Whites, it was a result of use of static risk factors that would have been predicted by the critics (Holder,

2014; Tonry, 2014). Blacks scored significantly higher than Whites on the YLS/CMI current offense/dispositions domain, which is based entirely on official juvenile history. Because this scale did not differentially predict reoffending by race, there was no appreciable test bias. On the other hand, the SAVRY historical scale items related to history of delinquent behavior did not show racial disparities, which lends credibility for the inclusion of prior offending information in risk assessment tools based on self-report rather than sole reliance on official records. Self-reported offending is likely to be a better indicator of actual risk for future delinquent behavior (see trajectories of delinquency in community samples; e.g., Moffitt, 1993; Odgers et al., 2012; Piquero et al., 2012), and is less susceptible to racial bias than only behaviors for which a youth has an official record (Elliott, 1994; Farrington, Loeber, & Stouthamer-Loeber, 2003). The only other instance where Black youth scored significantly higher than White youth was on the SAVRY community disorganization item, indicating they were more likely to come from a high crime area with poverty. Again, this did not have a differential impact in overall risk ratings or prediction of reoffending.

White youth were significantly more likely to have a number of other static risk factors than Black youth (e.g., history of suicide or self-harm attempts, exposure to violence in the home, exposure to physical abuse or neglect), and were more likely to be rated as having substance abuse difficulties on both instruments. Many of the differences we observed were consistent with Chapman et al. (2006) and previous work examining racial differences in risk factors and delinquent behavior (Farrington et al., 2003). In Farrington and colleagues' (2003) Pittsburgh Youth Study, for example, analyses were based on self-report and official delinquency records of boys aged 10 with a follow-up of 5.8 years. Their results indicated that there were some racial differences in the prevalence of risk factors (e.g., 47% of Blacks lived in a disadvantaged neighborhood as compared to 1% of Whites; 21% of Black boys had a court petition compared to 1% of White boys) and relevance of risk factors to later delinquency.

Since the inception of the Juvenile Justice and Prevention Act of 1974, which aimed to ensure fair and equal treatment for every youth in the justice system, regardless of race or ethnicity, policymakers have become increasingly concerned with the overrepresentation of minorities within the juvenile justice system. Policymakers, directors, judges, and attorneys have opposed the inclusion of risk assessments out of fear that they may exacerbate disparities (Holder, 2014). Results reported here and in prior research (Vincent et al., 2011), including studies of other risk instruments (McCafferty, 2016; Skeem & Lowenkamp, 2016), strongly support the use of validated risk instruments for both Black and White individuals who come into contact with the law. Clearly, it is possible that use of these instruments in community samples would lead to greater disparate impact and potential for bias. However, once youth enter the juvenile justice system, the presence of particular risk factors becomes much more similar.

Due to the diverse nature of the youth within the JJ system and the fact that youth of color are overrepresented in the system, it is critical that professionals only use risk assessment tools for real world decision making if research exists that specifically has examined test bias when predicting reoffending among different racial and ethnic groups. The current findings contribute to this small but growing critical research base and support for the SAVRY and YLS/CMI. A key point foundational to this recommendation is that assessment is a process that goes beyond use of a particular risk assessment tool. People make decisions, not tools. The assessment process can be compromised because of evaluator bias (bias in an evaluator's use of tool) despite the use of a test with evidence to indicate it is not biased from a psychometric perspective. This underscores the importance of high-quality training for professionals, including JPOs, who assess and manage risk for reoffending. More generally, enhancing cultural competency should be a priority for all professionals in this field (see Aggarwal, 2012).

The research community bears an important responsibility moving forward to continue to study racial bias. Broadly speaking, bias in assessment tools can be a function of problems with conceptual equivalence (differences in the meaning or relevance of the construct an instrument was developed to measure), structural equivalence (differences in the manifestation of the construct), metric equivalence (differences in the quantitative measurement of the construct), and/or predictive equivalence (differences in the prognostic value of quantitative measurements of the construct; see Hart, 2016; van deVijver & Tanzer, 2004). Of course, bias in the assessment process more generally also can result from racial (or more broadly, cultural) bias on the part of evaluators. Ideally, samples studied in future research should be large and culturally and racially heterogeneous. Modern test theory methods, such as item response theory, should be used to study measurement bias, including structural bias; analytic approaches such as logistic regression or event history analysis should be used to evaluate prediction bias (see Hart, 2016; van deVijver & Tanzer, 2004).

What is the Contribution of Dynamic Risk Factors?

As expected, results provided evidence that the dynamic risk factors were more predictive of future reoffending than static risk factors on both the SAVRY and YLS/CMI. Regarding the SAVRY, the historical scale was a significant predictor of all three types of recidivism on its own; however, once the dynamic scales were entered into the regression model, the historical scale was no longer significant and the individual/clinical scale took precedence. These results are consistent with prior research that has indicated dynamic domains have incremental predictive validity over the historical domain (Vincent, Chapman et al., 2011), or are more predictive (as evidenced by larger AUC values) of violent recidivism than the historical domains (Gammelgard, Koivisto, Eronen, & Kaltiala-Heino, 2015; Guy, 2008; Hilterman et al., 2014). Unlike the SAVRY historical scale, surprisingly the YLS/CMI prior and current offenses/disposition scale alone did not predict general reoffending, but was associated with violent reoffending. The YLS/CMI offense history scale had fairly wide confidence intervals around its predictive validity estimates, indicating there may have been more error relative to SAVRY indices. The dynamic scales most predictive of reoffending were similar for both instruments. On the YLS/CMI, these were peer relations, substance abuse, and personality/behavior. On the SAVRY, the dynamic scale with the strongest prediction was the individual/clinical scale, which comprises many of the same characteristics as these YLS/CMI scales (e.g., impulsivity and risk-taking, attention problems, negative attitudes, substance abuse, lack of remorse and empathy) with the exception of peer relations. The only dynamic

scales or items that showed racial disparities pertained to substance abuse, which was significantly more common among Whites than Black youth.

There are a few takeaway messages from these findings. First, dynamic risk factors improve estimates of risk of reoffending and may be more meaningful than static risk factors in this respect. Second, dynamic risk factors are less susceptible to having a disparate impact for Black versus White youth. Third, risk-need-responsivity (RNR)-related studies refer to the "Central 8" areas of risk factors, often referencing what are known as the "Big Four" to describe which risk factor domains are the most influential to reoffending: criminal history, antisocial personality (e.g., stimulation seeking, low self-control), antisocial attitudes (e.g., procriminal cognitions or thinking styles and values), and antisocial associates (Andrews & Bonta, 2006). The Big Four were derived from theory and research on adult offenders, and are often extended downward to adolescents. However, there has been little evidence to justify this downward extension. Results of this study suggest this downward extension is premature until more research is conducted with adolescents.

Consistent with the evidence for the Big Four with adults, this study, along with data from other youth studies (Haqanee et al., 2015; Holsinger, Lowenkamp, & Latessa, 2006; Thompson & McGrath, 2012), indicated personality and disruptive behavioral problems are one of the critical risk domains for youth. Another similarity was the influence by negative or criminal peers, which meta-analyses have in fact demonstrated to be one of the strongest risk factors among older adolescents (Lipsey & Derzon, 1998).

Inconsistent with the evidence from adult research, the criminal history items did not appear to be among the "Big Four" in this study given the dynamic domains were stronger predictors of reoffending and accounted for most of the variance. Another discrepancy from the adult literature was the prominence of the substance abuse scale, which is among the Central 8 risk factor areas for adults but is not described as among the Big Four. It is quite plausible that substance abuse among adolescents is a more significant risk factor than among adults. Indeed, the Pathways to Desistance study of serious adolescent offenders found the only juvenile justice intervention to reduce reoffending was substance abuse treatment, and substance abuse problems magnified the presence of other risk factors (Chassin, Knight, Vargas-Chanes, Losoya, & Naranjo, 2009). Finally, unlike the adult research, the current study found attitudes/orientation was inversely related to risk. Given that our results are inconsistent with those of Haqanee, Peterson-Badali, and Skilling (2015), it may be the case that JPOs in the present study were either misled by the youth who were the most antisocial in their thinking, and consequently rated them too low, or were not asking the right questions to rate these items accurately.

In sum, results of this study add to the evidence supporting the essential inclusion of dynamic risk factors in risk assessment instruments by also demonstrating that these risk factors are not as susceptible to racial differences within the JJ setting as are items based on juvenile history. Moreover, the assessment of dynamic or variable risk factors is an essential component of any risk assessment if one wishes to guide treatment and service strategies. The concept of a variable factor is synonymous with a criminogenic need, defined as a factor that "when changed, (is) associated with changes in the probability of recidivism" (Andrews & Bonta, 2003, p. 261). Monahan and Skeem (2014) noted that there is a difference between a variable risk factor and a variable risk marker, which can change as a result of the passage of time. Although it is uncertain whether a particular variable item is a risk marker or a risk factor in a particular individual, it is clear that inclusion of these items is particularly important when assessing adolescents, a period in which we would expect more change in risk simply as a result of development and maturation (Borum & Verhaagen, 2006; Hoge & Andrews, 1996; Steinberg & Schwartz, 2000).

Limitations

The findings should be considered in light of three notable limitations. First, due to the location of the probation offices the majority of youth in the SAVRY sample were Black and the majority of youth in the YLS/CMI sample were White. These samples are representative of the JJ populations in these states, but the unequal sample sizes in the analyses may be contributing to the lack of significant differences for some risk factors.

A second limitation was the use of only official court records, which can be particularly problematic when investigating racial differences in the predictive validity of a risk assessment tool. Official court records of delinquent activities may exaggerate differences in base rates of violence between White and Black youth, relative to self-report information, perhaps as a result of policing or court practices (Farrington et al., 2003). In addition, relying on only one source of recidivism may underestimate base rates, as at least some antisocial acts go undetected by the police (Douglas & Ogloff, 2003; Farrington, Auty, Coid, & Turner, 2013). Surprisingly, there was not a significant racial difference in the base rates of reoffending in either sample, but that does not mean there would not have been self-reported differences (such that White youth had higher rates of reoffending for behaviors that were undetected). Future studies of racial bias in the predictive validity of risk assessment tools should incorporate self-report offending data.

A third issue that can affect our interpretation of the results is that we examined the predictive validity of the initial risk assessments conducted just before disposition for each youth. Many youths did not reoffend until up to 5.88 months later in the SAVRY sample and 7.05 months later in the YLS/CMI sample. Conceivably, these risk assessments guided case management practices in a manner that reduced youths' risk and decreased rates of reoffending. This especially may be the case in this study because JPOs and probation offices underwent intensive policy changes and training related to the RNR approach to improve case management (Andrews & Bonta, 2003, 2010; Andrews, Bonta, & Hoge, 1990). In fact, results indicated that risk level guided dispositions, service allocation, and supervision decisions in the jurisdictions used in this study (Vincent, Guy, Perrault, & Gershenson, 2016).

If case management was conducted well, we actually would hope for no association between risk level and reoffending because higher risk youth should be prevented from reoffending. The impact of appropriate and individualized case management guided by risk assessment to mitigate risk has been demonstrated, with some mixed findings, in the adult (e.g., Belfrage et al., 2012) and adolescent (Luong & Wormith, 2011; Singh et al., 2011; Vieira,

Skilling, & Peterson-Badali, 2009) literature. This concept underlying the reduction approach to risk assessment (vs. the prediction approach; see Heilbrun, 1997) is the cornerstone of the SPJ model and also is a key feature of the YLS/CMI. That is, findings from the individual risk assessment should be used to directly influence the types of risk management strategies in service of reducing risk for reoffending. The dynamic nature of risk, particularly among adolescents, can affect results of field predictive validity where risk assessments were implemented in ways that would not be observed in lab-based predictive validity studies. A stronger research design for investigating predictive validity would have been to use the risk assessment completed closest in time to each youth's reoffending. Such a design was not feasible because too few youths were reassessed in the present study, even though policy in each jurisdiction dictated routine reassessments every 6 months. This remains an important area for future research.

References

Aggarwal, N. K. (2012). Adapting the cultural formulation for clinical assessments in forensic psychiatry. *The Journal of the American Academy of Psychiatry and the Law, 40,* 113–118.

American Educational Research Association, American Psychological Association, and National Council on Measurement Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson Publishing Company.

Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct* (4th ed.). Cincinnati, OH: Anderson Publishing Company.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16,* 39–55. http://dx.doi.org/10.1037/a0018362

Andrews, D. A., Bonta, J., & Hoge, R. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19–52. http://dx.doi.org/10.1177/0093854890017001004

Baird, C. (2009, February). *A question of evidence: A critique of risk assessment models used in the justice system. Special report*. San Diego, CA: National Council on Crime and Delinquency.

Barnes, A. R., Campbell, N. A., Anderson, V. R., Campbell, C. A., Onifade, E., & Davidson, W. S. (2016). Validity of initial, exit, and dynamic juvenile risk assessment: An examination across gender and race/ethnicity. *Journal of Offender Rehabilitation, 55,* 21–38. http://dx.doi.org/10.1080/10509674.2015.1107004

Bechtel, K., Lowenkamp, C. T., & Latessa, E. (2007). Assessing the risk of re-offending for juvenile offenders using the Youth Level of Service/Case Management Inventory. *Journal of Offender Rehabilitation, 45,* 85–108. http://dx.doi.org/10.1300/J076v45n03_04

Belfrage, H., Strand, S., Storey, J. E., Gibas, A. L., Kropp, P. R., & Hart, S. D. (2012). Assessment and management of risk for intimate partner violence by police officers using the Spousal Assault Risk Assessment Guide. *Law and Human Behavior, 36,* 60–67. http://dx.doi.org/10.1037/h0093948

Borum, R. (2000). Assessing violence risk among youth. *Journal of Clinical Psychology, 56,* 1263–1288. http://dx.doi.org/10.1002/1097-4679(200010)56:10<1263::AID-JCLP3>3.0.CO;2-D

Borum, R., Bartel, P. A., & Forth, A. (2006). *Structured Assessment of Violence Risk in Youth* (SAVRY). Lutz, FL: Psychological Assessment Resource, Inc.

Borum, R., & Verhaagen, D. (2006). *Assessing and managing violence risk in juveniles*. New York, NY: Guilford Press.

Chapman, J. F., Desai, R. A., Falzer, P. R., & Borum, R. (2006). Violence risk and race in a sample of youth in juvenile detention: The potential to reduce disproportionate minority confinement. *Youth Violence and Juvenile Justice, 4,* 170–184. http://dx.doi.org/10.1177/1541204006286316

Chassin, L., Knight, G., Vargas-Chanes, D., Losoya, S. H., & Naranjo, D. (2009). Substance use treatment outcomes in a sample of male serious juvenile offenders. *Journal of Substance Abuse Treatment, 36,* 183–194. http://dx.doi.org/10.1016/j.jsat.2008.06.001

de Vogel, V., & de Ruiter, C. (2006). Structured professional judgment of violence risk in forensic clinical practice: A prospective study into the predictive validity of the Dutch HCR-20. *Psychology, Crime & Law, 12,* 321–336. http://dx.doi.org/10.1080/10683160600569029

Douglas, K. S., & Belfrage, H. (2014). Interrater reliability and concurrent validity of the HCR-20 Version 3. *International Journal of Forensic Mental Health, 13,* 130–139. http://dx.doi.org/10.1080/14999013.2014.908429

Douglas, K. S., & Ogloff, J. R. P. (2003). Violence by psychiatric patients: The impact of archival measurement source on violence base rates and risk assessment accuracy. *Canadian Journal of Psychiatry, 48,* 734–740.

Elliott, D. S. (1994). Serious violent offenders: Onset, developmental course, and termination. The American Society of Criminology 1993 presidential address. *Criminology, 32,* 1–21. http://dx.doi.org/10.1111/j.1745-9125.1994.tb01144.x

Farrington, D. P., Auty, K. M., Coid, J. W., & Turner, R. E. (2013). Self-reported and official offending from age 10 to age 56. *European Journal of Criminal Policy and Research, 19,* 135–151. http://sci-hub.cc/10.1007/s10610-012-9195-x

Farrington, D. P., Loeber, R., & Stouthamer-Loeber, M. (2003). How can the relationship between race and violence be explained? In D. F. Hawkins (Ed.), *Violent crime: Assessing race and ethnic differences* (pp. 213–237). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511499456.014

Gammelgard, M., Koivisto, A. M., Eronen, M., & Kaltiala-Heino, R. (2015). Predictive validity of the structured assessment of violence risk in youth: A 4-year follow-up. *Criminal Behaviour and Mental Health, 25,* 192–206. http://dx.doi.org/10.1002/cbm.1921

Gutierrez, L., Wilson, H. A., Rugge, T., & Bonta, J. (2013). The prediction of recidivism with aboriginal offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology and Criminal Justice, 55,* 55–99. http://dx.doi.org/10.3138/cjccj.2011.E.51

Guy, L. (2008). *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey* (Unpublished doctoral dissertation). Simon Fraser University, Burnaby, BC.

Guy, L. S., Douglas, K. S., & Hart, S. D. (2015). Risk assessment and communication. In B. Cutler & P. Zapf (Eds.), *APA handbook of forensic psychology* (pp. 35–86). Washington, DC: American Psychological Association.

Guy, L. S., Nelson, R. J., Fusco-Morin, S. L., & Vincent, G. M. (2014). What do juvenile probation officers think of using the SAVRY or YLS/CMI for case management, and do they use the instruments properly? *International Journal of Forensic Mental Health, 13,* 227–241. http://dx.doi.org/10.1080/14999013.2014.939789

Guy, L., & Vincent, G. M. (2011, March). *Field inter-rater reliability of the SAVRY and YLS/CMI with probation officers*. Presentation at the 2011 annual meeting of the American Psychology Law Society, Miami, FL.

Hannah-Moffat, K. (2012). Actuarial sentencing: An unsettled proposition. *Justice Quarterly, 1,* 1–27.

Hannah-Moffat, K., & Maurutto, P. (2010). Re-conceptualizing pre-sentence reports: Risk and race. *Punishment & Society, 12,* 262–286. http://dx.doi.org/10.1177/1462474510369442

Haqanee, Z., Peterson-Badali, M., & Skilling, T. (2015). Implementation of risk assessment in case management for youth in the justice system. *Criminal Justice and Behavior, 42,* 304–320. http://dx.doi.org/10.1177/0093854814549595

Harcourt, B. E. (2010, September). *Risk as a proxy for race*. University of Chicago Law & Economics Olin Working Paper 535; University of Chicago Public Law Working Paper 323. Retrieved from http://ssrn.com/abstract=1677654

Hart, S. D. (2016). Culture and violence risk assessment: The case of *Ewert v. Canada*. *Journal of Threat Assessment and Management, 3,* 76–96. http://dx.doi.org/10.1037/tam0000068

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. *Law and Human Behavior, 21,* 347–359. http://dx.doi.org/10.1023/A:1024851017947

Heilbrun, K. (2010). *Violence risk assessment in adults*. Oxford, UK: Oxford University Press.

Hilterman, E. L. B., Nicholls, T. L., & van Nieuwenhuizen, C. (2014). Predictive validity of risk assessments in juvenile offenders: Comparing the SAVRY, PCL: YV, and YLS/CMI with unstructured clinical assessments. *Assessment, 21,* 324–339. http://dx.doi.org/10.1177/1073191113498113

Hoge, R. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior, 29,* 380–396. http://dx.doi.org/10.1177/0093854802029004003

Hoge, R. D., & Andrews, D. A. (1996). *Assessing the youthful offender: Issues and techniques*. New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4757-9296-6

Hoge, R. D., & Andrews, D. A. (2006). *youth level of service/case management inventory: User's manual*. North Tonawanda, NY: Multi-Health Systems.

Holder, E. (2014, August). National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference. Philadelphia, PA.

Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2006). Predicting institutional misconduct using the Youth Level of Service/Case Management Inventory. *American Journal of Criminal Justice, 30,* 267–284. http://dx.doi.org/10.1007/BF02885895

Jessor, R., Van Den Bos, J., Vanderryn, J., Costa, F. M., & Turbin, M. S. (1995). Protective factors in adolescent problem behavior: Moderator effects in developmental change. *Developmental Psychology, 31,* 923–933. http://dx.doi.org/10.1037/0012-1649.31.6.923

Jung, S., & Rawana, E. P. (1999). Risk and need assessment of juvenile offenders. *Criminal Justice and Behavior, 26,* 69–89. http://dx.doi.org/10.1177/0093854899026001004

Juvenile Justice and Delinquency Prevention Act of 1974 (JJDPA) (P. L. 93–415, 88 Stat. 1109) and subsequent reauthorizations (1997, 1980, 1984, 1988, 1992, 2002).

Lipsey, M. W., & Derzon, J. H. (1998). Predictors of violent or serious delinquency in adolescence and early adulthood: A synthesis of longitudinal research. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 86–105). Thousand Oaks, CA: Sage.

Luong, D., & Wormith, J. (2011). Applying risk/need assessment to probation practice and its impact on the recidivism of young offenders. *Criminal Justice and Behavior, 38,* 1177–1199. http://dx.doi.org/10.1177/0093854811421596

McCafferty, J. (2016). Unjust disparities? The impact of race on juvenile risk assessment outcomes. *Criminal Justice Policy Review*. Advance online publication. http://dx.doi.org/10.1177/0887403416634163

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46. http://dx.doi.org/10.1037/1082-989X.1.1.30

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100,* 674–701. http://dx.doi.org/10.1037/0033-295X.100.4.674

Monahan, J., & Skeem, J. L. (2014). The evolution of violence risk assessment. *CNS Spectrums, 19,* 419–424. http://dx.doi.org/10.1017/S1092852914000145

Mossman, D., & Somoza, E. (1991). ROC curves, test accuracy, and the description of diagnostic tests. *The Journal of Neuropsychiatry and Clinical Neurosciences, 3,* 330–333. http://dx.doi.org/10.1176/jnp.3.3.330

National Council on Crime and Delinquency. (2014). *NCCD compares juvenile justice risk assessment instruments: A summary of the OJJDP-Funded Study*. San Diego, CA: NCCD.

National Resource Council. (2013). *Reforming juvenile justice: A developmental approach*. Washington, DC: The National Academics Press.

Odgers, C. L., Caspi, A., Russell, M. A., Sampson, R. J., Arseneault, L., & Moffitt, T. E. (2012). Supportive parenting mediates neighborhood socioeconomic disparities in children's antisocial behavior from ages 5 to 12. *Development and Psychopathology, 24,* 705–721. http://dx.doi.org/10.1017/S0954579412000326

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior, 36,* 329–353. http://dx.doi.org/10.1177/0093854809331457

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment, 26,* 156–176. http://dx.doi.org/10.1037/a0035080

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 (YLS/CMI). *Criminal Justice and Behavior, 35,* 474–483. http://dx.doi.org/10.1177/0093854807313427

Piquero, A. R., Farrington, D. P., Fontaine, N. M., Vincent, G. M., Coid, J., & Ullrich, S. (2012). Childhood risk, offending trajectories, and psychopathy at age 48 Years (in the Cambridge Study in Delinquent Development). *Psychology, Public Policy, and Law, 18,* 577–598. http://dx.doi.org/10.1037/a0027061

Puzzanchera, C., Adams, B., & Hockenberry, S. (2012). *Juvenile court statistics 2009*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC analysis, Cohen's *d,* and *r*. *Law and Human Behavior, 29,* 615–620. http://dx.doi.org/10.1007/s10979-005-6832-7

Schmidt, F., Campbell, M. A., & Houdling, C. (2010). Comparative analyses of the YLS/CMI, SAVRY, and PCL:YV in adolescent offenders: A 10-year follow-up into adulthood. *Youth Violence and Juvenile Justice, 9,* 23–42. http://dx.doi.org/10.1177/1541204010371793

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,* 329–344. http://dx.doi.org/10.1177/0093854804274373

Schmidt, F., Sinclair, S. M., & Thomasdottir, S. (2011). Predictive validity of the youth level of service/case management inventory with youth who have committed sexual and non-sexual offenses. *Criminal Justice and Behavior, 43,* 413–430. http://dx.doi.org/10.1177/0093854815603389

Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31,* 449–462. http://dx.doi.org/10.1007/s10979-006-9071-7

Shepherd, S., Luebbers, S., & Dolan, M. (2013). Gender and ethnicity in juvenile risk assessment. *Criminal Justice and Behavior, 40,* 388–408. http://dx.doi.org/10.1177/0093854812456776

Shepherd, S. M., Luebbers, S., Ogloff, J. R. P., Fullam, R., & Dolan, M. (2014). The predictive validity of risk assessment approaches for young Australian offenders. *Psychiatry, Psychology and Law, 21,* 801–817. http://dx.doi.org/10.1080/13218719.2014.904262

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., Otto, R. K. (2014). International Perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13,* 193–206. http://dx.doi.org/10.1080/14999013.2014.922141

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31,* 499–513. http://dx.doi.org/10.1016/j.cpr.2010.11.009

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, & recidivism: Predictive bias and disparate impact. *Criminology, 54,* 680–712. http://dx.doi.org/10.1111/1745-9125.12123

StataCorp. (2015). *Stata statistical software: Release 14*. College Station, TX: StataCorp LP.

Steinberg, L., & Schwartz, R. (2000). Developmental psychology goes to court. In T. Grisso & R. Schwartz (Eds.), *Youth on trial* (pp. 9–31). Chicago, IL: University of Chicago Press.

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior, 36,* 345–355. http://dx.doi.org/10.1037/h0093930

Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter, 26,* 167–176. http://dx.doi.org/10.1525/fsr.2014.26.3.167

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54,* 119–135. http://dx.doi.org/10.1016/j.erap.2003.12.004

Vieira, T., Skilling, T., & Peterson-Badali, M. (2009). Matching court-ordered services with treatment needs: Predicting treatment success with young offenders. *Criminal Justice and Behavior, 36,* 385–401. http://dx.doi.org/10.1177/0093854808331249

Vincent, G. M., Chapman, J., & Cook, N. E. (2011). Risk-needs assessment in juvenile justice: Predictive validity of the SAVRY, racial differences, and the contribution of needs factors. *Criminal Justice and Behavior, 38,* 42–62. http://dx.doi.org/10.1177/0093854810386000

Vincent, G. M., Guy, L. S., Fusco, S. L., & Gershenson, B. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior, 36,* 225–236. http://dx.doi.org/10.1037/h0093974

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). *Assessment in juvenile justice: A guidebook for implementation*. Chicago, IL: John D. and Catherine T. MacArthur Foundation.

Vincent, G. M., Guy, L. S., Perrault, R. T., & Gershenson, B. (2016). Risk assessment matters, but only when implemented well: A multisite study in juvenile probation. *Law and Human Behavior, 40,* 683–696. http://dx.doi.org/10.1037/lhb0000214

Vincent, G. M., Paiva-Salisbury, M., Cook, N. E., Guy, L. S., & Perrault, R. T. (2012). Impact of risk/needs assessment on juvenile probation officers' decision making: Importance of implementation. *Psychology, Public Policy, and Law, 18,* 549–576. http://dx.doi.org/10.1037/a0027186

Wachter, A. (2015). *Statewide risk assessment in juvenile probation. JJGPS StateScan*. Pittsburgh, PA: National Center for Juvenile Justice.

Received March 2, 2016
Revision received December 14, 2016
Accepted December 15, 2016 ■