Six of one, half a dozen of the other: The benefits and challenges associated with risk classification tools*

Heather L. Pfeifer Ph.D., Faye S. Taxman Ph.D., and Douglas Young M.S.

Volume 2 - No. 2 - Fall 2005

* *A previous version of this article was presented at the annual meeting of the American Society of Criminology, Atlanta, GA, November, 2001. This project is funded by the Maryland Governor's Office of Crime Control and Prevention. All opinions are those of the authors and do not represent those of GOCCP.*

Abstract

The use of risk-assessment tools in the juvenile justice system has introduced more standardized methods in how cases are processed and individual offenders are treated. Yet, their development remains extremely challenging. Despite the diverse nature of the juvenile justice population, agencies often adopt a 'one-size fits all' approach when developing these instruments, which has led to a significant proportion of misclassified youth. Consequently, these instruments have been challenged on numerous grounds. This paper presents a case study of a classification tool currently under development for the Maryland Department of Juvenile Justice to illustrate some of the challenges researchers face when developing and implementing a standardized risk classification instrument for a juvenile population. A number of recommendations are presented on how researchers can improve these instruments' predictive efficiency, and on how agencies can overcome obstacles in their implementation.

About the Authors

Heather L. Pfeifer is an assistant professor in the Division of Criminology and Criminal Justice at University of Baltimore. Address all correspondence to Heather L. Pfeifer, University of Baltimore, Division of Criminology and Criminal Justice, 1420 N. Charles Street, Baltimore, MD 21201; email: hpfeifer@ubalt.edu

Douglas Young is a research associate at the Bureau of Governmental Research [BGR] at the University of Maryland, College Park.

Faye S. Taxman is the Director of BGR and an associate research professor in the Department of Criminology and Criminal Justice at the University of Maryland, College Park. She is the PI for NIDA's Criminal Justice Drug Abuse Treatment Studies (CJ-DATS) Coordinating Center.

Six of one, half a dozen of the other:
The benefits and challenges associated with risk classification tools

Introduction

Over the past two decades, there has been an increased awareness among researchers and policymakers of the need to improve decision-making practices within the juvenile justice system. Historically, juvenile justice agencies have relied upon a clinical approach to decision-making, collecting information through unstructured interviews and case file reviews. Researchers have criticized this approach, however, because it allows a significant amount of discretion in the interpretation and application of the information. As a result, it is all but impossible to determine which factors play a significant role in the decision-making process (Gottfredson, Wilkins, & Hoffman, 1978; Gottfredson & Gottfredson, 1986; Andrews & Bonta, 1994). Consequently, there has been mounting pressure for juvenile justice agencies to devise a more structured method to decision-making, particularly one that incorporates the use of empirically-based risk assessment instruments.

While the traditional concept of "risk" has centered on the relationship between generally "static" factors (e.g., prior history, age) and future deviant behavior, in the past two decades, researchers have begun to expand the concept to include a set of dynamic factors that have been shown to contribute to or diminish an individual's likelihood of re-offending (Bonta, 1996; Van Voorhis & Brown, 1997). These may include such things as the youth's performance at school, involvement with drugs and alcohol, and the nature of family relationships. While there have been notable strides in the conceptualization as to which factors should be included on a risk instrument, their predictive efficiency has remained relatively limited (Gottfredson & Gottfredson, 1980; Copas & Tarling, 1986; Farrington, 1987; Gottfredson, 1987; Ashford & LeCroy, 1988; Ashford & LeCroy, 1994; Jones, 1996; Van Voorhis & Brown, 1997).

Despite the diverse nature of the juvenile justice population, agencies have often adopted a "one-size fits all" approach when developing risk-assessment instruments. Given the significant amount of time and resources required to develop a risk classification instrument, some agencies have tried to cut corners either by taking an extant instrument developed elsewhere without first validating it on their own population or by developing an instrument for one population and then applying it unilaterally (Andrews & Bonta, 1994; Ashford & LeCroy, 1988; Kemshall, 1998; Pfeifer, Young, Bouffard, & Taxman, 2001). Not surprisingly, this has resulted in the misclassification of a significant proportion of youth. In particular, some instruments produce a high proportion of Type I errors (e.g., false positives), while others produce a disproportionate number of Type II errors (e.g., false negatives). Each of these raises a number of practical, as well as ethical issues.

Instruments that produce a high proportion of false positives will trigger a "net-widening" effect, causing a greater number of low-risk youth to be assigned more stringent or intrusive conditions of supervision. In the worst case scenario, such an error could result in a youth being detained unnecessarily. On a practical level, such errors deplete agency resources that need to be allocated for high-risk youth (Van Voorhis & Brown, 1997). On an ethical level, "the imprisonment of an offender…because it is predicted that he will commit offenses at high rates in the community is repugnant essentially because it is undeserved" (Farrington, 1987, p.90). In contrast, instruments that produce a high proportion of false negatives will inappropriately release high-risk youth back into the community, raising serious concerns from a public-safety perspective. As a result of the failure to remedy these errors, many juvenile justice practitioners have questioned the face validity of these instruments. Consequently, line staff within juvenile justice agencies often approach standardized risk assessment practices with a great deal of reluctance.

To address some of the challenges associated with the development and implementation of a standardized risk instrument for a juvenile population, the current research presents a case study of a classification tool currently under development for the Maryland Department of Juvenile Justice.[1] A number of recommendations are presented on how to improve these instruments' predictive efficiency, as well as on how to overcome obstacles in their implementation.

History of Risk Assessment

In the broadest context, risk assessment refers to the method used to estimate an individual's likelihood of engaging in a variety of future negative behaviors. While the most common outcome measure adopted has been recidivism, these instruments have also been used to estimate the likelihood that an individual will fail to appear for assigned court dates, escape from custody, or violate conditions of supervision in the community (Van Voorhis & Brown, 1997). The proposed value of risk assessment is that it can bring greater degree of validity, structure, and consistency to an agency's decision-making process and help allocate resources more efficiently (Gottfredson & Gottfredson, 1984; Farrington, 1987; Gottfredson, 1987; Bonta, 1996; Jones, 1996; Van Voorhis & Brown, 1997). While most advances in risk assessment have taken place within the adult offending population, juvenile justice officials have also recognized the utility of this methodology. Accordingly, a number of agencies have begun to integrate this practice throughout the various stages of juvenile case processing (Van Voorhis & Brown, 1997).

Traditionally, agencies have relied upon either subjective (e.g., "clinical") or objective (e.g., "actuarial") criteria to help guide their decision-making (Bonta, 1996). Those that have adopted a clinical approach to risk assessment collect information through unstructured interviews and case file reviews, and then attempt to interpret the information in some meaningful way. In contrast, agencies that have pursued an actuarial approach to risk assessment collect data on a set of offender characteristics empirically shown to have a relationship with a specific outcome measure. Offenders are "scored"

on each item, and those scores are then summed to form a composite risk score that is used to classify offenders into different levels of risk (i.e., low, medium, high).

While the application of an actuarial instrument appears straightforward, researchers have discovered its development is often not so simple. From the outset, researchers are faced with a number of methodological issues, such as: What items should be included on the instrument; what types of information should be used to collect the data; and, how should the items be scored. The following discussion addresses each of these issues in greater detail.

The Challenges of Risk Assessment

One of the first challenges researchers face is to identify which risk factors to include on their instrument. All too often, researchers have failed to look beyond a relatively small group of static factors, such as the youth's current offense, age of first referral, or number of prior adjudications (Bonta, 1996). While these items may be useful in guiding decisions concerning the level of freedom an offender should be granted, they can not provide direction as to how to *treat* the youth (Bonta, 1996).

In response to this shortcoming, researchers have recognized the need to develop instruments that include a wider array of information that can be used by agency officials to develop a youth's individual service plan. Such instruments typically include a set of dynamic factors, as well as static risk factors. Often referred to within the literature as "criminogenic needs," these items typically focus on a set of factors within the youth's life that are amenable to change, such as school performance, family functioning, peer relationships, and drug use. Researchers and practitioners have argued that modifying or diminishing the effects of these factors can reduce an individual's probability of re-offending (Andrews & Bonta, 1994, Bonta, 1996; Van Voorhis & Brown, 1997).

Additional methodological issues which researchers must consider include deciding how the requisite information should be collected. Typically, risk assessment instruments utilize two types of data, official records and youths' self-reports. The

former provides information pertaining to the youth's official contact with the juvenile justice system and may also document specific services the youth received while in the care of the agency (e.g. mental health, substance abuse, education, etc.). Unfortunately, the quality of these records is highly variable and is often missing a significant amount of data (Gottfredson & Gottfredson, 1980; Jones, 1986; Farrington, 1987; Gottfredson, 1987; Van Voorhis & Brown, 1997; Pfeifer et al., 2001). Therefore, some instruments now require that data be collected from multiple sources, such as official records, youth interviews, school records, and interviews with the youth's family.

Since the inception of risk assessment, researchers have debated the various scoring strategies utilized by these instruments. In the earliest method developed by Burgess (1928) in his study of adult parolees, items were constructed dichotomously, with subjects assigned a score of 0 or 1 on each of the predictors. Values were then summed to create a composite score. Burgess hypothesized that the higher the score, the greater the likelihood the individual would violate his parole. This approach came under criticism, however, for its assumption that all of the risk factors were equally predictive and for its inability to detect possible redundancies in variables due to interaction effects (Copas & Tarling, 1986; Gottfredson, 1987a; Gottfredson, 1987b; Bonta, 1996; Jones, 1996; Van Voorhis & Brown, 1997).

A more sophisticated method of scoring, initially developed by Sheldon and Eleanor Glueck (1950), assigns different weights to each item based upon its relationship with the specified outcome measure. While in theory this approach is believed to be a superior method for developing prediction models, in practice, numerous studies have failed to find significant improvements in models based upon multiple regression techniques than those that utilized the Burgess method (Simon, 1971; Gottfredson & Gottfredson, 1979; Farrington & Tarling, 1985; Gottfredson & Gottfredson, 1985; Tarling & Perry, 1985; Farrington, 1987; LeBlanc 2000). Moreover, LeBlanc (2000)

asserts that additive models are easier for agency workers to score, therefore, help to reduce problems when implementing a standardized risk instrument.

While most of the research on risk assessment has occurred within the adult offending population, there has been an increased interest within the juvenile justice field to incorporate this practice into various stages of decision-making. To date, at least 16 states have developed some type of formal risk assessment measure for juveniles (Towberman, 1992). While some states employ these tools during early stages of case processing to determine supervision levels for youth released back into the community, other states do not utilize the instruments until after a youth has been committed to one of their institutions (Towberman, 1992). In the late 1990s, Maryland's Department of Juvenile Justice began to take steps to develop a more formal approach to risk assessment to help guide placement decisions for adjudicated youth.

The current study provides a synopsis of how the Department's approach to risk assessment has evolved over the past five years and highlights some of the challenges researchers have faced when developing an empirically-based instrument. The paper concludes by outlining some of the issues researchers should address when evaluating the predictive efficiency of any risk assessment instrument.

Risk Assessment Within the Maryland Department of Juvenile Justice

In early 1997, the Maryland Department of Juvenile Justice [DJJ] contracted with the National Council of Crime and Delinquency [NCCD] to conduct a validation study of a "consensus-based" risk assessment measure DJJ had created independently to help guide placement decisions for adjudicated youth.[2] By choosing items based on what "made sense," DJJ committee members selected the specific items to be included in the instrument, and decided how each of those items should be weighted. The committee then selected specific cut-off scores or "cut-points" to determine supervision levels. While NCCD found the model's scale scores to be significantly related to recidivism rates, it did not produce meaningful distinctions across risk levels. Using additional data

provided within individual youth's case records, NCCD then went on to develop an empirically-based risk assessment instrument for the agency.

The revised instrument, comprised of nine items, was a combination of static and dynamic risk factors that were identified as significant predictors of recidivism. Using this scale, youth were then classified into four risk classification groups: low, medium, high, and very high. NCCD found that this revised scale not only predicted different types of recidivism (e.g., felony referral and violent crime referral), but also provided meaningful distinctions across risk levels.

While the instrument developed by NCCD was an improvement over the consensus-based model developed by DJJ, the findings presented in NCCD's report were subsequently met with a significant degree of caution. In reviewing the construction of NCCD's instrument, both internal and external reviews of the study's methodology raised a number of concerns, the most important of which centered on the sampling strategies that were employed. Specifically, NCCD stated both of their "validation" studies were conducted using a random sample of 833 youth placed on probation or released to aftercare. However, upon closer examination of the data, reviewers found less than ten percent of this sample was drawn from the aftercare population, challenging the relative stability of the findings for this particular sub-population. In addition, NCCD failed to include a sufficient number of females and youth from the smaller jurisdictions in the state to provide any meaningful information. Consequently, reviewers challenged whether the findings reported by NCCD could be generalized to a more representative sample within the DJJ population. Reviewers were also critical of the fact that NCCD did not validate its own tool on a separate sample of adjudicated youth. Rather, the sample NCCD used to construct the instrument was the same sample used to "test" its efficiency.

Given these concerns, DJJ contracted with the Bureau of Governmental Research at the University of Maryland, College Park [BGR] to evaluate the effectiveness of the

NCCD instrument using a more representative sample, and to determine if the instrument could somehow be improved upon. After careful evaluation, NCCD's model was found to produce less than a one percent improvement over what was predicted by chance.[3] As a result, additional analyses were undertaken to identify whether any additional risk factors existed that could improve NCCD's instrument.

Methods

Using automated records provided by the agency for fiscal year 1997, a stratified random sample of 694 adjudicated juveniles was selected. To ensure a more representative sample was captured than constructed in the NCCD study, committed youth, female offenders, and youth from the Eastern Shore were over sampled. The sample was stratified based upon jurisdiction, disposition status (e.g., commitment vs. probation), gender, and race. Within each region, cases were then randomly selected to ensure similar dispositional, racial, and gender distributions were captured.

Forty-four percent of the youth were committed to the Department of Juvenile Justice, while the remaining fifty-six percent were assigned to probation. As illustrated in Table 1, the majority of the sample is comprised of minority, mid-adolescent males. Approximately forty percent of the youth were from Baltimore City. Half of the youths' current sample offense was for a felony, with the greatest proportion attributed to either a felony drug distribution or an auto theft charge. The majority of the youth had at least one prior referral to DJJ and almost half of them had at least one prior adjudication.

In addition to their offending history, data were collected from youths' case files on over forty items pertaining to their school performance, mental health, drug/alcohol use, peer associations, and family functioning at the time of the sample offense. To capture as much information as possible, coders relied upon any social history report, psychological assessment, or field notes completed within that time frame. Items were coded dichotomously (Yes = 1; No = 0).[4] Missing data were handled by replacing it with the mean value for all cases present on that variable.

Table 1. Descriptive Statistics of 694 Adjudicated Youth[56]

Item	Construction (N = 343)	Validation (N = 351)	Total Sample (N = 694)
Demographics			
Male	.82	.80	.81
White	.34	.36	.35
Non-White	.66	.64	.65
Age at time of offense (Mean)	15.15	15.29	15.22
Age Range	9-18	9-18	9-18
Residence[7]			
a. Area 1	.41	.37	.39
b. Area 2	.20	.19	.20
c. Area 3	.13	.14	.14
d. Area 4	.06	.07	.07
e. Area 5	.19	.23	.21
Offense History			
Disposition Status		.	
Commitment	.46	42	.44
Probation	.54	.58	.56
Current Offense			
Misdemeanor	.50	.50	.50
Felony	.50	.50	.50
a. Violent Felony[8]	(.09)	(.11)	(.10)
b. Auto theft	(.19)	(.15)	(.17)
c. Drug distribution	(.16)	(.15)	(.16)
d. Property	(.07)	(.09)	(.08)
Prior History			
No prior involvement	.25	.28	.26
Prior misdemeanor adjudiction	.48	.48	.48
Prior felony adjudication	.43	.52	.48

Table 1 (continued)

Needs			
School Issues (%)			
Attendance problems	.49	.56	.52
Behavioral problems	.67	.65	.66
Assigned to special ed.	.35	.32	.33
Failure of at least one grade	.59	.57	.58
Mental Health (%)			
Rec'd outpatient treatment	.21	.22	.22
Rec'd inpatient treatment	.12	.13	.12
Prescribed psychiatric meds.	.20	.20	.20
Substance Abuse (%)			
Rec'd outpatient treatment	.17	.17	.17
Rec'd inpatient treatment	.10	.12	.11
Youth's reported/suspected level of use (Range: 1 'experimentation' to 4 'daily use')	2.45	2.51	2.48
Peers (%)			
Associates w/deviant peers	.80	.75	.77
Family Functioning (%)			
Youth living w/ single parent	.64	.66	.65
History of drug addiction among either parent	.48	.47	.48
Siblings have been referred to DJJ	.19	.15	.17
Youth reported to have been abused or neglected	.19	.17	.18
At least one biological parent is deceased	.16	.14	.15

Not surprisingly, a significant degree of dysfunction was found in these youths' lives. Specifically, two thirds of the sample was reported to have exhibited moderate to serious behavioral problems at school. In addition, twenty-two percent had sought some type of outpatient mental health services in the previous year, and seventeen percent had received outpatient services for alcohol and/or drug abuse. Finally, over three quarters of the youth were reported to associate with deviant peers and nearly half had at least one parent who had a history of drug abuse.

Youth were "tracked" over a standardized 15 month follow-up period to identify which youth recidivated. To be consistent with the NCCD study, two dichotomous outcome measures were selected: a) whether a youth was re-referred to DJJ (e.g., "any referral") and b) whether a youth returned on a felony charge (e.g., "any felony referral"). The former, however, was subsequently dropped from the analysis because it was discovered that such a broad definition resulted in over an 80 percent recidivism rate for the sample. Moreover, an examination of the distribution of charge types for first re-offense revealed that nearly 20 percent of the youth were subsequently re-referred for status offenses, technical violations, or other low-level misdemeanors.[9] Therefore, a third outcome measure was constructed. This resulted in only those youths charged with more serious misdemeanor and felony offenses being coded as recidivists ("any referral II"). Using this definition, 65 percent of the sample was found to have recidivated within the 15 month follow-up period. Using the outcome variable "any felony referral," 45 percent of the youth were found to have recidivated within the same time frame.

A risk instrument was then constructed using these two outcome measures. The following discussion outlines the procedures employed to construct the instrument, and examines how it was used to classify the sample youth into different groups. It concludes with a discussion of some of the hurdles researchers encounter when attempting to evaluate the predictive efficiency of classification models and sets forth specific ways to improve such practices in the future.

Analysis and Findings

To construct the instrument, the population was randomly split into construction (N= 343) and validation (N= 351) samples. Both were found to be comparable on nearly

all items.[10] Items were first examined on a bivariate level to assess how strongly they correlated with each outcome measure. For "any referral II," twenty-five items were identified; for "any felony referral," twenty items were identified. These items were then entered into logistic regression models.[11] As illustrated in Table 2, six items were found to be significant predictors of "any referral." However, after testing these same items on the validation sample, only two remained significant (e.g., youth's involvement with deviant peers and parental drug abuse). Both of these items were found to more than double a probability of youth's odds of being re-referred to DJJ.

Following the same procedure to estimate the second model, six items were found to be significant predictors of "any felony referral." However, after testing these same items on the validation sample, only three remained significant: current felony offense, school attendance/behavioral problems, and grade failure. Items that predicted either outcome measure among the validation sample were subsequently added to the instrument.[12] In addition, three policy-based items were added that agency officials had requested (e.g., age at first referral, current violent felony offense, history of drug treatment). In sum, eight risk factors were selected for the final instrument. A copy of this measure is provided in Appendix A.

Each of the items were assigned equal weight and coded dichotomously (Yes = 1; No = 0). Items were then summed to create a composite risk score. To examine the instrument's predictive efficiency, the instrument was tested on the entire adjudicated sample. Table 3 provides the distribution of these scores.

Table 2. Odds-Ratios for Recommended Model – Total Adjudicated Sample

	Any referral II	**Any felony referral**	**Any referral II**	**Any felony referral**
	Construction	Validation	Construction	Validation
Gender (CONTROL)	1.97 ***		2.69 ***	8.48 ***
Race (CONTROL)				
Age at time of offense (CONTOL)		.87 ***		.83 ***
Current offense is felony auto theft (CONTROL)	2.52 ***			2.16 ***
At least one parent has significant mental health problem (CONTROL)	.10 ***			
Total felony adjudications	1.43 ***		1.89 ***	
Total violent felony adjudications			.25 ***	
Current offense is a felony			3.62 ***	2.94 ***
Youth has had significant school problems in past year (e.g., attendance or behavior)			1.47 *	2.35 ***
Youth has failed at least one grade			1.68 **	1.63 **
Youth has received either in- or outpatient mental health treatment	.51 **			
Youth has been, or currently is on psychiatric medication	6.43 ***			
Youth associates with deviant peers	2.00 ***	2.31 ***		
Youth has, or is currently expecting a child	.41 **			
At least one parent has had a significant drug abuse problem within the past year	1.81 **	2.83 ***	1.77 **	
Model Statistics -2 Log	761.75	906.12	773.91	793.74

* p < .1 ** p < . 05 *** p <.01

Table 3. Distribution of Risk Scores Among Adjudicated Sample

Risk Score	*Frequency*	*Valid Percent*	*Cumulative Percent*
0	55	7.9	7.9
1	103	14.8	22.8
2	166	23.9	46.7
3	157	22.6	69.3
4	120	17.3	86.6
5	60	8.6	95.2
6	26	3.7	99.2
7	7	1.0	100.0
Mean 2.72	N = 694	100.0	100.0

Because the ultimate utility of the tool is to identify "high risk" youth, a statistical technique was used to create classification schemes using the scores from the instrument. The statistical tool used for these purposes was the "Relative Improvement Over Chance" [RIOC] test statistic (Loeber & Dishion, 1982; Gottfredson & Gottfredson, 1986), which tests the efficiency of the instrument by determining how a model performs relative to its expected performance and its best possible performance. Scores range from zero, indicating no improvement, to one, indicating absolute improvement. The closer the score is to one, the more efficient the model.

One of the advantages of the RIOC is that it takes into account the model's base rate and selection ratio, thereby allowing researchers to assess the predictive efficiency of models whose criterion variables have skewed distributions (Copas & Tarling, 1986; Gottfredson & Gottfredson, 1986; Jones, 1996). The weakness of this statistic, however, is that it creates a two-way classification scheme assigning youth to only "high" or "low" risk categories, in contrast to many classification systems that have three to four levels. Moreover, the RIOC score can be manipulated by changing the "cut-off" point between

these two classifications. While this allows researchers to fine-tune the overall predictive efficiency of their model, they should do so cautiously.

The current study illustrates the problems associated with selecting a cut-off point. In particular, when predicting future felony referrals, the model produces an RIOC score ranging from .23 to .36 depending on where the cut-point is placed. As illustrated in Table 4, when youth are classified as high-risk with a score of three or above, the model produces an RIOC score of .23. When the cut-point is raised to four, the RIOC score remains relatively unchanged. However, when the cut-point is raised to five, the RIOC score increases to .36. While this latter model demonstrates the greatest amount of predictive efficiency, all of the models produce a significant proportion of false positives and false negatives.

Table 4. Predictive Efficiency of "Any Felony" Model

	Cut-point	False Positives	False Negatives	RIOC Score
Model I	3	32%	58%	.226
Model II	4	32%	61%	.228
Model III	5	29%	62%	.359

With Model I (cut-point 3), the resulting classification scheme misclassified fifty-eight percent of the youth as low-risk and thirty-two percent as high-risk. Similarly, Model II (cut-point 4) misclassified sixty-one percent as low-risk and thirty-two percent as high-risk. Finally, Model III (cut-point 5) misclassified sixty-two percent as low-risk and twenty-nine percent as high-risk. This degree of prediction error requires that

researchers further develop the instrument and explore additional strategies to improve its overall efficiency.

One of the greatest challenges researchers face when developing risk classification instruments is related to the quality of the data provided by agencies (Gottfredson & Gottfredson, 1980; Jones, 1986; Farrington, 1987; Gottfredson, 1987; Van Voorhis & Brown, 1997; Pfeifer, Young, Bouffard, & Taxman, 2001). This issue was of particular concern to the present study. Due to the lack of uniformity in the type and quality of data recorded in each youth's case file, researchers were prevented from collecting information on all the variables that might have been useful in developing the tool.[13] For example, little information was found on recent drug use by either of the youths' parents.[14] Consequently, to maximize the number of cases for analysis, a standard technique was used to replace missing data by using the mean of that particular variable (Neuman, 2000).[15] The weakness of this resolution, however, is that it prevents researchers from understanding some of the distribution of different variables. Further, because this instrument is comprised of only eight items, missing data may have significantly impaired the researchers' ability to adequately assess its efficiency. Consequently, it is important that data be collected from multiple sources (e.g., agency records, youth self-reports, parent interviews, school records, etc.) in the future stages of the instrument's development.

Another factor that may have limited the current instrument's predictive efficiency is its reliance on dichotomous risk factors and outcome measures. Several scholars have argued that risk instruments should focus on continuous, rather than dichotomous items because the latter treats risk items as a simple "yes" or "no"

phenomena and assumes all factors are equally predictive (Bonta, 1996; Gottfredson, 1987a; Gottfredson, 1987b; Jones, 1996; Van Voorhis & Brown, 1997). This may be especially problematic when developing prediction models with juveniles because adolescence is a time marked by significant physical, intellectual, and emotional changes. This makes it more difficult for researchers to identify a set of "static" risk factors that accurately predict future behavior (Altschuler & Armstrong, 1991; Kazdin, 2000). When dichotomous items are adopted, researchers should assign different weights to each item, depending on the relative strength of the relationship each shares with the criterion variable (Gottfredson, 1987a; Gottfredson, 1987b; Van Voorhis & Brown, 1997).

On a similar note, a number of critics have argued against the use of dichotomous outcome measures. They contend that by defining these types of behavior in absolute terms, such as "success" or "failure," researchers limit their ability to understand the nuances of different behaviors (Blumstein, Cohen, Roth, & Visher, 1986; Gottfredson, 1987a; Jones, 1996; Van Voorhis & Brown, 1997). In fact, Gottfredson (1987a) argues that since any "behavior of interest is almost always [a] matter of degree" (p.15), researchers should try to adopt continuous measures of recidivism, such as the length of time to re-arrest or seriousness of re-offense, to enable instruments to measure change in behavior over time. Similarly, researchers should pay careful attention to which outcome measure(s) they chose to build their instrument. If the base-rate of a behavior is particularly high or low (e.g., violent crime), it will be difficult for the instrument to make more accurate predictions using a statistically-based model, than if the behavior was predicted by chance alone (Gottfredson, 1987a).

The prior discussion highlights just a few of the complexities researchers face when developing standardized risk instruments. Each of these problems can significantly influence an instrument's predictive efficiency. The key to developing these tools is to recognize that risk assessment is a process, rather than a task that can be completed in a short period of time. By identifying different factors that may have contributed to the level of prediction error in the current instrument, researchers will have a better idea of the type of adjustments that need to be made in the subsequent stages of its development.[16]

Discussion

Scholars have touted the importance of valid risk screening tools for the adult and juvenile systems for nearly two decades. During this period, advances have been made both in the development of these instruments, and in the way in which their efficiency is measured. Their primary strength is that if implemented correctly, they help reduce the amount of discretion in the decision making process by limiting the variables that should be taken into consideration to only those that are theoretically and statistically correlated with failure (negative outcomes). Consequently, it makes it easier for the agency to train its staff. Finally, such practices have enabled agencies to clearly articulate their policies with regard to detention and confinement practice.

The current study reviews the process researchers followed to create a standardized risk instrument for the Maryland Department of Juvenile Justice to help classify adjudicated youth. While researchers were able to construct an instrument, a number of challenges presented themselves when devising a classification scheme. Though researchers were able to utilize the RIOC statistic to identify the model that

produced the greatest amount of predictive efficiency, the statistic was only able to create a two-way classification scheme assigning youth to only "high" or "low" risk. In addition, the proposed model produced a significant proportion of false positives and false negatives. The consequences of both types of error can not be understated. Misclassifying youth as high-risk may lead the agency to impose more stringent and intrusive conditions of supervision. In the worst case scenario, such an error could result in detaining a youth unnecessarily. Given that agency resources are limited, it is important that available resources be reserved for the most serious offenders. Assigning more intensive measures to low-risk youth is not only fiscally irresponsible, but it may also cause more harm in the long run (Van Voorhis & Brown, 1997). Similarly, misclassifying a youth as low-risk could result in the agency's releasing the youth back into the community where he or she could pose a threat to public safety. In the worst case scenario, the youth could commit a serious, violent crime (e.g., "Willie Horton incident"). This, in turn, may cause the public to question the efficiency of the agency itself.

Ultimately, both of these errors jeopardize the current instrument's validity from the perspective of the stakeholders and the line-staff who are responsible for implementing the protocol. If researchers remain insensitive to the fiscal, policy, and ethical costs associated with these errors, the agencies will become more resistant against adopting these practices. In turn, it will become even more challenging for researchers to develop sound prediction models.

Conclusion

Developing a risk tool that is used to produce a classification scheme is a mix of art and science. In particular, the science can help identify the variables that should be included in the instrument based on the strength of the correlation they share with various undesirable outcomes, as well as the theoretical values of such variables. The art lies in how researchers use the RIOC scores to develop the most efficient model. If they ignore the proportion of false positives and false negatives the classification scheme produces, the overall utility of the instrument will be reduced. As stated earlier, a risk tool that overclassifies youth will result in net widening policies, while a tool that under classifies them will cause an agency to fail to fulfill its mission of public safety. Despite such challenges, if developed carefully, these classification schemes will benefit both the agency and the juvenile population it serves.

Appendix A

BGR Risk Instrument

Youth Name: _____ **DOB:** _____/_____/_____

Gender: **1.** M **2.** F **Ethnicity:** **1.** Caucasian **2.** African-American **3.** Hispanic **4.** Other

County of Residence: _____ **County of Jurisdiction:** _____
Most Serious Current Adjudicated Offense: _____
Date of Assessment: _____/_____/_____

_____

		Score
1. Age at first referral to the Department of Juvenile Justice (Age: ____)		
a. 11 or younger	1	
b. 12 and older	0	_____
2. Is current adjudicated offense a felony?		
a. Yes	1	
b. No	0	_____
3. Is current adjudicated offense a violent felony? (Include all violent person-to-person Type I and II offenses. Do **NOT** include simple assaults)		
a. Yes	1	
b. No	0	_____
4. Has youth had moderate to severe school problems within the prior 12 months? (Youth has been reported: a) truant >30 days, b) dropped out, c) suspended at least once in the past 12 months, **OR**, d) expelled)		
a. Yes	1	
b. No	0	_____
5. Has youth ever failed a grade?		
a. Yes	1	
b. No	0	_____
6. Does the youth's peer group consist of some delinquent friends? (Youth associates with others who have been involved in delinquent/criminal activity, or has more than three arrests involving co-defendants)		
a. Yes	1	
b. No	0	_____
7. Has either of the youth's parents (or caretakers) reported a major substance abuse problem within the prior 12 months? ('Major' substance abuse problem means it has interfered with their ability to hold a job, or to provide for their children's welfare, or to supervise the youth effectively. It may also include any addiction, drug distribution, or any other criminal activities associated with substance abuse)		
a. Yes	1	
b. No	0	_____
8. Has youth ever received either in- or outpatient drug treatment?		
a. Yes	1	
b. No	0	_____

Total Risk Score: _____

Appendix B

Predictive Efficiency of "Any Felony" Model

A. Cut-off point 3 and above

TN 138	**FN** 186
FP 120	**TP** 250

N = 694

% FN = .58 % FP = .32 RIOC = .226

B. Cut-off point 4 and above

TN 189	**FN** 292
FP 69	**TP** 144

N = 694

% FN = .61 % FP = .32 RIOC = .228

C. Cut-off point 5 and above

TN 231	**FN** 370
FP 27	**TP** 66

N = 694

% FN = .62 % FP = .29 RIOC = .359

References

Altschuler, D., & Armstrong, T. (1991). "Intensive aftercare for the high risk juvenile parolee: Issues and approaches in reintegration and community supervision." In T. Armstrong (Ed.), *Intensive interventions with high risk youth: Promising approaches in juvenile probation and parole*. Monsey, NY: Willow Tree Press.

Andrews, D. A., & Bonta, J. (1994). *The psychology of criminal conduct*. Cincinnati, OH: Anderson Publishing.

Ashford, J. B., & LeCroy, C. W. (1988). "Predicting recidivism: An evaluation of the Wisconsin Juvenile Probation and Aftercare Instrument." *Criminal Justice and Behavior 15* (2), 141-151.

-------- (1990). "Juvenile recidivism: A comparison of three prediction instruments." *Adolescence 25* (98), 441-450.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18-32). Thousand Oaks, CA: Sage.

Champion, D. J. (1994). *Measuring offender risk: A criminal justice sourcebook*. Westport, CT: Greenwood Publishing.

Copas, J. B., & Tarling, R. (1986). Some methodological issues in making predictions. In A. Blumstein, J. Cohen, J. Roth, & C. Visher (Eds.), *Criminal careers and'Career criminals'* (pp. 291-355). Washington, DC: National Academy Press.

Gottfredson, D. M. (1987a). "Prediction and classification in criminal justice decision making." In D. Gottfredson & M. Tonry (Eds.), *Crime and justice: A review of research* (pp.1-20). Chicago: University of Chicago Press.

Gottfredson, S. D. (1987b). "Prediction: An overview of selected methodological issues." In D. Gottfredson & M. Tonry (Eds.), *Crime and justice: A review of research*(pp.21-51). Chicago: University of Chicago Press.

Gottfredson, S. D., & Gottfredson, D. M. (1980). "Data for criminal justice evaluation: Some resources and pitfalls." In M. Klein & K. Teilmann (Eds.), *Handbook of criminal justice evaluation*. Beverly Hills, CA: Sage.

-------- (1986). "Accuracy of prediction models." In A. Blumstein, J. Cohen, J. Roth, & C. Visher (Eds.), *Criminal careers and 'Career criminals'* (pp. 212-290). Washington, DC: National Academy Press.

Howell, J. C., Krisberg, B., & Jones, M. (1995). "Trends in juvenile crime and youth violence." In J. C. Howell, B. Krisberg, J. D. Hawkins, & J. J. Wilson (Eds.), *Serious, violent, & chronic juvenile offenders* (pp.1-35). Thousand Oaks, CA: Sage.

Jones, P. R. (1996). "Risk prediction in criminal justice." In A.T. Harland (Ed.), *Choosing correctional options that work* (pp.33-68). Thousand Oaks, CA: Sage.

Kemshall, H. (1998). *Risk in probation practice*. Aldershot, England: Ashgate.

LeBlanc, M. (2000). "Review of screening, decision-making, and clinical assessments strategies and instruments for adolescent offenders." Paper presented at NATO Advanced Research Workshop, Crakow, Poland.

Loeber, R., & Dishion, T. (1983). "Early predictors of male delinquency: A review." *Psychological Bulletin* 94, 68-99.

Loeber, R., Farrington, D., & Waschbusch, D. A. (1999). "Serious and violent juvenile offenders." In R. Loeber, & D. Farrington (Eds.), *Serious and violent juvenile offenders – Risk factors and successful intervention* (pp.13-39). Thousand Oaks, CA: Sage.

Kazdin, A. E. (2000). "Adolescent development, mental disorders, and decision making of delinquent youths." In T. Grisso & R. G. Schwartz (Eds.), *Youth on trial: A developmental perspective on juvenile justice* (pp. 33-65). Chicago: University of Chicago Press.

Neuman, W. L. (2000). *Social research methods: Qualitative and quantitative approaches (4th ed.)*. Boston: Allyn Bacon.

Pfeifer, H. L., Young, D., Bouffard, J., & Taxman, F. (2001). *Department of JuvenileJustice Risk Screening Project*. College Park, MD: Bureau of Governmental Research.

Snyder, H. N. (1999). "Serious, violent, and chronic juvenile offenders – An assessment of the extent of and trends in officially recognized serious criminal behavior in a delinquent population." In R. Loeber, & D. Farrington (Eds.), *Serious and violent juvenile offenders – Risk factors and successful intervention* (pp.428-444). Thousand Oaks, CA: Sage.

Van Voorhis, P., & Brown, K. (1997). *Risk classification in the 1990s*. Washington, DC: National Institute of Corrections.

Weitekamp, E. G., Kerner, H. J., Schindler, V. & Schubert, A. (1995). "On the 'dangerousness of chronic/habitual offenders': A re-analysis of the 1945 Philadelphia birth cohort data." *Studies in Crime and Crime Prevention: Annual Review, 4*, 159-175.

Wiebush, R., Johnson, K., & Wagner, D. (1997). *Development of an empirically based risk assessment instrument and placement matrix for the Maryland Department of Juvenile Justice. Final report*. Washington, DC: National Council on Crime and Delinquency.

Notes

[1] For a copy of the report on the development of the Maryland Risk Assessment Instrument, send a request to Dr. Faye Taxman at the Bureau of Governmental Research, University of Maryland, 4511 Knox Road Suite 301, College Park, MD 20740.

[2] For more information see R. Wiebush, K. Johnson, & D. Wagner (1997). *Development of an empirically-based risk assessment instrument and placement matrix for the Maryland Department of Juvenile Justice. Final report*. Washington, DC: National Council on Crime and Delinquency.

[3] For an in-depth description of this evaluation, see H. Pfeifer, D. Young, J. Bouffard, & F. Taxman (2001). *Department of Juvenile Justice Risk Screening Project*. College Park, MD: Bureau of Governmental Research.

[4] Three items, however, were not coded dichotomously: a) 'youth's reported/suspected level of drug use in the past twelve months' coded 1 (only once or twice) to 4 (daily, or almost daily); b) 'number of siblings under the age of 18 that live in home,' coded continuously; and c) 'summary characterization of family functioning at the time of the 1997 sample disposition/release,' coded 1 (no problem) to 3 (major disorganization or dysfunction).

[5] Columns were rounded up to the nearest percentage and therefore may not be equivalent to 100 percent.

[6] Prior to analysis, the sample was randomly divided into two samples: a construction and a validation sample. The former was used to 'build' the instrument (e.g., identify the most significant predictors of the outcome variable), and the latter to 'test' the instrument's validity. Table 1 presents the descriptive statistics for the total sample, as well as each of the sub-samples.

[7] At present, the Maryland Department of Juvenile Justice is divided into five separate areas: Baltimore City (Area One); Baltimore, Howard, Carroll, and Howard counties (Area 2); Montgomery, Washington, Frederick, Allegany, and Garrett counties (Area 3); Eastern Shore (Area 4); Anne Arundel, Charles, Calvert, St. Mary's, and Prince George's counties (Area 5).

[8] "Violent felony" includes those offenses defined by Article 27, Section 643B in the Maryland Code. These include the following: Abduction; Arson (1st degree); Assault (1st degree); Carjacking; Child Abuse; Incest; Kidnapping; Manslaughter (except involuntary manslaughter); Mayhem and Maiming; Murder; Rape; Robbery; Sexual offense (1st and 2nd degree); and, any attempt to commit the aforementioned acts.

[9] Low-level misdemeanors incorporate those offenses characterized by the agency as 'Type II Misdemeanor," such as: disorderly conduct, disturbing the peace, failure to obey a lawful order, fireworks violations, loitering, pager at school, tampering, telephone misuse, traffic violations, etc.

[10] Significant differences were only found on two items (current offense is felony auto theft and youth's mother had a significant mental health problem), therefore, they were entered as control variables when the logistic models were run.

[11] Along with the items identified at the bivariate level, five control variables were included in the logistic models (youth's race, gender, age at the time of current offense, current auto-theft offense, and parental mental health problem). Other than the control variables, all the independent variables were entered into each model as its own 'block' using forward method of variable entry.

[12] A total of five items were found to be significant predictors of the two outcome measures: youth's involvement with deviant peers, parental drug abuse, current felony offense, school attendance/behavioral problems, and grade failure.

[13] While researchers attempted to collect data from the youth's case file on over forty items including his or her school performance, mental health, drug/alcohol use, peer associations, and family functioning, data was found to be missing between 8 –58 percent of the time, depending on the variable. In particular, information pertaining to the youth's mental history, level of drug use, and parent(s) drug and/or criminal history (particularly for the father) was unable to be located between 30-58 percent of the time. For the five social contextual variables included on the risk instrument, the percentage of missing data were as follows: a) school problems (14%); b) grade failure (25%); involvement with deviant peers (26%); parental drug problem (28%); youth drug treatment (20%).

[14] To score a '1' on this item, there had to be evidence that *either* of the youth's parents/caretakers had had a significant substance abuse problem within the previous 12 months. Information pertaining to the youth's mother was unable to be located in 28 percent of the files, while 43 percent of the case files did not have this information about the youth's father.

[15] The authors recognize that this strategy jeopardizes the validity of the instrument, but as noted previously, if all incomplete cases were excluded from analysis, the resulting sample would have been too small.

[16] Given the prior concerns outlined in the paper, the current instrument is undergoing additional analysis and is being tested on additional samples of adjudicated youth. A variety of outcome measures are being tracked over the next twelve months to explore the utility of developing multiple risk assessment instruments for different types of offenders, as well as at different decision-points within the system. Because many of the decisions made within the juvenile justice process often elicit different concerns, it is unlikely a single instrument will perform equally well for every stage of the juvenile justice system. However, by systematically collecting data at multiple points in the system, with different types of offenders, researchers will be in a better position to develop more reliable prediction models and to bring greater uniformity and efficiency to the system.