Crime & Delinquency

http://cad.sagepub.com

Applying a Generic Juvenile Risk Assessment Instrument to a Local Context: Some Practical and Theoretical Lessons

Joel Miller and Jeffrey Lin

Crime Delinquency 2007; 53; 552
DOI: 10.1177/0011128706293689

Crime & Delinquency
Volume 53 Number 4
October 2007 552-580

10.1177/0011128706293689
http://cad.sagepub.com
hosted at
http://online.sagepub.com

Applying a Generic Juvenile Risk Assessment Instrument to a Local Context

Some Practical and Theoretical Lessons

Joel Miller
University of Málaga, Spain
Jeffrey Lin
University of California, Irvine

This article examines issues raised by the application of a generic actuarial juvenile risk instrument (the Model Risk Assessment Instrument) to New York City, a context different from the one in which it was developed. It describes practical challenges arising from the constraints of locally available data and local sensibilities and highlights important differences between locally relevant recidivism predictors and generic tool predictors. The analysis shows that the generic tool is less predictive than a locally developed risk-assessment tool and also performs less well than unassisted clinical judgment. This is true even after the generic tool has been validated and optimized on local data. This is because the tool does not include key demographic variables relevant to the New York City context.

Keywords: *risk assessment; actuarial; clinical judgment; juvenile recidivism; family court*

The principles underpinning actuarial risk assessment in justice populations were established many decades ago when research began to demonstrate that the characteristics of offenders were correlated with their subsequent behavior (Burgess, 1928; Glueck & Glueck, 1950). These insights form the basis of actuarial risk assessment tools, which have in common a core set of principles. Based on statistical analysis of offender characteristics and subsequent behavior, items or questions about a person's legal, psychological, and social characteristics are scored and combined to form a scale that is indicative of "risk"—such as the risk of rearrest or reconviction, the risk of absconding while on bail, or the risk of violation of parole or probation conditions (Clear, Wasson, & Rowland, 1988; M. R. Gottfredson & Gottfredson, 1986; S. D. Gottfredson & Moriarty, 2006). From the 1970s onward, these insights began to work their way into systematic practice, as structured predictive tools emerged within criminal justice settings as a basis for decision making (Andrews, 1989).

This article is concerned with risk assessment within the juvenile justice system, which has also been a fertile area for development in recent years. Today, the use of juvenile assessment tools is relatively widespread (Office of Juvenile Justice Delinquency Planning [OJJDP], 1995), and the development of empirically based risk assessments has been described in a number of studies and reports (Baird, 1984; Johnson, Wagner, & Matthews, 2002; LeCroy, Krysik, & Palumbo, 1998; Risler, Sutphen, & Shields, 2000; Schwalbe, Day, & Fraser, 2002; Washington State Institute for Public Policy, 2004). Specifically, we examine issues raised by the application of a generic juvenile risk instrument to a context different from the one in which it was originally developed. It applies an "off-the-shelf" risk tool to data generated by research in New York City (NYC). In doing so, the article explores practical challenges that accompany such a task and examines the predictive efficacy of the generic tool, both before and after it has been subject to validation and adjustment using local data.

Actuarial Risk Assessment

Research in a range of contexts, including mental health and behavior (Dawes, Faust, & Meehl, 1989), child welfare (Gambrill & Shlonsky, 2000), and the justice system (Andrews, Bonta, & Wormith, 2006; Carroll, Wiener, Coates, Galegher, & Alibrio, 1982; Glaser, 1955, 1962; D. M. Gottfredson & Beverly, 1962; Holland, Holt, Levi, & Beckett, 1983), attests to the fact that actuarial assessment tools make more accurate classifications than subjective or "clinical" judgments of professionals. Most definitively, Grove, Zald, Lebow, Snitz, and Nelson (2000) report on a meta-analysis of 130 studies of human health and behavior, including those in criminal justice, which showed that actuarial techniques were about 10% more accurate than clinical prediction on average. Actuarial prediction substantially outperformed clinical prediction in 33% to 47% of studies examined, and in only 6% to 16% of studies was it substantially less accurate.

Despite the advantages of actuarial tools, they come with some limitations. Notably, actuarial methods are not a golden bullet for anticipating who will and who will not reoffend, all having substantial margins of error in this regard (Ashford & LeCroy, 1990; M. R. Gottfredson & Gottfredson, 1984; S. D. Gottfredson, 1987; S. D. Gottfredson & Moriarty, 2006; Klein & Caggiano, 1986; Wiebush, Baird, Krisberg, & Onek, 1995). As such, contemporary risk assessment is more often concerned with classification of offenders into groups of offenders with different rates of recidivism than with prediction of offending in individuals (Ashford & LeCroy, 1990; Baird, 1984; Juvenile Sanctions Center, 2002; Loeber & Stouthamer-Loeber, 1987; Marczyk, Heilbrun, Lander, & DeMatteo, 2003; OJJDP, 1995; Wiebush et al., 1995). Efforts to improve the predictive efficacy and practical utility of actuarial tools have also come to emphasize the importance of the dynamic changeable factors—such as drug use, peer groups, family relationships, or employment—in addition to the static, unchangeable factors of an offender—such as his or her age, gender or previous convictions (Andrews et al., 2006; Gendreau, Goggin, & Little, 1996; S. D. Gottfredson & Moriarty, 2006). The movement from clinical tools to actuarial tools emphasizing static factors and, subsequently, to actuarial tools emphasizing dynamic factors has been characterized in terms of first-, second-, and third-generation assessment tools (Bonta, 1996).

Another very important limitation of risk assessment tools—and one with which we are concerned in this article—is their local validity. It is clear from existing research that the validity of any given risk assessment tool will vary from place to place and across time. For example, Clear et al. (1988) show how an adult risk screening instrument developed by the Wisconsin Bureau of Community Corrections and applied elsewhere transfers well to some jurisdictions but not to others. Wright, Clear, and Dickinson (1984), examining the same tool applied to NYC probationers, showed that that tool had significant weaknesses and included variables that did not predict risk for the NYC sample. In a juvenile justice context, Krysic and LeCroy (2002) show how a previously empirically validated assessment tool for juveniles in Arizona was outstripped by probation officers' subjective judgment several years later because of the changes in juvenile justice population over time. Despite these kinds of problems, the wholesale transfer of risk assessment tools to jurisdictions other than those in which they were validated apparently remains common practice (S. D. Gottfredson & Moriarty, 2006).

Risk Assessment in the Juvenile Justice System

There is reasonable consensus on the predictors of juvenile recidivism. The OJJDP (1995) provides direct comparison among 8 separate risk scales empirically derived both from county and state populations. School functioning was found on all instruments, and age at first referral, number of priors, substance abuse, peers, and family function were each found on at least 5 out of 8 instruments. Other factors included in half or fewer were (in decreasing order of frequency) current offense type, prior out-of-home placements, gender, runaway history, prior assault, victim of abuse and neglect, other factors, special education, and mental health problems.

Probably the most definitive empirical analysis of juvenile recidivism predictors is a meta-analysis of 22 separate studies of youth within the juvenile justice system (Cottle, Lee, & Heilbrun, 2001). They found a total of 23 factors statistically significant in their prediction of recidivism, of which the 10 most powerful were age at first commitment, age at first contact with the law, nonsevere pathology (e.g., stress and anxiety), family problems, conduct problems, effective use of leisure time, delinquent peers, length of first incarceration, number of out-of-home placements, and number of prior commitments. These findings have a close resemblance to the instruments reviewed by the OJJDP (1995), detailed above, though the analysis also highlights factors that are more difficult to measure in justice settings (as against research contexts) and hence are absent from the OJJDP tools. Such factors include, for example, nonsevere pathology and conduct problems.

There is an inevitable synergy between recidivism predictors and the broader canon of juvenile delinquency research and theory, though it is important to note that recidivism predictors are relevant to a specific subset of the general youth population—one that has already come into contact with the justice system and is therefore already in the throes of a delinquent career. For example, the prominence of age at first offense as a predictor of recidivism provides an important marker that helps distinguish between what Moffit (1993) calls “adolescent-limited” delinquents who start their antisocial behavior later in life have shorter-lived offending careers and “life-course-persistent” delinquents who have conduct problems that start earlier in life and have more persistent offending careers. Other factors such as family function, peer group, mental health problems, and conduct problems all show up in studies of risk factors for antisocial behavior and delinquency in the general youth population (Hawkins et al., 1998; Herrenkohl et al., 2000; Lipsey & Derzon, 1998; Loeber, Farrington, Stouthamer-Loeber, Moffitt, & Caspi, 1998; McCord, Widom, & Crowell, 2001; Tremblay & LeMarquand, 2001; Wasserman & Seracini, 2001).

Of particular interest to the current article is the emergence of a number of generic juvenile risk assessments that are promoted and used across a range of different populations. They include, for example, The Youth Level of

Service/Case Management Inventory (Hodge & Andrews, 1999), the Youth Assessment and Screening Instrument (Orbis Partners, 2006)**,** Back on Track! (Assessments.com, 2004), the Global Risk Assessment Device (Gavazzi et al., 2003), and the Model Risk Assessment Instrument (MRAI; Juvenile Sanctions Center, 2002). Typically, instruments to measure juvenile risk make use of no more than a dozen questions to generate an overall risk classification and draw on variables from readily available information, such as investigatory interviews and reviews of official records (OJJDP, 1995). Inevitably, these off-the-shelf tools are validated, in the first instance, on different populations from many of those on which they may later be applied. The premise of these tools seems to be that, when applied in new contexts, they should help predict recidivism even without local validation (e.g., Juvenile Sanctions Center, 2002; OJJDP, 1995; Orbis Partners, 2006) but can only be used with full confidence once validation and score adjustment (if necessary) have been performed in the new context (Flores, Travis, & Latessa, 2004; Juvenile Sanctions Center, 2002; OJJDP, 1995; Orbis Partners, 2006; Turner, Fain, & Sehgal, 2005). For example, the Juvenile Sanctions Center (2002), in discussing their MRAI, assert,

> Its use of factors that consistently appear on other validated instruments suggest that it is a tool that can be expected to work reasonably well for any jurisdiction that chooses to adopt it. It is imperative, however, that the adopting agency eventually . . . conducts the research necessary to validate the model scale on the local population. (p. 4)

The Current Study

This article explores the challenges of applying a generic juvenile risk assessment tool to a different context from the one in which it was initially developed and validated. The article exploits the opportunity afforded by the development and validation of a local, homegrown NYC juvenile risk assessment tool (hereafter referred to as the NYC tool), which involved the collection of baseline and recidivism data on 730 juvenile delinquents. The risk assessment tool was developed for use by family court probation officers writing predisposition reports on youth adjudicated as juvenile delinquents. It was designed to provide a risk classification of the juveniles according to their relative likelihood of recidivism, which could be used to guide whether a recommendation of institutional placement with the state or of community disposition (primarily probation) was made—and the intensity of programming and supervision associated with the latter. Importantly, it was constructed from items that matched information that was readily available to the NYC probation officers at the time of the predisposition report.

In effect, the article simulates and evaluates the process that proponents of generic tools recommend to local jurisdictions: First, introduce a tool validated elsewhere and use it straightaway to guide decision making. Subsequently, validate the tool based on the recidivism of the first cohort of offenders assessed and refine the tool to further increase its predictive efficacy.[1] The NYC research, as well as providing data to do this, offers two comparative benchmarks against which predictive efficacy of a generic tool can be compared (both pre- and post-validation). First, the risk scale of the locally developed NYC tool provides a key reference point. Second, the NYC data include an approximation of a clinical risk assessment of probation officers. Recommendations they make to court about whether a youth should be sent to institutional placement or receive a community disposition such as probation can be seen as an approximation for clinical assessments of risk. This comes with an important caveat. In practice, these recommendations are a product not only of perceived risk to the public but also of concerns about the welfare of the youth or even other bureaucratic considerations (Carter & Wilkins, 1967; Hood, 1966; Neubrauer, 1974; Rosecrance, 1985). Together, these other factors will tend to dilute and compromise the measure as an indication of the clinical risk assessment of probation officers, meaning that their recommendation will probably understate rather than overstate their actual ability to distinguish high- and low-risk youth. As such, a probation officer recommendation ought to be an easier threshold for an actuarial tool to beat than a more accurately measured clinical assessment of risk.

The generic risk assessment tool assessed in this article is the Juvenile Sanctions Center's (2002) MRAI. It was chosen as an example to serve the purposes of this case study not because it was the only tool of its kind. However, it was felt to be particularly suitable to the NYC context both because of its pedigree and its design. First, the tool was developed through the work of the National Council on Crime and Delinquency, a key agency that has been involved in developing juvenile risk assessments across different U.S. jurisdictions for many years. The tool actually draws on factors that occurred repeatedly across separate research validated juvenile risk assessment tools in 13 jurisdictions across the United States (a group of items that closely resemble those highlighted in research discussed above).[2] Published research on two of the validations, by LeCroy et al. (1998) and Johnson et al. (2002), highlights robust evaluation methodologies tracking juvenile re-referral rates during 12-month follow-up periods. As such, the generic tool should be strong: It represents the accumulated validation experience from a diverse range of settings, focusing on factors common across these settings while disregarding factors found only infrequently.

The tool was also chosen as a comparison because of its functional similarity to the NYC tool and because of its relative simplicity to apply retrospectively. Importantly, it is explicitly targeted at those at the "front-end" (p. 1) of the juvenile justice system and includes among its possible uses a role in "structuring dispositional decisions for adjudicated youth" (p. 6)—closely resembling our focus in NYC. Also, like our NYC tool, it produces a single (risk) score that relies on a handful of scored items.

The research asks,

- Are there any obstacles to the implementation of the MRAI generic tool in the NYC juvenile justice context?
- How well does the nonvalidated MRAI perform compared, respectively, to a locally developed and validated tool (the NYC tool) and to clinical judgment?
- How well does the MRAI perform against these benchmarks after it has been validated locally?
- What lessons can we therefore learn about the value of generic instruments applied in contexts different from which they were developed and validated?

Background: Data Collection and Development of the NYC Tool

We first highlight the steps we took to collect data for the development of the NYC tool and the methods we used to create the tool as a context to the testing of the MRAI.

Data Collection

We collected data on a total of 730 juvenile delinquents who received dispositions in the New York City Family Court system during the spring of 2000. In New York, a juvenile delinquent is a person between the ages of 7 and 15 (at the time of the offense) who is charged with committing an act that would constitute a crime if committed by an adult.[3] Study participants were identified from family court calendars for the key juvenile delinquency parts all five NYC boroughs and included juvenile dispositions found for each day in April, May, and June of 2000. Files were retrieved, on average, for 87% of identified cases (the retrieval rate in each borough ranged from 80% to 93%).[4]

The design of our coding schedule was heavily influenced by the literature on juvenile recidivism and the structure of existing risk assessment tools, as described above. Variables collected included demographic information, charges and justice processing information, legal history, family circumstances, school attendance and performance, community and peer relationships, drug and alcohol use, mental health, and victimization history. However, the items on the data collection tool relevant to recidivism sometimes evolved beyond the categories received from other tools and literature. Pilot work, which involved reviewing files and refining coding protocols, helped establish locally relevant versions of measures or locally available proxies for measures that could not be directly measured from available information. This meant that measures of concepts highlighted by the existing literature and assessment tools evolved to take a locally relevant form that could be applied in the NYC context. This issue is discussed in more detail later in the article.

One of the main sources of information in the files was the probation department's "investigation and recommendation" reports, written by probation officers, recommending the most appropriate disposition in each case and submitted to family court judges. They contain fairly comprehensive legal, social, and psychological information for delinquent youth. These reports were complemented by a series of other documents sometimes available on file, including probation intake reports, mental health reports, assessments for intensive probation, school records, court petitions, and arrest reports.

To measure recidivism, we chose rearrest (of any kind) because of its greater frequency compared to other measures, such as reconviction. The measure is consistent with a number of other juvenile risk assessment studies focused on recidivism that have focused either on rearrest (Marczyk et al., 2003; Rodríguez-Labarca & O'Connell, 2005) or the closely related re-referral to juvenile justice intake (Johnson et al., 2002; LeCroy et al., 1998), which tends to follow after a juvenile is arrested. We did not wish to include probation or other supervision violations because this would have introduced biases as, unlike arrests, they are dependent on being under supervision while in the community. Nor did we include summonses. We obtained arrest data (both juvenile and adult) from a comprehensive database maintained by NYC's Criminal Justice Agency on all police arrests. However, to control for time spent by juveniles out of the community in our outcome measure (because of institutional confinement), we also needed data on their patterns of confinement after disposition. Data from the New York State Office for Children and Family Services, the agency responsible for children in placement by the juvenile justice system, and data from the city's Department of Juvenile Justice, responsible for detention of youth, were also added to our sample cases. Based on these data, we found that 97% of the sample had spent a full 18 months in the community, or at least had been rearrested before the end of this 18-month period had passed, during a 3-year follow-up period. Thus, rearrest within 18 months spent in the community (across a 3-year follow-up period) became our key recidivism measure. Within this time, about half our total sample (48%) was rearrested and about one third (34%) was rearrested for a felony offense. Table 1 provides more details.

Tool Construction

To create and validate an assessment tool, Clear et al. (1988) highlight a number of steps that we closely followed. These include the development of a study sample containing measures and potential correlates of failure. The process further involves the division of that sample into a "construction" subsample, on which to derive a statistical model that provides a basis for the assessment tool, and a "validation" subsample, on which to test the predictive reliability of the model, to offset the possibility of shrinkage of statistical power when a model is applied to a different sample from the one with which it was derived (S. D. Gottfredson, 1987). When splitting the sample, we deliberately chose a larger construction subsample (n = 499) than a validation subsample (n = 231) because we wanted enough cases in the construction subsample to produce a robust model.[5] We used a randomly generated number to split the sample so the characteristics of the two subsamples are very similar in terms of key demographic and recidivism variables. Table 1 highlights the characteristics of the two samples.

We drew on a number of independent variables for inclusion in an initial logistic regression model that was subsequently refined through backward stepwise procedures.[6] The initial variable list was largely consistent with, but slightly more expansive than, those present elsewhere in risk assessment tools (OJJDP, 1995; Washington State Institute for Public Policy, 2004) and those highlighted by meta-analysis (Cottle et al., 2001). They included demographic variables (sex, age), school variables (attendance, conduct, attainment), legal history variables (previous arrests, previous violent arrests, previous adjudications), family (obedience to parents, parental supervision), current offense (severity, offense type), community (negative peer group, gang association), drug use, and attitude to the current offense. Most of the variables were collapsed down into binary categories following bivariate analysis with the dependent variable, with the exception of school attendance, which was a three-category variable, and age, which was included as a continuous variable (with some adjustments).[7]

Table 1
Characteristics of Estimation and Validation Samples

	Estimation Sample (%)	Validation Sample (%)
Male	80	78
16 and older (at disposition)	17	17
14-15	66	61
13 or younger	18	22
White	6	6
Black	63	63
Hispanic	29	27
Other	2	4
Prior or other arrest	42	39
Prior disposition for delinquency	15	15
Prior delinquency placement	4	4
Current felony adjudication	40	38
Current adjudication for violent or weapons offense	50	46
Placement disposition	51	49
Rearrest (18 months)	49	47
Felony rearrest (18 months)	34	34
Violent felony rearrest (18 months)	22	21
2+ rearrests (12 months)	16	22
3+ rearrests (12 months)	9	8
n (minimum)	457	212

There are a few points to note in our initial variable list. First, we excluded variables (including race, geography, family structure, and experiences of abuse and neglect) where we, along with NYC stakeholders, felt that inclusion would be discriminatory, unfair, or unpalatable among practitioners. We did, however, include gender, which other tool developers have often excluded for ethical reasons (LeCroy et al., 1998; OJJDP, 2005) because of its very profound associations with recidivism.[8] In addition, we did not include age at first referral, a very important predictor in the literature, principally because of its high correlation with current age, which we did include (which made it somewhat redundant). We refer in more detail to this issue later in the article.

Secondly, we had some missing data (averaging about 8% in the final model variables but ranging as high as 24% for the school misconduct variable), but we did not want to lose cases to missing data—we felt this would be an unrealistic approach for probation officers to take in the field. We therefore grouped missing values with appropriate nonmissing values according to their predictive characteristics measured through bivariate analysis.[9]

Table 2
Logistic Regression Model to Develop Locally Derived New York City Tool (and Version Without Sex)

Locally Derived Items	Full Model	Model Without Gender
Male	1.38**	
School attendance		
Absences (10%-90% present)	0.17**	0.06*
Rarely attends or not enrolled	1.14**	0.82*
Intervention by school for misconduct		0.41
Age—standardized from 0 (young) to 1 (old)	1.06*	1.21**
Prior or other arrests	0.60**	0.71**
Felony A or B adjudication	2.39*	2.52*
Substance use	0.46*	0.43*
Negative peer group	0.52*	0.55*
Constant	–3.23**	–2.32**
Nagelkerke R^2	.25	.19

Note: $n = 483$.
$*p < .05$. $**p < .01$.

Table 2 provides the logistic regression results that formed the basis of our risk score development. We present here our main model that was actually used in the development of the NYC risk assessment tool. The table also provides an alternative model that does not use a sex variable for comparative purposes with the MRAI, which excludes gender. In both cases, backwards stepwise log likelihood logistic regression was used, and in both cases the models retained seven predictor variables.

The most powerful predictor of both models is severity of offense, though it only applies to a small number of cases (14 out of 15 of cases in the estimation subsample adjudicated as Felony A or B were rearrested). Gender too is a very powerful predictor, as we can see in the first model. Other important variables include age, school attendance, and, to a lesser extent, previous arrests, peer group, and substance abuse. School misconduct also is important for the second model, once gender is not included (the variable is significant at $p < .10$). These predictors show both overlaps with and differences from variables found in other literature already reviewed (Cottle et al., 2001; OJJDP, 1995) For example, the NYC tool contains similar variables for about 4 out of the 6 core variables found in half or more instruments reviewed by OJJDP (1995): school attendance, prior arrests, substance abuse, and negative peers. By contrast, the NYC tool did not include the core variables of age at first referral and family functioning. Interestingly, those variables with the greatest weights on the NYC tool (offense severity, sex, and age) are actually absent from the core list of variables from other jurisdictions. Although gender has often been excluded in other tools on reasons of principle, offense severity and age presumably lack predictive power in other contexts despite their importance in NYC.

Coefficients from the final logistic regression model were each multiplied by 10 and rounded for each to produce the scoring system that was incorporated into the NYC tool. The full score for any individual, therefore, represents a direct index of the probability of rearrest. For the purposes of this article only, we also produced a second scoring system based on our logistic regression model that excluded sex as a predictor.

Tool Validation

To test both model scoring systems, we applied them to the validation subsample and, for each scoring system, divided the validation sample into approximate quartiles, in increasing order of risk. Both produced highly significant ($p < .01$) differentiation among quartiles, though the full model was better overall. The full model NYC tool produced recidivism rates for quartiles of 21%, 41%, 59%, and 71%, respectively, and the model without sex produced recidivism rates across quartiles of 22%, 49%, 46%, and 67% (slightly less discriminatory in the middle range). Figure 1 visually displays these results alongside the validation results for the MRAI tool.

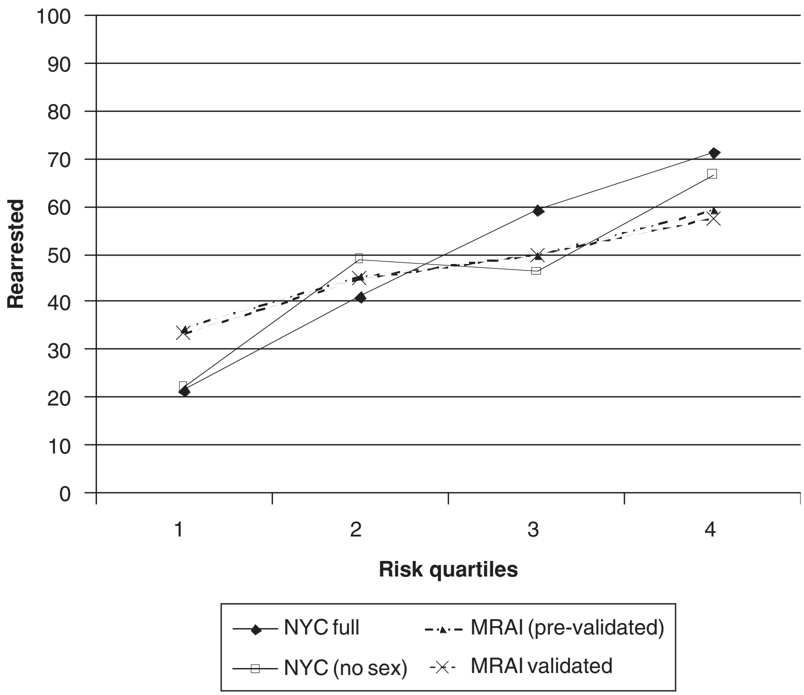
Fitting the Generic MRAI Tool to NYC Data

In this section, we discuss the application of the MRAI generic tool to the NYC data, which forms our core analysis for this article.

Table 3 shows the questions and points available for the nonvalidated MRAI. Each of the tool items typically encompasses a number of subcategories associated with different subscores. For example, total number of referrals is grouped into three categories of one referral (0 risk points), two or three referrals (1 point), and four or more referrals (3 points). Peer relationships covers four categories: "friends provide positive influence" (0 points), "some delinquent friends with negative influence" (1 point), "most friends are delinquent; strong negative influence" (3 points), and "gang member/associate" (4 points).[10]

Although in theory applying the scoring system to the NYC data would be a straightforward exercise, in practice there were some difficulties.

Figure 1
Quartile Recidivism Rates for the Four Risk Scores From the New York City (NYC) and Model Risk Assessment Instrument (MRAI) Risk Tools in Validation Subsample



Note: $n = 225$.

Many of them strike at the heart of the challenge of introducing a generic tool into a new and different setting from the context in which it was developed. They are discussed in turn below.

Administrative Limitations in Available Data

Limitations in locally available data set some practical limits to our ability to complete the MRAI items, at least following a strict interpretation. This meant we were forced to make some conceptual compromises and use some proxy measures that did not exactly map the item specification. To

Table 3
Questions for Model Risk Assessment Instrument Risk Assessment and Their Risk Scores

Question	Allocated Points
Peers	4
Age at first referral	3[a]
Total number of referrals	3
School discipline or attendance	3
Substance abuse	2
Number of out-of-home placements	2
Parental supervision	2
Referrals for violence or assault	1
Parent or sibling criminality	1
Victim of abuse or neglect	1
All questions	22

Note: On the New York City (NYC) population, only 2 points are available for age at first offence. This is because juvenile delinquents only go to age 15 in NYC. Sixteen and older would be associated with an extra risk point in other contexts where 16-year-olds qualify as juvenile delinquents.

some extent, these limitations were a product of the choices we had already made as researchers in our data collection protocol (we had not collected data with the MRAI in mind). However, more substantially, they reflected limitations in the data that were available to probation officers at the point of their investigation—which our research data closely mirrored. These tended to reflect the quality and character of records available to probation officers, in particular those passed to them by other agencies such as education, child welfare, and justice agencies.

For example, the MRAI includes an item measuring history of abuse and neglect. It reads: "Victim of child abuse or neglect (based on report to child welfare agency, substantiated or not)." It accords a score of 1 to a yes and 0 to a no. Based on the information routinely available to probation officers from the NYC child welfare agency, we could identify open or closed child welfare cases relating to the youth's family—and this information was captured in our data set. However, it was not clear from this information whether these cases were for abuse or neglect or other child welfare issues. Nor did we know whether they were related to the young person being assessed or another family member. This posed a problem that we resolved through using an imperfect compromise measure. We chose to include any child welfare case that was open, closed, or with a finding as an indicator of child abuse and neglect—knowing that this measure would be wrong in an unknown proportion of cases.

In some instances, we had to make minor adaptations to items more directly because of choices we had taken in our data collection protocol. Yet even in most of these cases, it remains doubtful whether the full MRAI data would have been consistently available on file had we tried to collect them—our measures largely reflecting the character of data readily available. For example, the MRAI tool asks about "school discipline/attendance during the prior 12 months," whereas we had collected the data for the 3 months prior to arrest—which we used to fill out the MRAI. In practice, we felt these two measures were likely highly correlated and felt reasonably comfortable relying on the 3-month measure. However, discipline and attendance data are rarely available to a probation officer for a whole year.[11]

Missing data also presented a problem: The average rate of missing cases across measures was 7% and ranged up to 17% for the item measuring peer group associations. As with the development of the NYC tool, we did not want to ignore cases where we had missing data because this would be an unrealistic approach to completing a risk assessment instrument in the NYC (or probably any) context. Yet the MRAI provides no guidance on how to resolve the problem. We therefore developed an approach that involved substituting a guessed value based on the group norm. Sometimes this was simply the modal value for the remaining sample on a question. In other cases, we had some limited information that allowed us to focus on a subset of cases. For example, when scoring someone on his or her number of referrals, we often knew that he or she had at least one prior referral, but we did not know how many. We therefore took the modal value only for those with at least one prior referral (the modal value in this example was 1).

Ethical Considerations

The development of the NYC tool had already highlighted to us that ethical sensibilities among stakeholders in different settings likely vary. However, ethical dilemmas were more acutely raised through the simulated implementation of the MRAI in the NYC context. For example, the MRAI tool included a variable on abuse and neglect, even though we had deliberately excluded this type of variable on ethical grounds in the development of the NYC tool. Similarly, an MRAI item that asked about parent or sibling criminality (specifically whether they had been incarcerated or on probation in the past 3 years) was one that we felt would have clashed with the sensibilities of stakeholders we worked with within NYC and hence would have presented a problem had our attempts at implementation of the MRAI been real rather than simulated.

Examining the Predictive Efficacy of the MRAI

In this section, we look in detail at the performance of the MRAI tool based on the NYC data. In doing so, we carry out a validation exercise in which we simulate the approach recommended by proponents of generic tools. That is, we assume the tool has been used as a basis for data collection (in our case it has actually been fitted retrospectively to existing data), and we test whether the tool really is predictive. On the basis of this, we make adjustments to the scoring of the existing MRAI items to maximize its predictive efficacy. In addition, we analyze the prevalidated and validated MRAI by assessing their performance against two comparative reference points: the predictive validity of the homegrown NYC tool and the predictive validity of the clinical judgments of probation officers.

Examining MRAI Items

To take a look at how the individual MRAI items behave as predictors in the NYC data, we first entered each of the separate item scores into a logistic regression model of the NYC estimation subsample. The overall score of an item is taken as a continuous variable, even though some items have different values assigned for each of a number of levels (e.g., for substance abuse, "no problem" is given 0 points, use "sometimes" is given 1 risk point, and use "frequently" is given 2 risk points). For illustrative purposes, we have also standardized the range of each item to between 0 and 1 so that the coefficients indicate the relative importance of each item. The first model in Table 4 illustrates what we found.

The second model in Table 4 details an optimized predictive model based once again on the estimation subsample. The model simulates the outcome of a validation exercise, using the data that would have been produced if the MRAI had been used on a cohort of youth in NYC, to produce the optimum local scoring system. The model was produced by excluding nonpredictive variables through a backwards stepwise procedure. Also, in creating the optimized model, original item scores provided by the original MRAI tool were not used, and instead items were introduced to logistic regression as categorical variables to optimize their predictive value. Some categories were collapsed

Table 4
Logistic Regression Models Using Model Risk Assessment Instrument (MRAI) Predictors

MRAI Items Prevalidated		Optimized Model of Validated MRAI Items	
Original Items (Standardized 0 to 1)		Adjusted Items (All Comparisons Standardized 0 to 1)	
Age at first referral (1 is younger than 0)	–0.60**	Age at first referral (1 is younger than 0)	–0.68**
Total referrals	1.29**	Total referrals (1)**	
		2 or 3	0.95**
		4 or more	1.10**
Referrals for violence or assaults	–0.19		
Prior out-of-home placements	–0.71		
School discipline or attendance	0.81*	School discipline or attendance	0.75**
Substance abuse	0.37		
Peer relationships	0.77**	Peer relationships*	
		Some delinquent or negative	0.62
		Most delinquent or negative	0.72*
		Gang associate	0.87**
Victim of abuse or neglect	0.82**	Victim of abuse or neglect	0.76**
Parental supervision	–0.06		
Parent or sibling criminality	–0.12		
Constant	–1.29**		–1.33**
Nagelkerke R^2	.16		.18

Note: n = 483.

together where they added no extra explanatory value. The scoring system based on this model is hereafter referred to as the validated MRAI.

What is striking from Table 4 is that many of the items do a poor job predicting recidivism despite their importance in the literature (Cottle et al., 2001; OJJDP, 1995). Referrals for violence or assault, prior placements, parental supervision, and parent or sibling criminality do nothing to predict recidivism (and even have negative coefficients) in a multivariate model. It is possible that parental supervision and parent or sibling criminality are difficult to characterize and perhaps unreliably measured (relying as they do primarily on interviews between the probation officer and the youth and parent), which helps explain their lack of explanatory power. However, referrals for violence and previous placement episodes are typically based on simpler and more systematically available information, suggesting that these two variables are simply not predictive in a multivariate context. Substance abuse has a positive coefficient that echoes its predictive power in the NYC risk tool models, though in this case it falls short of significance set against the other predictors included in the model.

Particularly notable is the fact that age at first referral is statistically significant but predicts recidivism in entirely the opposite fashion of that anticipated by the MRAI and indeed other juvenile risk assessments and juvenile recidivism literature generally. That is, those who have a younger age of first referral are less likely to recidivate than those who are older at their first referral. One possible explanation is that a narrow and young age spectrum of juvenile delinquents in NYC undermines the common association between age at first offense and recidivism. Youth in NYC only qualify as juvenile delinquent up to age 15, in contrast with many other jurisdictions, where they are older. Indeed, in all the jurisdictions on which the MRAI instrument was originally based, juvenile status extends at least to 16 and in most cases to 17. It is possible that the narrower age distribution in the NYC youth leaves less variation among offenders in the cohort than the age at first arrest variable normally signifies. In the NYC data, age at first referral also closely maps current age (the Pearson correlation coefficient between these two measures is .69), perhaps because this is a relatively young cohort, which may make it a better indicator of where people are in their offending careers (because it closely resembles their current age) than what type of offender they are. A different kind of explanation may relate to distinctive institutional practices in the NYC setting. We noted in the course of our research that a large number of arrests take place within schools (though we did not quantify this). It could be that children who tend to go to school rather than play truant may be disproportionately arrested precisely because they remain at school. This could mean that those getting arrested younger are also those more likely to have been in school. This would be precisely the group who were less prone to delinquency in the longer term because of their attachment to institutions.

MRAI Predictive Performance Compared With NYC Tool Scores

To test the MRAI in both its prevalidated and validated forms, we applied it to the validation subsample of the NYC data. In creating quartiles for the two measures, we found that both were predictive of recidivism but did not achieve high levels of statistical significance on our validation sample. The prevalidated MRAI only achieved significance at the $p < .10$ level, producing quartiles associated with recidivism of 35%, 45%, 50%, and 59%, respectively. Surprisingly, the validated MRAI produced almost identical outcomes ($p < .10$) of 33%, 45%, 50%, and 58%.

Figure 1 displays the NYC tool measures along with the MRAI measures, based on quartile recidivism rates in the validation subsample. In interpreting the graph, a steeper slope for any of the lines indicates more discriminatory power. It is clear from this graph that both NYC tools (with and without the sex variable) perform better than the MRAI tool, regardless of whether it has been validated (though the NYC no-sex tool is weak in the middle risk score range). It is also clear that the validated MRAI shows no improvement on the prevalidated version.

It is necessary to ask why the NYC tool continues to perform better than the MRAI, even after the latter has been validated, and indeed why the validated MRAI tool offers little improvement over the prevalidated version. A review of Tables 2 and 4 suggests that the main explanation relates to the role of demographics. The NYC tool's reliance on age and sex (and the second NYC tool's reliance still on age) adds substantially to its predictive power. This is true despite the fact that a crude proxy for age is effectively used by the validated MRAI in the form of the correlated age at first referral—which overall appears as a less powerful predictor. Overall, it seems as if, within NYC, demographic differences dominate others in predicting recidivism, and these are not adequately captured by the MRAI items. Thus, even after validation, the MRAI does not improve substantially because it is not adequately capturing these items.

There are other differences between the tools which, although important to reflect on, apparently did not in this instance help explain the differences in predictive power between the tools. Notably, the NYC tool includes a variable relating to the severity of the current offense—a powerful predictor not covered by the MRAI nor found in the literature on juvenile recidivism generally. Although this enhances the NYC tool, in practice it only compensates for omitting (because of local ethical judgments) a variable relating to abuse and neglect that is found on the MRAI and that is also an important predictor (if we remove the current offense severity variable from the NYC tool model, it reduces the Nagelkerke R^2 from .25 to .23; however, adding the abuse and neglect variable into the model raises its value back to .25).

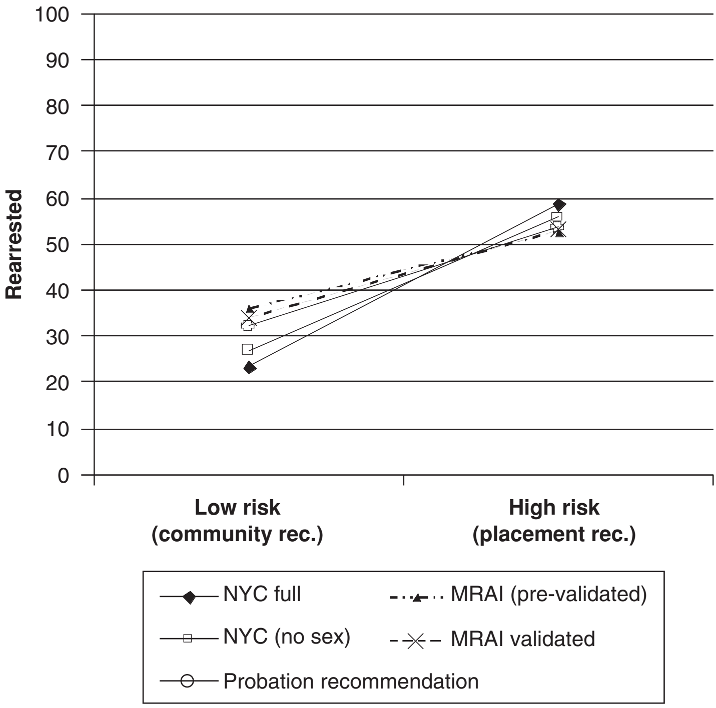
We were also interested to see whether the poorer predictive power of the MRAI might be attributable to differences in the specification of similar items between tools. For example, both the validated MRAI and the NYC tool have items on school functioning issues, previous referrals, and peer groups, which have some differences in specification. However, we found, overall, that this did not help explain differences either (replacing the three NYC items with the three similar MRAI items in the NYC tool model actually increases the Nagelkerke R^2 from .25 to .26, suggesting that the MRAI variables are no worse predictors, overall, and perhaps may be better).

That said, some interesting lessons that are relevant to the application of generic risk assessment tools did emerge from the latter comparison. Notably, the school attendance variable used in the NYC model is a better predictor than the generic MRAI school discipline or attendance variable when we look at them in isolation (if we substitute the NYC school attendance item into the MRAI validated model in place of the original school functioning item, we find that it also improves the Nagelkerke R^2 from .18 to .19). Furthermore, we suspect that this reflects specific strengths and weaknesses of locally available data: School attendance is the best documented of all school information on NYC probation officer files because it is based on computerized records of school attendance consistently provided by the schools. The NYC tool categories read: "Attends classes regularly (at least 90% of the time)," "Regular absences and/or regular cutting (in class between 50 and 90% of time)," "Present in class between 10% and 50% of the time," "Not enrolled/rarely attending (less than 10% of time)." By contrast, the MRAI item wraps multiple school concepts into one variable (attendance, behavior, and school intervention) and in doing so relies on data that are not as consistently or reliably recorded on NYC files. The MRAI categories read: "Enrolled, attending regularly, no suspensions; or, graduated or GED," "Some truancy; suspended 1-2 times; considered somewhat disruptive," "Major truancy or dropped out; suspended 3+ times; considered seriously disruptive." This suggests that it may be important to focus on stronger locally available data when adapting item constructs to local circumstances.

MRAI Predictive Performance Compared to Clinical Judgment

As we have already noted, our ability to make a comparison with clinical judgment is qualified by the fact that our benchmark comparison is not a pure subjective assessment of risk. Instead, we rely on probation officers' recommendations, which reflect concern in significant part with risk but also with welfare and perhaps bureaucratic considerations, which may dilute the measure. Figure 2 provides a measure of recidivism rates of the probation

Figure 2
Recidivism Rates, According to Probation Officer (PO) Recommendation, Compared With New York City (NYC) and Model Risk Assessment Instrument (MRAI) Scales in Validation Subsample



Note: n = 203. Split to match distribution of PO recommendations.

officer according to whether they recommend placement (considered here high risk) or community sanctions (considered here low risk). To make a comparison with this judgment, the various NYC and MRAI scales have each been divided to match, as closely as possible, the distribution the probation officer recommendation (62% high risk, and 38% low risk).[12] Once again, steeper lines are indicative of more discriminatory power.

Figure 2 shows that both NYC scores (with and without sex) trump the probation officer recommendation as a predictor of recidivism, as we would expect. However, the probation officer recommendation, in turn, trumps the

MRAI both prior to and after it has been validated. This is surprising and less than reassuring. We would expect the MRAI to outperform the probation officer's judgment, at least after it has been validated and adjusted. This failure is particularly damning given that, in this case, our measure of probation officer judgment, if anything, understates true clinical risk assessment. Likely, the reasons for this are those documented above: The MRAI fails to include variables that are particularly relevant to the NYC context.

Conclusions

This article has explored the practical challenges that would likely be associated with introducing a generic juvenile risk assessment tool—the MRAI—into NYC. It has highlighted some important issues that could usefully guide researchers and practitioners in jurisdictions that may be contemplating such an approach and some broader theoretical lessons.

Importantly, this article has put to the test the proposition, often suggested or implied by proponents of generic tools (Juvenile Sanctions Center, 2002; OJJDP, 1995; Orbis Partners, 2006), that taking a generic juvenile risk assessment tool in a prevalidated form and applying it in a new context probably does a reasonable job of predicting risk in that context and that it will—after validation and adjustment—live up to the expectations of actuarial risk assessment (i.e., it will predict recidivism with greater accuracy than clinical judgment). In this case, that promise has not been fulfilled.

A key observation of this research is that local NYC data are not always available in the precise form required by the generic MRAI tool, so adaptations and proxies were required to make the tool fit the local context. In NYC, this reflects real limitations to the kind of data that are available to probation officers when they are required to assess juveniles and in turn reflects the character of records passed to them by others (e.g., education, child welfare, and justice agencies). There were also cases where information is simply missing—a normal occurrence in the imperfect environment in which probation officers work—and this too presents challenges for fitting the tool to the local environment. We also speculate that local sensibilities about what kinds of information are appropriately used in an assessment tool might also have clashed with the requirements of the imported tool were we to have tried to implement it for real.

Importantly, analysis shows that the MRAI generic tool applied to locally available information does not match the predictive power of a locally developed NYC risk assessment tool, even after the former has been validated and adapted in the local context. Indeed, local validation gave rise to few gains in predictive power over the prevalidated version. More disturbingly, the generic tool, in both pre- and postvalidated forms, performs less well than the probation officers' clinical judgment as embodied by their dispositional recommendation court, a measure that probably understates their true ability to assess risk.

These shortcomings appear to reflect the fact that predictor variables used by the generic MRAI tool—reflecting as they do the conventional literature on juvenile recidivism (e.g., Cottle et al., 2001; OJJDP, 1995)—have some overlap with, but also important differences from, locally effective predictor variables. Notably, the MRAI excludes age, gender, and severity of offense, which it turns out are locally powerful predictors. On the other hand, the MRAI (in its prevalidated form) includes variables that have no predictive value or that predict recidivism in the opposite direction to that expected (age at first referral): Younger age at first referral in NYC is actually a marker of lower rather than higher recidivism, contrary to that assumed by the MRAI and the juvenile recidivism literature more generally. This is a curious finding that may reflect distinctive characteristics of the NYC juvenile delinquent population.

Like all case studies, it is not clear whether these findings would generalize to other places and situations. However, case studies can alert us to some important theoretical concerns—and this study is no exception. First of all, the findings suggest that a tendency to look toward a universal model of juvenile recidivism may actually obscure some of its important local characteristics, to which risk assessment needs to pay attention. To put this in practical terms, a reliance on only a set of core variables that predict recidivism across settings—such as those identified by meta-analysis (Cottle et al., 2001) or those that tend to occur already on validated tools (OJJDP, 1995)—may actually deflect attention from locally important, and perhaps idiosyncratic, predictors. These might reflect legal differences (e.g., the legally defined age group of juvenile delinquents), but they might also relate to local institutional arrangements, such as the way in which the local schooling environment intersects with the juvenile justice system.

A further theoretical issue is the issue of item validity and reliability. In addition to examining the predictive power of risk items, much research is devoted to the operationalization of item constructs to enhance validity and reliability, for example, focusing on the interrater reliability of measures, that is, the extent to which different raters come to the same conclusions about the same cases (e.g., Austin, Coleman, Peyton, & Johnson, 2003; Lowenkamp, Holsinger, Brusman-Lovins, & Latessa, 2004; Schmidt, Hoge, & Gomes, 2005). However, fewer studies acknowledge or examine variations in reliability and validity across settings and jurisdictions—and in particular how different institutional contexts may provide better or worse informational resources on which raters can form judgments. This seems to be an area that the risk assessment field could significantly benefit. Once again, this may point theoretical development away from a rigid adherence to universalized item constructs and toward locally specific adaptations rooted in local practices and informational resources.

On a practical level, our findings caution us against the wholesale uncritical adoption of generic risk assessment tools into new environments, even with a commitment to downstream validation. This may still leave a jurisdiction unable to improve on clinical judgment, even after validation work has been completed. We suggest a couple of precautions to help avoid this outcome. First of all, taking the opportunity to capture a broader range of data in a prevalidation phase than a single tool provides might help—perhaps drawing on the hunches of local justice officials rather than the academic literature alone. Second, practitioners should avoid a blind adherence to generic tool items. Not only may it be unfeasible to complete the items with available information, but even with some available data, the generic measures may not always reliably get at an underlying concept of interest. Instead, efforts may be best directed to adapting generic tool items to locally available information and practices—being mindful of which data, approximating concepts of interest, are most reliable.

Notes

1. It is important to note in reading this article that there has been no practical need for a generic tool in the New York City (NYC) context given the development of the NYC local tool. Rather, this article uses data generated by the NYC research to retrospectively examine both likely implementation issues and predictive efficacy of a generic tool, had one been introduced.

2. These settings include Arizona, Cuyahoga County (Ohio), District of Columbia, Indiana, Maryland, Michigan, Missouri, Nebraska, New Mexico, Oklahoma, Rhode Island, Travis County (Texas), and Virginia.

3. The sample of juvenile delinquents does not include the most serious offenders (termed juvenile offenders in New York) who commit 1 or more of 15 specified felonies (e.g., murder) and whose cases are waived to adult criminal court.

4. We should acknowledge two caveats about the representativeness of this sample. First, we know, in practice, that a few juvenile cases were processed in other parts that were not specialized in juvenile delinquency. Based on our conversations with court officials, however, we came to believe that this is true in only a small minority of cases—though we could not actually quantify

it. A second constraint is that the 3-month period in question (at the end of the school year) may be affected by seasonal variation, which means it may look slightly different from a sample drawn from a whole year. We have no reason to believe the differences across the year are profound, but insofar as there is variation, the period we sampled is probably more typical of the year in general than, say, a period when children would have been on vacation from school.

5. Clear, Wasson, and Rowland (1998) state that an estimation subsample should aim to include at least 50 cases for each predictor variable. This meant our subsample would provide sufficient statistical power to support a model with up to 10 predictor variables.

6. Various other approaches have been taken to produce scoring systems for risk prediction (S. D. Gottfredson, 1987). These include the Burgess method, in which predictor variables from bivariate correlations are each given a single point, with a final risk score representing the sum of these scores (Burgess, 1928). Other approaches rely on applying a standard unitary score for predictor values based on regression techniques, regardless of coefficient size (Wainer, 1976). Others have also used clustering techniques to differentiate different categories of offenders (e.g., Brennan, 1987). In practice, the research literature suggests there is little difference in predictive power among these and other methods (S. D. Gottfredson & Gottfredson, 1979; Simon, 1971; Tarling & Perry, 1985).

7. We examined age as a categorical and continuous variable, but although the variable was always significant, individual binary variables representing different age ranges, as part of a categorical variable, typically were not. We therefore chose to include it as a continuous variable, with some modifications. Importantly, in our bivariate analysis of age, we found that, despite an increase in recidivism with age, the relationship between the two variables was slightly *S* shaped, with recidivism rates level for 11 years and younger and also for 15 years and older. For this reason, the continuous variable was recoded so that all cases older than 15 years were represented as 15 years, and all cases younger than 11 years were represented as 11 years.

8. Omitting gender would have led to many girls with low risk of recidivism being recommended for institutional confinement because of their misleading assessment as high risk.

9. For example, for 17% of cases, we had missing school attendance information. We gave these cases the same variable value as those who attended school more than 10% of the time but less than 90% of the time because their recidivism rates were very similar: 44% of the missing group were rearrested within 18 months compared to 47% among this category of truants.

10. In practice, these scales had some negative values. In the article, we describe the lowest values, whether negative or zero, as having no points, and the other point allocations are described relative to the lowest value. This means that if a question includes a category of –1 and a second category as 1, here we report this as 0 and 2.

11. Similarly, with out-of-home placements, we were not able to distinguish in our data between separate placement episodes. Instead, we were able to count placements of different types (i.e., child welfare placement, state placement in state facilities, state placement in contract facilities), which we felt was likely to be a close approximation of the number of placements. Once again, however, systematic data that counted different placement episodes were rarely available on file, and accounts of placement episodes with child welfare were primarily anecdotal rather than based on systematically recorded data. Again, it is unlikely had we gone back to the original files that we would have been able to fill out the Model Risk Assessment Instrument (MRAI) with much greater precision than we could from our research data set.

12. In practice exact cutpoints varied a little because the distributions of each scale score did not allow being cut precisely. Thus, the high-risk category ranged from 63% of cases (in the NYC no-sex tool) to 56% in the prevalidated MRAI.

References

Andrews, D. A. (1989). Re-arrest is predictable and can be influenced: Using risk assessments to reduce re-arrest. *Forum on Corrections Research, 1,* 11-18.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*, 7-27.

Ashford, J. B., & LeCroy, C. W. (1990). Juvenile recidivism: A comparison of three prediction instruments. *Adolescence, 98*, 441-450.

Assessments.com. (2006). *Assessment catalog*. Retrieved June 7, 2006, from http://www.assessments.com/purchase/category.asp?cid=3&cl=2

Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument.* Washington, DC: George Washington University.

Baird, S. C. (1984). *Classification of juveniles in corrections: A model systems approach*. Madison, WI: National Council on Crime and Delinquency.

Bonta, J. (1996). Risk-Needs Assessment and Treatment. In A. T. Harland (Ed.), *Choosing correctional interventions that work: Defining the demand and evaluating the supply* (pp. 18-32). Thousand Oaks, CA: Sage.

Brennan, T. (1987). Classification methods in criminology. In D. M. Gottfredson & M. Tonry (Eds.), *Prediction and classification: Criminal justice decision making, Vol. 9, Crime and justice: A review of research* (pp. 201-248). Chicago: University of Chicago Press.

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* (pp. 221-234). Springfield: Illinois State Board of Parole.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review*, *17*, 199-228.

Carter, R. M., & Wilkins, L. T. (1967). Some factors in sentencing policy. *Journal of Criminal Law, Criminology, and Police Science, 58*, 503-514.

Clear, T., Wasson, B. F., & Rowland, J. (1988). Statistical prediction in corrections. *Research in Corrections, 1,* 1-52.

Cottle, C. C., Lee, R. J., & 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,* 1644-1673.

Flores, A. W., Travis, L. F., & Latessa, E. J. (2004) *Case classification for juvenile corrections: An assessment of the Youth Level of Service/Case Management Inventory (YLS/CMI), Executive summary* (98-JB-VX-0108). Washington, DC: U.S. Department of Justice.

Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review*, *22*, 813-839.

Gavazzi, S. M., Slad, D., Buettner, C. K., Partridge, C., Yarcheck, C. M., & Andrews, D. W. (2003). Toward conceptual development and empirical measurement of global risk indicators in the lives of court-involved youth. *Psychological Reports, 92,* 599-615.

Gendreau, P., Goggin, C., & Little, T. (1996). A meta-analysis of the predictors of adult offender recidivism: What works? *Criminology*, *34*, 575-607.

Glaser, D. (1955). The efficacy of alternative approaches to parole prediction. *American Sociological Review*, *20*, 283-287.

Glaser, D. (1962). Prediction tables as accounting devices for judges and parole boards. *Crime & Delinquency*, *8*, 239-258.

Glueck, S., & Glueck, E. (1950). *Unraveling juvenile delinquency*. New York: Commonwealth.

Gottfredson, D. M., & Beverly, R. F. (1962). *Development and operational use of prediction methods in correctional work. Proceedings of the Social Statistics Section*. Washington, DC: American Statistical Association.

Gottfredson, M. R., & Gottfredson, D. M. (1984). *Decision making in criminal justice: Toward a rational exercise of discretion*. New York: Plenum.

Gottfredson, S. D. (1987). Prediction: Methodological issues. In D. M. Gottfredson & M. Tonry (Eds.), *Prediction and classification: Criminal justice decision making, Vol. 9, Crime and justice: A review of research* (pp. 29-33). Chicago: University of Chicago.

Gottfredson, S. D., & Gottfredson, D. M. (1979). *Screening for risk: A comparison of methods*. Washington, DC: National Institute of Corrections.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency, 52,* 178-200.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19-30.

Hawkins, J. D., Herrenkohl, T. L., Farrington, D. P., Brewer, D., Catalano, R. F., & Harachi, T. W. (1998). A review of predictors of youth violence. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 106–146). Thousand Oaks, CA: Sage.

Herrenkohl, T. L., Maguin, E., Hill, K. G., Hawkins, J. D., Abbott, R. D., & Catalano, R. F. (2000). Developmental risk factors for youth violence. *Journal of Adolescent Health, 26*, 176–186.

Hodge, R. D., & Andrews, D. A. (1999). *The Youth Level of Service/Case Management Inventory and manual.* Ottawa, Ontario, Canada: Carleton University, Department of Psychology.

Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology*, *68*, 203-211.

Hood, R. (1966). A study of the effectiveness of pre-sentence investigations in reducing recidivism. *British Journal of Criminology, 6*, 303-310.

Johnson, K., Wagner, D., & Matthews, T. (2002). *Missouri juvenile risk assessment re-validation report.* Wisconsin: National Council on Crime and Delinquency.

Juvenile Sanctions Center. (2002). Structured decision making for graduated sanctions. *Training and Technical Assistance Program, 1*, 2.

Klein, S. P., & Caggiano, M. N. (1986). *The prevalence, predictability, and policy implications of recidivism*. Santa Monica, CA: RAND.

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.

LeCroy, C. W., Krysik, J., & Palumbo, D. (1998). *Empirical validation of the Arizona Risk/Needs instrument and assessment process.* Phoenix, AZ: LeCroy & Milligan.

Lipsey, M., & Derzon, J. H. (1998). Predictors of 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.

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., Moffitt, T. E., & Caspi, A. (1998). The development of male offending: Key findings from the first decade of the Pittsburgh Youth Study. *Studies in Crime and Crime Prevention*, *7*, 141–172.

Loeber, R., & Stouthamer-Loeber, M. (1987). Prediction. In H. C. Quay (Ed.), *Handbook of juvenile delinquency* (pp. 325-383). New York: John Wiley.

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the Level of Service Inventory Revised. *Federal Probation, 68,* 34-38.

Marczyk, G. R., Heilbrun, K., Lander, T., & DeMatteo, D. S. (2003). Predicting juvenile recidivism with the PCL-YV, MAYSI, and YLS-CMI. *International Journal of Forensic Mental Health, 2*, 7-18.

McCord, J., Widom, C. S., & Crowell, N. A. (Eds.). (2001). *Juvenile crime, juvenile justice. Panel on juvenile crime: Prevention, treatment, and control.* Washington, DC: National Academy Press.

Moffit, T. (1993). Adolescent-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.

Neubauer, D. W. (1974). *Criminal justice in middle America*. Morristown, NJ: General Learning Press.

Office of Juvenile Justice Delinquency Planning. (1995). *Guide for implementing the comprehensive strategy for serious, violent and chronic juvenile offenders.* Washington, DC: U.S. Department of Justice

Office of Juvenile Justice Delinquency Planning. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts.* Washington, DC: U.S. Department of Justice.

Orbis Partners. (2006). *Welcome*. Retrieved June 7, 2006, from http://www.orbispartners.com/frame.htm

Risler, E. A., Sutphen, R., & Shields, J. (2000). Preliminary validation of the juvenile first offender risk assessment index. *Research on Social Work Practice, 10,* 111-126.

Rodríguez-Labarca, J., & O'Connell, J. P. (2005). *Delaware juvenile recidivism: 1994-2004 Juvenile Level III, IV and V Recidivism Study.* Dover: State of Delaware, Office of the Budget, Statistical Analysis Center.

Rosecrance, J. (1985). The probation officers' search for credibility: Ball park recommendations. *Crime & Delinquency, 31*, 539-554.

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.

Schwalbe, C., Day, S., & Fraser, M. (2002). *Risk assessment in the juvenile justice system (Web report).* Chapel Hill: University of North Carolina.

Simon, F. (1971). *Prediction methods in criminology.* London: HMSO.

Tarling, R., & Perry, J. (1985). Statistical models in criminological prediction. In D. Farrington & R. Tarling (Eds.), *Prediction in criminology* (pp. 210-231). Albany: State University of New York.

Tremblay, R. E., & LeMarquand, D. (2001). Individual risk and protective factors. In R. Loeber & D. P. Farrington (Eds.), *Child delinquents: Development, intervention, and service needs* (pp. 137–164). Thousand Oaks, CA: Sage.

Turner, S., Fain, T., & Sehgal, A. (2005). *Validation of the risk and resiliency assessment tool for juveniles in the Los Angeles County probation system.* Santa Monica, CA: RAND.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 312-317.

Washington State Institute for Public Policy. (2004). *Assessing risk for re-offense: Validating the Washington State juvenile court assessment*. Olympia: Author.

Wasserman, G. A., & Seracini, A. G. (2001). Family risk factors and interventions. In R. Loeber & D. P. Farrington (Eds.), C*hild delinquents: Development, intervention, and service needs* (pp. 165–189). Thousand Oaks, CA: Sage.

Wiebush, R. G., Baird, C., Krisberg, B., & Onek, D. (1995). Risk assessment and classification for serious, violent, and chronic juvenile offenders. In J. C. Howell, B. Krisberg, J. D. Hawkins, & J. J. Wilson (Eds.), *A sourcebook: Serious, violent, & chronic juvenile offenders* (pp. 171-210). Thousand Oaks, CA: Sage.

Wright, K. N., Clear, T. R., & Dickinson, P. (1984). Universal applicability of probation risk assessment instruments. *Criminology, 22,* 113-134.

Joel Miller is a visiting professor of criminology at the University of Málaga. He has previously held senior research positions at the Vera Institute of Justice, New York, and the Home Office of the UK government. He has carried out research on police accountability, juvenile justice, probation, and crime and violence prevention. He has a PhD in sociology from the University of Surrey, England.

Jeffrey Lin is a postdoctoral scholar at the University of California, Irvine's Center for Evidence-Based Corrections. His research interests include parole supervision, offender reentry, and juvenile justice. He previously worked as a research associate at the Vera Institute of Justice, where he conducted research on juvenile justice, adult corrections, and family homelessness. He has a PhD in sociology from New York University.