Predictive Validity of the Joint Risk Matrix With Juvenile Offenders: A Focus on Gender and Race/Ethnicity

Craig S. Schwalbe, Mark W. Fraser and Steven H. Day

Criminal Justice and Behavior 2007; 34; 348
DOI: 10.1177/0093854806292244

PREDICTIVE VALIDITY OF THE JOINT RISK MATRIX WITH JUVENILE OFFENDERS

A Focus on Gender and Race/Ethnicity

CRAIG S. SCHWALBE
Columbia University School of Social Work

MARK W. FRASER
STEVEN H. DAY
University of North Carolina at Chapel Hill

Actuarial risk assessment instruments promise to increase decision-making accuracy and equity in settings such as the juvenile justice system, but both aims presume high levels of predictive validity. Prior research suggests that the predictive validity of some juvenile justice risk assessment instruments differs across gender and race/ethnicity. The Joint Risk Matrix (JRM) described herein is an instrument developed to increase the predictive validity of risk assessment for the diverse populations served by the nation's juvenile courts. The predictive validity of the JRM was estimated on a sample of 536 court-involved juveniles. The instrument demonstrated acceptable levels of validity across all juveniles (AUC = .710). Gender-based differences were explained by gendered patterns of referral to out-of-home placements. Differences by race/ethnicity were reduced compared with previous reports. The findings suggest that risk assessment can be improved by including measures related to the behavior and demeanor of offenders and the cooperation of their parents or caretakers.

Keywords: risk assessment; juvenile justice; delinquency

Juvenile courts in the United States increasingly use actuarial risk assessment to gauge the chances that youths will commit new offenses. The use of risk assessment by juvenile justice systems more than doubled between 1990 and 2003 (Griffin & Bozynski, 2003; Towberman, 1992). Central to the success of risk assessment is predictive validity, the degree to which an instrument predicts future offending. Instruments with high levels of predictive validity increase the capacity of the courts to identify high-risk offenders and then to allocate court resources to these offenders. To the extent that court services are effective, risk assessment may contribute to reductions in recidivism. In the same vein, risk assessment protocols and instruments that lack predictive validity may have high error rates and misdirect court resources.

The predictive validity of risk assessment is related to the generalizability of risk factors to the diverse population of juveniles in the justice system. When predictive validity differs markedly across gender or race/ethnicity, for instance, errors in classification may differ by gender or race/ethnicity. Depending on the nature of these errors, risk assessment holds the potential to exacerbate gender and race/ethnicity disparities in court dispositions. One statistical solution is to treat demographic variables as risk factors in risk assessment instruments.

AUTHORS' NOTE: *Correspondence concerning this article should be addressed to Craig S. Schwalbe, Columbia University School of Social Work, 1255 Amsterdam Avenue, New York, NY 10027; e-mail: css2109@columbia.edu.*

CRIMINAL JUSTICE AND BEHAVIOR, Vol. 34, No. 3, March 2007 348-361
DOI: 10.1177/0093854806292244

However, to the extent that risk assessment instruments influence judicial dispositions regarding confinement and the restriction of civil liberties, this approach raises intractable ethical problems. Gottfredson and Jarjoura (1996) advocate including demographic variables in statistical risk models but only as an intermediate step to remove the effects of race on risk assessment predictive validity. In this study, we develop an alternative strategy.

The focus of this study is the North Carolina Assessment of Risk (NCAR), a brief, nine-item index. Two earlier studies showed that it classifies juveniles according to their risk of recidivism (Schwalbe, Fraser, Day, & Arnold, 2004; Schwalbe, Fraser, Day, & Cooley, 2006). However, these same studies indicated that its predictive validity is related to both gender and race/ethnicity; it predicts recidivism better for males and non-Latino White juveniles than for females and African American juveniles. Thus, it appears that the NCAR provides information for judicial decision making that is more useful for some groups than for others. Moreover, a recent meta-analysis demonstrated that its predictive validity was lower than most instruments reported in the literature (Schwalbe, 2005). Thus, the purpose of this study was to improve risk assessment for female and African American offenders in North Carolina.

HISTORY OF RISK ASSESSMENT

Risk classification in criminology dates to the 1920s (Monachesi, 1950). Ernest Burgess (1928) was first to show that an additive index of risk factors could predict future criminality. Burgess examined the records of 3,000 adult parolees to identify case characteristics associated with parole failure, defined as either a technical violation or new criminal behavior. His final scale consisted of 21 dichotomous equally weighted risk factors. Parole failure rates ranged from 1.5% for men with fewer than 5 risk factors to 76% for men with more than 17 risk factors. Burgess argued then, as contemporary authors argue today, that accurate risk assessment was essential for the efficient allocation of resources to prevent recidivism (Burgess, 1928, 1936; Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Howell, 2003).

The actuarial approach developed by Burgess (1928) remains the gold standard for risk assessment development. Actuarial risk assessment begins with the identification of risk factors through empirical research. Hoge (2002) defines two types of risk factors: static factors and criminogenic needs. Static factors are empirical markers of future delinquent behavior that cannot improve over time. Prototypical static factors include early age at first offense and other history of delinquent behavior (Cottle, Lee, & Heilbrun, 2001; Lipsey & Derzon, 1998). Criminogenic needs, on the other hand, are markers of future delinquent behavior that can show improvement over time. Prototypical criminogenic needs include psychological traits and problems, parent or family problems, and delinquent peer associations (Cottle et al., 2001; Lipsey & Derzon, 1998). Risk assessment instruments often combine static factors and criminogenic needs into an additive scale that is validated on an independent sample. After trimming weak items, the result is usually a brief risk assessment instrument that includes only the stronger predictors of delinquency (Shlonsky & Wagner, 2005).

Krysik and LeCroy (2002) developed such an instrument using classic actuarial methodology. They randomly divided a sample of adjudicated juvenile offenders into separate estimation (n = 7,001) and validation (n = 4,754) samples. A five-item instrument was

empirically derived using file information from juveniles in the estimation sample and cross-validated with juveniles in the validation sample. The predictive validity of Krysic and LeCroy's instrument (AUC = .684) was greater than the median risk assessment instrument (AUC = .649) in a recent meta-analysis of 40 validation studies (Schwalbe, 2005).

Not all risk assessment instruments have been developed with such a classic approach. Some risk assessment instruments are adaptations of other instruments, such as the Model Risk Assessment Instrument recommended by the Office of Juvenile Justice and Delinquency Prevention (OJJDP; Howell, 1995). Although these instruments have the appearance of actuarial instruments (similar risk factors, brevity), they are not based on the empirical rigor that characterizes actuarially based instruments. The NCAR is a prototypical example. Its nine items are a mix of static risk factors (age at first offense, number of prior referrals, most serious prior adjudication, prior assaults, runaway) and criminogenic need factors (substance abuse, school behavior problems, peer relations, parental supervision). It was constructed by juvenile justice professionals in North Carolina who worked in consultation with the National Council on Crime and Delinquency and researchers at the University of North Carolina at Chapel Hill. Despite the nonactuarial developmental process, 8 of 9 risk factors predicted recidivism at a statistically significant level in the most recent validation study (Schwalbe et al., 2006).

Other risk assessment instruments do not share the brevity of the classic actuarial approach. These instruments are akin to measurement scales common in the field of psychometrics and psychology (DeVellis, 2003). The Youth Level of Service Inventory/Case Management Inventory (YLS/CMI) measures risk in eight domains using 42 items (Hoge & Andrews, 2001). Seven domains (offense history, family circumstances/parenting, education/employment, peer relations, substance abuse, leisure/recreation, personality/behavior) focus on static risk and criminogenic needs factors measured by many risk assessment instruments. The last domain, attitudes/orientation, is unique. It measures what Hoge (2002) called "responsivity factors," individual characteristics that predispose juveniles to respond unfavorably to court intervention. Across 10 validation studies, its average predictive validity (r = .28) approaches that of risk assessment instruments in adult forensic settings (Schwalbe, 2005).

RACE AND GENDER EFFECTS

Longer instruments such as the YLS/CMI may have advantages over brief actuarial instruments and the OJJDP model instrument. Because they measure a more comprehensive range of risk factors, they may be more sensitive to variations in risk across gender and race. Two published studies (Jung & Rawana, 1999; Schmidt, Hoge, & Gomez, 2005) showed no gender effects on YLS/CMI predictive validity. Jung and Rawana (1999) also found no differences across race/ethnicity (Native Canadian vs. White). Moreover, the predictive validity of another longer risk assessment instrument, the 76-item Young Offender Level of Service Inventory, showed no differences across gender (N = 164; Ilacqua, Coulson, Lombardo, & Nutbrown, 1999). In contrast, shorter instruments such as the 10-item Orange County risk assessment (Sharkey, Furlong, Jimerson, & O'Brien, 2003) and the NCAR (Schwalbe et al., 2004, 2006) both had higher levels of predictive validity for males than for females, and the NCAR had higher levels of predictive validity for non-Latino White youths

than for African American youths. Although longer instruments take more time to complete, the data suggest that they have higher across-group predictive validity when compared with shorter instruments.

REVISING THE NCAR

Two strategies were employed in expanding the NCAR to better address across-group risk factors. First, the content of the instrument was changed. Following trends in the literature and large-scale meta-analyses of juvenile recidivism, the scope of measured risk was expanded. To the risk factors already measured by the NCAR (offense history, running away, substance use or abuse, school behavior problems, delinquent peer associations, and parental supervision) were added seven additional variables that measured individual or psychological risk, family criminality, and responsivity factors. Relative brevity was maintained by offsetting additions with reductions in the number of offense history variables. The final revised risk assessment included 14 items.

The MacArthur Study of Mental Disorder and Violence suggested a second strategy (Monahan et al., 2001). In a study of risk assessment for violence among adults discharged from psychiatric hospitals ($N = 951$), Monahan et al. (2001) showed that predictive validity could be increased through the application of multiple assessments. This is because multiple assessments increase confidence in low- and high-risk ratings when agreement among risk assessment instruments is high. In effect, multiple assessments enable a test of convergent validity; high levels of agreement among risk assessment instruments suggest that they measure the same construct: risk of recidivism. In the revised NCAR, this strategy was operationalized by dividing the risk factors into two separate scales: a static factors scale and a dynamic factors scale. Results from each scale were cross-classified into a Joint Risk Matrix. From the Joint Risk Matrix, a final classification into low-, medium-, and high-risk groups was made. A more detailed description of the Joint Risk Matrix is presented below.

The purpose of this study is to describe the predictive validity of the Joint Risk Matrix. Because of its expanded breadth, we expected the revised instrument to have improved predictive validity over the NCAR for female and African American youths. Moreover, we expected that overall predictive validity would improve because of gains made in convergent validity through the multiple assessment strategy.

METHOD

PARTICIPANTS

The predictive validity of the NCAR and the Joint Risk Matrix were assessed on a statewide proportional sample of delinquent juveniles from North Carolina. Each court district ($n = 39$) submitted risk and recidivism data for a preassigned quota of adjudicated juveniles. Quotas were calculated by multiplying the proportion of the total statewide population of delinquent offenders for each district by the target sample of 600 juveniles. Initial data collection began during February 2002. Follow-up data collection began during February 2003. In all, assessment and follow-up data were gathered for 590 youths. The majority of the juveniles were male (68%; $n = 402$). The sample was evenly balanced between African American and non-Latino White youths (45%, $n = 268$; and 47%, $n = 275$,

respectively) with the remaining youths (8%, $n = 47$) classified into other racial and ethnic groups. Because of their larger numbers relative to other groups, the sample was restricted to African American and non-Latino White youths. In addition, 7 cases were omitted because of missing assessment dates, making calculation of time to recidivism impossible. The final sample is composed of 536 juvenile offenders.

MEASURES

Joint Risk Matrix. The Joint Risk Matrix is based on three assessment tools. For each juvenile, court counselors completed the NCAR, a structured Needs Assessment instrument, and a supplemental assessment of additional risk and responsivity factors. These assessments were completed as a part of routine practice after an adjudication of delinquency (in this case, the equivalent of a finding of guilt for an adult) and prior to judicial disposition.

The NCAR and the Needs Assessment were developed concurrently and implemented in 2000 in juvenile courts throughout North Carolina. Both are similar to the model risk and needs assessment instruments recommended by OJJDP (Howell, 1995). The supplemental assessment was designed specifically for this study to expand available measures for the Joint Risk Matrix. In each assessment, individual risk factors were measured by two to five behaviorally anchored response options. Training required to complete the NCAR and Needs Assessment instrument is minimal. Court counselors, approximately 40% of whom were awarded college degrees with majors in criminal justice, received a formal, agency-sponsored orientation to the NCAR and Needs Assessment instrument when it was implemented in 2000.

The Joint Risk Matrix includes a total of 14 items. It includes 3 static factors (first offense prior to age 12, number of prior referrals, history of runaway) and 11 dynamic factors (alcohol or drug abuse; school behavior problems; peer delinquency; parental supervision; family criminality; hyperactivity, impulsivity, or attention problems; pattern of hostile or aggressive behavior; mental health problems; juvenile cooperation; expression of remorse; parental cooperation). Risk factor scores within each type were summed to create two separate risk scales: one for static risk and one for dynamic factors. The static risk scale has a theoretical range of 0 to 5 points. A cross-classification table of the static risk scale with recidivism rates identified three distinct risk classes: low risk (0-1), medium risk (2-3), and high risk (4-5). The dynamic scale has a theoretical range of 0 to 21 points. A cross-classification table of the dynamic risk scale with recidivism similarly identified three distinct risk classes: low risk (0-6), medium risk (7-12), and high risk (13+). These separate assessments were cross-classified into a 3-by-3 joint risk matrix. Patterns of recidivism rates, defined below, showed three stable risk classes (low, medium, high).

Recidivism. At the conclusion of the study (February 2003), court counselors reported recidivism and dates of new, if any, offenses for all juveniles. Recidivism is defined as any delinquency complaint subsequent to the initial adjudicated offense. Procedural violation of probation was excluded. Status offenses were rare (7%, $n = 9$). Sensitivity analysis demonstrated that the relationship between risk and recidivism was similar for these youths compared with juveniles who reoffended with delinquency offenses; therefore, status offenders were retained in the analysis.

Length of time in out-of-home placement. At the conclusion of the study, court counselors also reported the length of time in days that a juvenile spent in any type of out-of-home placement (foster care, group home, treatment hospital, juvenile correctional facility) during the period beginning with the initial assessment and concluding with the end date of the study or date of a new complaint (reoffense), whichever was earlier. Out-of-home placement is confounded with risk assessment as it restricts juvenile access to opportunities to reoffend. To date, no published study of risk assessment has controlled for time in placement.

ANALYSIS

The purpose of the analysis was twofold: (a) to compare the predictive validity of the Joint Risk Matrix with the NCAR and (b) to describe differences in predictive validity across gender and/or race/ethnicity. To accomplish these objectives, receiver operator characteristic curve (ROC) analysis and event history analysis were employed.

ROC analysis is increasingly used to describe predictive validity of risk assessment instruments. It analyzes the relationship between the sensitivity of a test (proportion of true positives) and its specificity (proportion of false positives) across the full range of risk assessment scores (Quinsey, Harris, Rice, & Cormier, 1998; Rice & Harris, 1995; Swets, 1996, 2000). Say, for example, that a risk assessment instrument measures risk on a scale from 0 to 10. Establishing a cut-point score of 9 or above to indicate high risk (i.e., predicted to reoffend) will result in a certain tradeoff between sensitivity and specificity. Compared to lower cut points, the higher cut point will result in higher rates of false negatives as more juveniles will be incorrectly predicted to not reoffend (lower sensitivity) while at the same time result in relatively lower rates of false positives (higher specificity). Traditionally, ROC analysis has served as a tool to establish optimal cut points. However, summary statistics, specifically the area under the curve (AUC), describe the overall predictive validity of a risk assessment instrument without regard to specific cut points, making it of interest to the present study. The AUC is interpreted as the probability that a randomly selected reoffender will have a higher risk score than a randomly selected nonreoffender. AUC scores close to .500 reflect an even trade-off between sensitivity and specificity, that is, predictive validity approaching chance. AUC scores close to 1.0 reflect high gains in sensitivity with small decreases in specificity, that is, predictive validity approaching perfection.

Kaplan-Meier survival curve analysis and Cox regression, forms of event history analysis (Hougaard, 2000), are used to describe recidivism (time to recidivism) and to explain any gender and race/ethnicity differences in predictive validity shown by the ROC analysis. These procedures are designed to control for censored observations that result from different lengths of follow-up. In the present study, median length of follow-up time was 294 days; 434 juveniles (81%) were followed for at least 270 days (9 months). Kaplan-Meier survival curve analysis is a nonparametric procedure that models time to recidivism (measured in days). The survivor function—denoted as $S(t)$—is an estimated probability that a juvenile will not have reoffended by time interval t; it is the inverse of the recidivism rate at time t. Cox regression is a semiparametric procedure. It estimates the hazard ratio, defined as the instantaneous probability of reoffending in any given time interval. Partial-maximum likelihood estimation is used to derive parameter estimates for linear predictors of the hazard of recidivism.

TABLE 1: Sample Characteristics

		Gender		*Race*	
	Total Sample[a]	*Male*[b]	*Female*[c]	*White*[d]	*African American*[e]
Male	.68	—	—	.68	.67
African American	.49	.49	.50	—	—
Mean age (*SD*)	14.6 (1.34)	14.5 (1.39)	14.7 (1.21)	14.5 (1.41)	14.6 (1.25)
Offense severity					
0[f]	.12	.09	.20	.15	.10
1[g]	.55	.54	.56	.56	.54
2[h]	.31	.35	.23	.27	.35
3[i]	.02	.02	.01	.02	.02
Offense type					
Person	.26	.23	.32	.21	.30
Property	.33	.35	.29	.33	.34
Drug	.10	.12	.07	.14	.07
Other	.41	.42	.39	.46	.36
Percent first-time offenders	42.5	44.5	41.6	46.7	38.3
Mean static risk scale score (*SD*)	1.38 (1.29)	1.35 (1.31)	1.44 (1.25)	1.22 (1.22)	1.55 (1.34)
Mean dynamic risk scale score (*SD*)	8.64 (3.70)	8.71 (3.74)	8.51 (3.61)	8.56 (3.72)	8.73 (3.68)
Percent out-of-home placement	22	21	22	23	20
Kaplan-Meier survivor function, S(270)	.77	.76	.79	.84***	.70***

a. $N = 536$.
b. $n = 363$.
c. $n = 173$.
d. $n = 272$.
e. $n = 264$.
f. Status offenses.
g. Misdemeanor offenses.
h. Felony offenses.
i. Serious felony offenses.
***$p < .001$.

RESULTS

Table 1 shows sample characteristics. As described previously, the sample was divided evenly between African American and non-Latino White juveniles. A majority of offenders (68%, $n = 363$) were male. Most offenders were adjudicated for a misdemeanor level offense. Two in 5 juveniles were first-time offenders, and about 1 in 5 was placed in an out-of-home placement during the follow-up period of the study. The Kaplan-Meier survivor function estimate for 9 months (270 days) shows that about 23% of juveniles were expected to reoffend during this period (i.e., 77% survived through the follow-up period without a new complaint).

Table 1 also shows similarities and differences by gender and race/ethnicity. Males and females did not differ in age, rates of out-of-home placements, percentage of first-time offenders, and rates of recidivism. They differed in offense severity and type. Compared with males, the odds of adjudication for a status offense were 2.5 times greater for females, and the odds of adjudication for a person-type offense were 1.6 times greater. Conversely, the odds of adjudication for a drug offense were 1.8 times greater for males than for females.

TABLE 2: Risk Factor Mean Values and Correlation With Recidivism

Risk Factors	M[a]	r
Static Risk Scale		
Age at first offense	.14	.12
Number of prior court referrals	.33	.15
History of running away from home or placement	.24	.06
Dynamic/Responsivity Scale		
Alcohol or drug abuse during past 12 months	.29	.08
School behavior problems during past 12 months	.79	.16
Peer delinquency or gang involvement	.43	.10
Parental supervision	.23	.14
Family criminality	.14	.10
Hyperactivity, impulsivity, attention	.32	.09
Pattern of hostile or aggressive behavior	.27	.08
Mental health problems	.62	.09
Juvenile noncooperation with court services	.18	.09
Juvenile lacks expression of regret or remorse	.56	.18
Parental noncooperation with court services	.09	.09

a. Risk assessment scores were transformed into a common metric (0-1) to permit comparisons.

African American and non-Latino White juveniles differed significantly by prior record and reoffense type. Fewer African American juveniles were first-time offenders and African American offenders had higher rates of recidivism. Odds of adjudication for a felony-level offense (odds ratio, OR = 1.5) and person offenses (OR = 1.6) were greater for African American juveniles; odds of adjudication for status offenses (OR = 2.5) and drug offenses (OR = 2.2) were greater for non-Latino White juveniles.

Table 2 shows the mean values for risk factors and their correlations with recidivism. To permit comparison of variable means for this analysis, risk factor scores were standardized on a common metric ranging from zero to one. Mean scores are elevated (> 0.50) for three variables: school behavior problems, mental health problems, and juvenile lacks expression of remorse or regret. Correlation coefficients show that effects are small for all variables. Risk factors with relatively larger effects on recidivism include juvenile lacks expression of remorse or regret, school behavior problems, number of prior offenses, and parental supervision.

THE JOINT RISK MATRIX

Table 3 shows the Joint Risk Matrix, a cross-classification table of the static risk scale, and the dynamic factors scale. The static risk scale most frequently rated juveniles in the low-risk category (59%, $n = 317$); the dynamic factor scale most frequently rated juveniles in the medium-risk category (57%, $n = 308$). On the diagonal, risk ratings of the static risk scale and dynamic factors scale have complete concurrence for 47% ($n = 255$) of juveniles. The association between the two scales was statistically significant ($r = .35$, $p < .001$). Of juveniles twice classified at low risk, 8.5% ($n = 11$) reoffended; 61.5% ($n = 8$) of juveniles twice classified at high risk reoffended.

Patterns in recidivism rates shown in Table 3 determined the final risk classification. In general, juveniles classified in the high-risk category for either assessment had the highest rates of recidivism and were grouped together as a single high-risk class ($n = 77$). This pattern did not hold for high static risk–low dynamic risk juveniles ($n = 4$), where the recidivism rate was relatively low (25%). Nevertheless, this small group of juveniles was classified into

TABLE 3: Joint Risk Matrix

	Dynamic Risk Level										
	Low (0-6)			*Medium (7-11)*			*High (12+)*			*Total*	
Static Risk Level	n	*Rate*	*Label*	n	*Rate*	*Label*	n	*Rate*	*Label*	n	*Rate*
Low (0-2)	129	(.085)	Low	173	(.249)	Medium	15	(.467)	High	317	(.192)
Medium (3-4)	41	(.219)	Medium	113	(.283)	Medium	23	(.435)	High	177	(.288)
High (5+)	4	(.250)	High	22	(.545)	High	13	(.615)	High	39	(.539)
Total	174	(.121)		308	(.283)		51	(.490)		536	(.250)

Note. n is juveniles classified, rate is recidivism rate, and label is the final risk classification.

TABLE 4: Receiver Operator Characteristic Analysis of the NCAR and Joint Risk Matrix by Gender and Race/Ethnicity

Sample	*Risk Instrument*	*AUC*	*95% Confidence Interval*
Full[a]	NCAR	.675	.616-.730
	Joint Risk Matrix	.710	.652-.764
Male[b]	NCAR	.692	.623-.755
	Joint Risk Matrix	.720	.651-.781
Female[c]	NCAR	.633	.518-.738
	Joint Risk Matrix	.689	.575-.787
Black[d]	NCAR	.633	.553-.708
	Joint Risk Matrix	.661	.581-.735
Non-Latino White[e]	NCAR	.706	.612-.788
	Joint Risk Matrix	.787	.698-.858

Note. NCAR = North Carolina Assessment of Risk.
a. $N = 536$.
b. $n = 397$.
c. $n = 186$.
d. $n = 264$.
e. $n = 269$.

the high-risk group to maintain conceptual clarity. The recidivism rate of juveniles twice classified as low risk ($n = 129$) was small compared with other groups (9%), leading us to classify these juveniles into a low-risk group. The recidivism rates of the remaining dynamic-static cross-classifications ranged from 22% to 28%, leading us to classify these juveniles into a medium-risk group ($n = 330$). The Kaplan-Meier survivor function shows that the three groups differed significantly in their rates of survival, $S(270)_{low} = .93$; $S(270)_{medium} = .76$; $S(270)_{high} = .55$; $\chi^2(536, 2) = 43.56$, $p < .001$.

COMPARATIVE PREDICTIVE VALIDITY

ROC analysis was used to compare the predictive validity of the Joint Risk Matrix with the NCAR for the full sample and across gender and race/ethnicity. As can be seen in Table 4, the Joint Risk Matrix had higher AUC scores than the NCAR. The increase for the full sample was 5%. Increases within demographic groups ranged from 4% for males and African American juveniles to 9% for female offenders to 11% for non-Latino White offenders. Despite these improvements, Joint Risk Matrix AUC scores remained highest for male and non-Latino White offenders, compared with female and African American offenders.

TABLE 5: Cox Regressions of Hazard of Recidivism on Risk Factors by Gender

	Males[a]		*Females*[b]	
Variable	*B*	SE	*B*	SE
African American	1.246****	0.290	0.957**	0.485
Out-of-home placement	–0.046***	0.015	0.021	0.034
Static risk scale	0.603**	0.275	1.728****	0.422
Dynamic factors scale	0.782****	0.165	0.577**	0.292
Static × race	–0.569*	0.330	–1.527****	0.505
Dynamic × placement	—	—	–0.064**	0.031
Model χ^2	$\chi^2(361, 5) = 63.37$, $p < .001$		$\chi^2(172, 6) = 25.14$, $p < .001$	

a. $n = 361$.
b. $n = 172$.
*$p < .10$. **$p < .05$. ***$p < .01$. ****$p < .001$.

Additional analyses were conducted to determine the degree to which gains in predictive validity were because of the addition of new risk factors versus the multiple assessment strategy. Two separate analyses were conducted. First, a third additive risk assessment instrument was constructed using all variables included in the static risk scale and the dynamic factors scale. Scores (range = 0–23, $M = 10$) were divided into risk classes (low, medium, high) based on their relationship to recidivism. Cut points were adjusted to maximize predictive validity. The best model achieved an AUC score lower than the NCAR (AUC = .669). This instrument classified a large proportion of juveniles into the medium-risk group (73%, $n = 391$) and few into the high-risk group (4%, $n = 21$). Next, a second Joint Risk Matrix was constructed using only static and dynamic variables from the NCAR. Several models were tested. The best model achieved an AUC score slightly higher than the NCAR but lower than the Joint Risk Matrix (AUC = .689). These analyses showed that the higher Joint Risk Matrix AUC scores were because of the combined effects of the increase in number of risk factors and of the multiple assessment strategy.

EXPLAINING GENDER AND RACE/ETHNICITY DIFFERENCES

To examine gender and race/ethnicity differences more closely, separate Cox regression models were estimated for males and females (see Table 5). Variables entered into each regression model included race, length of time in days spent in out-of-home placement, risk level on the static risk scale, risk level on the dynamic factors scale, and interaction effects. For both males and females, interpretation of the parameter estimate for race and of the parameter estimate for the static risk scale is confounded by a significant interaction between static risk and race. Both analyses show that the effect of static risk is smaller for African American juveniles than for non-Latino White juveniles. Gender differences emerged when comparing effects of dynamic risk and out-of-home placements. For males, the effects of out-of-home placements and dynamic risk were direct and distinct: Out-of-home placements reduced recidivism for all males, and higher levels of dynamic risk increased risk of recidivism. For females, on the other hand, interpretation of these parameter estimates was confounded by a significant interaction between out-of-home placement and dynamic risk so that longer time spent in out-of-home placement reduced the

effect of dynamic risk on recidivism rates. In other words, dynamic risk became a less potent predictor of recidivism as length of time in out-of-home placements increased.

Patterns of out-of-home placements were analyzed further to clarify these gender differences. Among low-risk juveniles, the odds of placement were greater for males than females (OR = 1.72); among medium-risk juveniles, the odds of placement were even (OR = 1.06); among high-risk juveniles, the odds of placement were lower for males than for females (OR = 0.71). Moreover, correlations between length of time in out-of-home placement and the static risk scale (r_{female} = .27, r_{male} = .16) and between length of time in out-of-home placement and the dynamic factors scale (r_{female} = .30, r_{male} = .21) were larger for female offenders than for male offenders. These findings suggest that judicial decisions regarding out-of-home placement were more aligned with risk level for females than males, resulting in a more targeted effect on high-risk female offenders than similarly situated male offenders. However, this conclusion is attenuated by the lack of statistical significance for these observed differences ($p > .10$).

To assess the effect of placement on the predictive validity of the Joint Risk Matrix, risk scores were adjusted according to the length of time juveniles spent in out-of-home placement. Low-, medium-, and high-risk levels were assigned numerical values (0, 1, 2, respectively), and .10 was subtracted from risk scores for each month spent in out-of-home placement. ROC analysis was conducted on these adjusted risk scores. The AUC score increased 20% for the full sample (from .710 to .850), 23% for males (from .692 to .854), and 32% for females (from .633 to .837). This adjustment for out-of-home placements led to substantial increases in predictive validity estimates and eliminated gender differences.

DISCUSSION

Compared to the NCAR, the Joint Risk Matrix had higher levels of predictive validity with all offenders. This improvement was because of the expansion of measured risk and the development of the multiple assessment strategy. The statistically significant, though modest, relationship between the dynamic factors and static risk scales indicates that they measure overlapping, yet distinct, risk constructs. This finding would seem to suggest that both contribute to a valid assessment of risk of recidivism.

Yet racial/ethnic differences surrounding the predictive validity of the static risk scale present a puzzle. On its face, this finding suggests that risk factors such as offending history predict future offending for some groups but not others. However, this hypothesis finds little support in the literature that shows instead a consistent relationship between early offending and later offending irrespective of race/ethnicity (Ayers et al., 1999; Maughan, Pickles, Rowe, Costello, & Angold, 2000; Nagin & Tremblay, 1999; Schaeffer, Petras, Ialongo, Poduska, & Kellam, 2003). Another possibility involves differences in base rates of prior offending and recidivism. In this sample, African American juveniles were more likely to have prior offenses and recidivism than were non-Latino White juveniles. Higher base rates on both counts reduced the discriminating power of offending history for African American juveniles. One explanation is that African American youths offended at higher rates than did non-Latino White youths. Another explanation is that law enforcement surveillance and court referral patterns themselves increased the odds of detection of African American delinquency compared with non-Latino White delinquency. Although self-report

studies of delinquency tend to cast doubt on the former explanation (Hawkins, Laub, & Lauritsen, 1999; Peeples & Loeber, 1994), literature on the disproportionate confinement of minorities suggests that the latter is plausible (Brownfield, Sorenson, & Thompson, 2001; Leiber & Mack, 2003; Pope & Feyerherm, 1995; Pope & Snyder, 2003).

The statistical interaction between out-of-home placement and dynamic risk among females suggests a different explanation for gender differences. For females, the effect of out-of-home placements was exclusively targeted at higher risk offenders. For females who remain in home, the relationship between risk and recidivism increases linearly as expected, so that high-risk females had higher rates of recidivism than did medium- and low-risk females; for females placed in out-of-home placements, recidivism rates for low-, medium-, and high-risk groups became more similar as length of time in placement increased. In contrast, the effect of out-of-home placements for males was independent of risk level. Because of this difference, out-of-home placements suppressed predictive validity estimates more sharply for females than for males. In fact, when a crude adjustment was made to risk level for length of time in out-of-home placement, gender differences in predictive validity vanished. This finding suggests either: (a) that out-of-home placements suppress recidivism more effectively for female offenders than for male offenders or (b) that out-of-home placements are allocated differently for females—perhaps more efficiently—favoring a more targeted effect. The findings themselves hint at the latter explanation by showing a trend toward a more consistent relationship between risk level and out-of-home placement for females than for males. Studies with additional samples are required to confirm this tentative hypothesis, however.

Thus, differences across both race/ethnicity and gender may be artifacts of processes exogenous to the risk assessment instrument itself. For African American juveniles, they may be the same processes that underlie disproportionate minority confinement; for females, they may be gender differences in the allocation and/or efficacy of out-of-home placements.

This finding points to a weakness in risk assessment research: Few studies control for the effect of intervening variables. As we demonstrated, one type of intervening variable, out-of-home placement, can have dramatic effects on estimates of predictive validity. The same might be true were it possible to control for other exogenous variables, such as law enforcement and judicial decision-making practices implicated in disproportionate minority confinement. This illustrates that validation studies conducted in dynamic settings characterized by discretionary decision making and active efforts to prevent future offending may produce spurious estimates of predictive validity. Researchers should strive to model the effects of intervening variables when estimating risk assessment predictive validity, although this is a challenging prospect.

This study has an important limitation. Validation on multiple samples is required to test the stability of the predictive validity of the Joint Risk Matrix in the presence of sampling error. The extent that the ROC analysis benefited from overfitting the data, especially in the determination of risk classes in the static risk scale and the dynamic factors scale, can only be determined through further empirical study. Evidence favoring the Joint Risk Matrix is found in that the 95% confidence intervals show an increase in statistical significance (greater distance from AUC = .500) for all groups relative to the NCAR. Nevertheless, cross-validation is warranted.

This weakness notwithstanding, the results suggest that the Joint Risk Matrix may provide better information for use in judicial decision making. On balance, the Joint Risk Matrix had higher levels of predictive validity than do previous instruments. The findings suggest that differences in predictive validity across race/ethnicity and gender may be related to exogenous forces that are beyond the scope of risk assessment. In sum, the cross-classification of static and dynamic risk factors, as done in the Joint Risk Matrix, appears to hold promise for improving risk assessment in juvenile justice.

REFERENCES

Ayers, C. D., Williams, J. H., Hawkins, J. D., Peterson, P. L., Catalano, R. F., & Abbott, R. D. (1999). Assessing correlates of onset, escalation, deescalation, and desistance of delinquent behavior. *Journal of Quantitative Criminology, 15*, 277-305.

Brownfield, D., Sorenson, A. M., & Thompson, K. M. (2001). Gang membership, race, and social class: A test of the group hazard and master status hypotheses. *Deviant Behavior, 22*, 73-89.

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, & A. J. Harno (Eds.), *The workings of the indeterminate sentence law and the parole system in Illinois.* Springfield: Illinois State Board of Parole.

Burgess, E. W. (1936). Protecting the public by parole and by parole prediction. *Journal of Criminal Law and Criminology, 27*, 491-502.

Cottle, C. C., Lee, R. L., & Heilbrun, K. (2001). The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior, 28*, 367-394.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Gottfredson, S. D., & Jarjoura, G. R. (1996). Race, gender, and guidelines-based decision making. *Journal of Research in Crime and Delinquency, 33,* 49-69.

Griffin, P., & Bozynski, M. (2003). *National overviews: State juvenile justice profiles*. Retrieved November 5, 2003, from http://www.ncjj.org/stateprofiles/

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323.

Hawkins, D. F., Laub, J. H., & Lauritsen, J. L. (1999). Race, ethnicity, and serious juvenile offending. In R. Loeber & D. P. Farrington (Eds.), *Serious & violent juvenile offenders.* London: Sage.

Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior, 29*, 380-396.

Hoge, R. D., & Andrews, D. A. (2001). *The youth level of service/case management inventory (YLS/CMI): Intake manual and item scoring key*. Ottawa, Ontario, Canada: Carleton University.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer-Verlag.

Howell, J. C. (1995). *Guide for implementing the comprehensive strategy for serious, violent, and chronic juvenile offenders*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Howell, J. C. (2003). *Preventing & reducing juvenile delinquency: A comprehensive framework*. Thousand Oaks, CA: Sage.

Ilacqua, G. E., Coulson, G. E., Lombardo, D., & Nutbrown, V. (1999). Predictive validity of the Young Offender Level of Service Inventory for criminal recidivism of male and female young offenders. *Psychological Reports, 84*, 1214-1218.

Jung, S., & Rawana, E. P. (1999). Risk and need assessment of juvenile offenders. *Criminal Justice and Behavior, 26*, 69-89.

Krysik, J., & LeCroy, C. W. (2002). The empirical validation of an instrument to predict risk of recidivism among juvenile offenders. *Research on Social Work Practice, 12*, 71-81.

Leiber, M. J., & Mack, K. Y. (2003). The individual and joint effects of race, gender, and family status on juvenile justice decision-making. *Journal of Research in Crime and Delinquency, 40*, 34-70.

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

Maughan, B., Pickles, A., Rowe, R., Costello, E. J., & Angold, A. (2000). Developmental trajectories of aggressive and non-aggressive conduct problems. *Journal of Quantitative Criminology, 16*, 199-221.

Monachesi, E. D. (1950). American studies in the prediction of recidivism. *Journal of Criminal Law and Criminology, 41*, 268-289.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., et al. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. Oxford, UK: Oxford University Press.

Nagin, D., & Tremblay, R. E. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development, 70*, 1181-1196.

Peeples, F., & Loeber, R. (1994). Do individual factors and neighborhood context explain ethnic differences in juvenile delinquency? *Journal of Quantitative Criminology, 10*, 141-157.

Pope, C. E., & Feyerherm, W. H. (1995). *Minorities and the juvenile justice system* (2nd printing). Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Pope, C. E., & Snyder, H. N. (2003). *Race as a factor in juvenile arrests*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737-748.

Schaeffer, C. M., Petras, H., Ialongo, N., Poduska, J., & Kellam, S. (2003). Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology, 39*, 1020-1035.

Schmidt, F., Hoge, R. D., & Gomez, L. (2005). Reliability and validity analyses of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior, 32*, 329-344.

Schwalbe, C. S. (2005). *A meta analysis of juvenile justice risk assessment predictive validity*. Unpublished manuscript.

Schwalbe, C. S., Fraser, M. W., Day, S. H., & Arnold, E. M. (2004). North Carolina Assessment of Risk (NCAR): Reliability and predictive validity with juvenile offenders. *Journal of Offender Rehabilitation, 40*, 1-22.

Schwalbe, C. S., Fraser, M. W., Day, S. H., & Cooley, V. (2006). Classifying juvenile offenders according to risk of recidivism: Predictive validity, race/ethnicity, and gender. *Criminal Justice and Behavior*, *33*, 305-324.

Sharkey, J. D., Furlong, M. J., Jimerson, S. R., & O'Brien, K. M. (2003). Evaluating the utility of a risk assessment to predict recidivism among male and female adolescents. *Education and Treatment of Children, 26*, 467-494.

Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. *Children and Youth Services Review, 27*, 409-427.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Lawrence Erlbaum.

Swets, J. A. (2000). Enhancing diagnostic decisions. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd ed.). New York: Cambridge University Press.

Towberman, D. B. (1992). A national survey of juvenile risk assessment. *Family & Juvenile Court Journal, 43*, 61-67.