Psychological Services

The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study

Christopher T. Lowenkamp, James L. Johnson, Alexander M. Holsinger, Scott W. VanBenschoten, and Charles R. Robinson

Online First Publication, November 12, 2012. doi: 10.1037/a0030343

Psychological Services

In the public domain
DOI: 10.1037/a0030343

The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study

Christopher T. Lowenkamp
and James L. Johnson
Administrative Office of the United States Courts,
Washington, DC

Alexander M. Holsinger
University of Missouri-Kansas City

Scott W. VanBenschoten and Charles R. Robinson
Administrative Office of the United States Courts, Washington, DC

Offender assessment has been and remains the cornerstone of effective community supervision. This article presents the development of and tests the predictive validity of a 4th-generation risk assessment instrument designed for U.S. probation. A large administrative data set was used to create the assessment instrument and conduct an initial validation. Subsequent data generated from officer-completed assessments were used to conduct a prospective validation. Finally, data from case vignettes scored by trained officers were used to test the interrater agreement of the assessment instrument. Overall, analysis revealed that the assessment instrument predicted rearrest reliably when using the assessment results based on administrative data or officer-completed assessments. Analysis also revealed high rates of interrater agreement. Recommendations for future research and policy implications are presented.

Keywords: risk assessment, federal probation, validation, PCRA

From the perspective of a probation officer, effective risk classification identifies offenders most likely to violate the law or conditions of supervision, while also identifying factors that can be influenced in order to change the likelihood of recidivism (Van Voorhis & Brown, 1996). Identifying and working most intensively with the highest risk offenders (risk principle), identifying criminogenic needs (need principle), and identifying and compensating for potential barriers to treatment (responsivity principle) are three of the primary principles of effective classification (Andrews, Bonta, & Hoge, 1990).

To date, the most effective assessment tools use both static risk factors (e.g., criminal history factors that do not change over time) and dynamic factors (e.g., criminogenic needs such as current substance abuse) to accurately identify those offenders at greatest risk of reoffending and identify the needs present that put those offenders at risk (Andrews & Bonta, 1998). Although these principles shape practice in general, D. M. Gottfredson (1987) and Andrews and Bonta (1998) note that professional discretion is still a necessary part of the assessment process; however, professional discretion should be informed by the principles noted above and by the results of structured and objective risk assessment.

Risk assessment has evolved in a series of generations from basic to increasingly complex, with each generation using the best available methods to predict the risk of recidivism and then apply the results of the assessment to supervision strategies

Christopher T. Lowenkamp, James L. Johnson, Scott W. VanBenschoten, and Charles R. Robinson, Office of Probation and Pretrial Services, Administrative Office of the United States Courts, Washington, DC; Alexander M. Holsinger, Department of Criminal Justice and Criminology, University of Missouri-Kansas City.

This article is based, in part, on a technical report that appeared in *Federal Probation* in 2011 by Johnson, VanBenschoten, Robinson, and Lowenkamp.

Correspondence concerning this article should be addressed to Scott W. VanBenschoten, Administrative Office of the United States Courts, Office of Probation and Pretrial Services, One Columbus Circle, NE, Washington, DC 20544. E-mail: scott_vanbenschoten@ao.uscourts.gov

(Bonta & Wormith, 2007). This tradition continues today, with researchers continually refining their understanding of criminal behavior and the associated enhancements to risk/needs prediction tools (VanBenschoten, 2008).

For most of the 20th century, professional judgment or intuition was the most common method used to predict criminal behavior. This form of assessment, which is now referred to as the first generation, typically involved an unstructured interview with the offender and a review of official documentation (Andrews & Bonta, 2006; Bonta, 1996; Connolly, 2003; Van Voorhis & Brown, 1996). Although intuitively appealing, this method had certain weaknesses, such as an inability to distinctly determine how decisions are made, an inability to test interrater reliability, and the potential for personal bias to influence case management (Bonta & Andrews, 2007; Monahan, 1981; O'Rourke, 2008; Van Voorhis & Brown, 1996; Wardlaw & Millier, 1978). The first generation was typified by a complete lack of any actuarial assessment tools or methods of statistical prediction, let alone any other standardized processes.

Actuarial (i.e., research-based) instruments, which characterize the second generation of risk assessment, were not widely used in offender classification until the 1970s (although some of the first examples appeared in the 1920s; see, e.g., Burgess, 1928; O'Rourke, 2008). Generally, actuarial instruments assess risk using factors that the extant literature base reveals are related to recidivism (Bonta & Andrews, 2007). These factors are "counted" in some fashion and then summed to create a composite score, with higher scores indicating higher risk of recidivism. Second-generation risk assessments include the Statistical Index of Recidivism (SIR), and the Salient Factor Score (SFS). Like the Burgess (1928) scale noted above, the SIR and SFS and many other examples of second-generation assessments rely almost entirely (if not exclusively) on static (unchangeable) items and are heavily weighted toward information coming from criminal history (e.g., number of prior arrests, history of violence, prior incarceration). Practitioners typically did not (and in some cases do not) resist the use of these tools, as they include information that is relatively easy to collect and items with at least some face validity. Because of their reliance on static factors, these instruments by definition make case planning and the monitoring of progress/change impossible.

The third generation of risk offender classification incorporated dynamic (changeable) factors into actuarial assessments (Bonta & Wormith, 2007). These dynamic factors represent literature-supported criminogenic needs, which when appropriately targeted can reduce overall risk of recidivism. Third-generation actuarial assessments therefore allow for the incorporation of both the risk and need principles of effective intervention (Andrews & Bonta, 1995; Bonta & Andrews, 2007). In addition, third-generation assessments allow for comprehensive case planning and the assessment of change over time, while greatly increasing predictive validity (Van Voorhis & Brown, 1996). Some common examples of third-generation risk/needs assessments include the Wisconsin Risk and Needs Assessment (if both the Risk and Need scales are employed), the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), and the Level of Service Inventory—Revised (LSI–R). All three of these third-generation examples involve information that is gathered through a practitioner/client interview as well as through case file review and other sources of collateral information when available. In addition, all three use both static and dynamic predictors that have support in the extant literature base. All three also have support in the literature regarding their predictive validity (to varying degrees) as composite risk/need scales. However, each has its own limitations that indicate some room for improvement. For example, the Wisconsin assessment has scoring criteria that tend to inflate offenders' risk scores, and the COMPAS and LSI–R are limited in the extent to which they can be used on populations that have been incarcerated.

The fourth generation of actuarial assessments is set apart from the third generation by incorporating responsivity factors. Furthermore, these responsivity factors are addressed through the case-planning process, thereby enhancing responsiveness to treatment and supervision. Finally, these fourth-generation instruments explicitly link the identified needs (which put an offender at risk) to the case plan, which in turn increases the likelihood that meaningful criminogenic needs are identified and targeted (Bonta & Andrews, 2007; Bonta & Wormith, 2007; see also Andrews et al., 1990).

Risk classification has been a part of the correctional practitioner landscape for many decades. As noted above, the evolution of risk assessment has been (for the most part) scientifically driven, leading to better tools. Given the dynamic nature of both the offender population and the criminal justice system itself (e.g., changes in statutes and other legislation), every aspect of the system—including classification systems—should be consistently revisited and updated. The continuing effective implementation of the risk, need, and responsivity principles may depend on the development and validation (and revalidation) of new tools. Too often the "risk decision" is regarded as the most important if not sole decision point in offender case processing. The risk decision is not unimportant, but focusing solely on risk can cause officers to miss opportunities to intervene for offender change. To move beyond mere temporary control over or incapacitation of offenders, officers need to take into account all three principles (risk, need, and responsivity). Specifically, it is by these principles that officers make meaningful classification, formulate meaningful case plans, and execute meaningful attempts to compensate for barriers.

Meaningful case classification (i.e., risk assessment) entails a real and palpable agency or program response that is driven by risk. In short, it should be demonstrable that lower risk offenders experience intervention (whether through a controlling strategy such as supervision or through a rehabilitative program) at a level that is appropriate. The same holds true for higher risk offenders: Monitoring and adjusting the duration and intensity of the intervention are key for effective case processing.

Meaningful case planning means that relevant criminogenic needs drive the case plan, making it more likely that agencies will not waste resources targeting largely irrelevant circumstances. Meaningful case planning also requires the monitoring of change and progress. Likewise, meaningful attempts at removing barriers to effective treatment require going beyond merely identifying responsivity factors to using those factors in deciding how best to deliver needed services.

Post Conviction Risk Assessment Tool (PCRA): A Fourth-Generation Tool

For many years, the Risk Prediction Index was used in the federal probation and pretrial services system. External reviewers of the federal system suggested moving toward a third-generation risk assessment. Although a number of existing instruments were piloted and considered, the Administrative Office, with input from the field, opted to develop a fourth-generation risk and needs assessment based on federal data and guided by end-user needs and inputs. A detailed chronology of the need for the PCRA is provided in the "Construction and Validation of the Post Conviction Risk Assessment" by Johnson, VanBenschoten, Robinson, and Lowenkamp (2011).

The purpose of this study is threefold. First, we review the research presented in Johnson et al. (2011) with a more detailed discussion of methods and thereby introduce the validity of the PCRA to a new audience of practitioners, researchers, and academics. Second, we extend the initial study by introducing additional analysis with a new sample and data generated from training and certification procedures. Specifically, we assess interrater agreement of the instrument and provide a predictive validation using a small sample of assessments completed by trained officers. These new analyses enhance the knowledge of the validity of the PCRA. In addition, establishing an interrater method provides confidence in the implementation of the tool with a broader range of users. Third, we define the implications for the adoption and use of the PCRA and directions for future developments.

Method

Measures

As described in an earlier technical report (Johnson et al., 2011), multiple data sources and measures were used to construct and validate the PCRA. These data sources include federal presentence reports, existing risk assessments, criminal history record checks, and the Probation/Pretrial Services Automated Case Tracking System (PACTS; Johnson et al., 2011, p. 18).[1] Criminal history records or rap sheets were used to identify any new arrest after the start of supervision.

There are two sets of items included in the PCRA. The first consists of items rated by the

[1] PACTS is an electronic case management tool used by probation and pretrial services officers in all 94 federal districts to track federal defendants and offenders. At the end of each month, districts submit case data into a national repository that is accessible to the Administrative Office of the United States Courts, Office of Probation and Pretrial Services.

officer, which are given a numerical score that contributes to an offender's risk score. Rated and scored items used to develop the PCRA were theoretically derived and based on the extant empirical research in predicting criminal behavior (e.g., Andrews & Bonta, 2006; Gendreau, Little, & Goggin, 1996; Hubbard & Pratt, 2002; Simourd & Andrews, 1994). Based on a review of existing theoretical and empirical research, we selected data elements related to criminal history, peer associations, family, employment, substance abuse, and attitudes from PACTS. A series of bivariate and multivariate analyses were used to identify the factors to include in the initial instrument and the scoring of those items. Given the exploratory nature of the research, variables included on the PCRA had a significance level of .10 or below (see Table 1).

Ultimately, the scored items (see analyses below) originated from an exploration of five prevalent domains (prevalence determined by the existence of support in the extant literature): criminal history, education/employment, substance abuse, social networks, and cognitions. Several items within each domain were tested (see analyses below) for their individual predictive validity. The results of this testing indicated which items were the best candidates for inclusion in the domain subscales and which would in turn potentially contribute points to the overall composite scale. For example, criminal history examined factors related to number of arrests, history of violent offending (including domestic violence), whether or not there was a history of varied offending behavior on prior supervision, institutional adjustment, and age at intake to supervision. Items under education

Table 1
Multivariate Model Predicting Arrest During Initial Case Plan Period (Split Sample Construction Only)

Variable	*B*	*SE*	Wald	*df*	Sig.	Exp(*B*)
Community supervision violation	.34	.05	43.55	1	.00	1.41
Varied offending pattern	.23	.05	21.41	1	.00	1.25
Institutional adjustment	.23	.10	4.85	1	.03	1.26
Violent offending	.32	.08	16.31	1	.00	1.38
Unemployed	.37	.05	66.25	1	.00	1.45
Poor work outlook	.32	.06	27.50	1	.00	1.38
Alcohol problems	.48	.10	22.08	1	.00	1.62
Lacks social support	.27	.05	30.67	1	.00	1.31
Family problems	.19	.05	14.28	1	.00	1.21
Single	.10	.05	3.18	1	.08	1.10
Not motivated to change	.38	.05	59.80	1	.00	1.47
Drug problems	.71	.06	132.20	1	.000	2.03
Arrest history	.15	.02	50.54	1	.000	1.16
Age	.38	.03	136.61	1	.000	1.47
Educational attainment	.23	.05	27.20	1	.000	1.26
Mental health problems	.07	.05	1.92	1	.166	1.07
Gambling addiction	−.40	.28	1.95	1	.163	0.67
Criminal associates	−.08	.05	2.53	1	.11	0.92
Weapon concerns	−.09	.06	1.79	1	.18	0.92
Financial problems	−.07	.08	0.81	1	.37	0.93
Life skills deficiencies	−.02	.06	0.10	1	.75	0.98
Female	−.22	.06	13.59	1	.00	0.81
Race			3.11	4	.54	
Asian	.61	.49	1.57	1	.21	1.85
Black	.64	.47	1.87	1	.17	1.89
Native American/Eskimo	.67	.48	1.98	1	.16	1.95
White	.68	.47	2.15	1	.14	1.98
Constant	–4.54	.47	92.69	1	.00	0.01

Note. Model $\chi^2(26) = 1503.78$, $p < .000$; –2LL = 15868.80; Nagelkerke $R^2 = .12$. From "The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)," by J. Johnson, S. VanBenschoten, C. R. Robinson, and C. T. Lowenkamp, 2011, *Federal Probation*, *75*(2), p. 19. In the public domain. Reprinted with permission.

and employment included highest level of education achieved, degree of employment, number of jobs in the past 12 months, and general work history. Substance abuse included the extent to which substance use/abuse disrupted the work, school, or home environment; physically hazardous substance use; whether or not substance use was clearly related to legal problems; continuing use despite social or other problems; and a current alcohol problem and/or current drug problem. Social network included marital status, with whom the offender lives, lack of familial support, stability of the family situation, the criminogenic nature of peer networks, and lack of prosocial support systems. Cognitions included existence of antisocial attitudes/values and the offender's attitude toward supervision and/or change (a measure of motivation).

The second set of data elements comprises elements that are rated but not scored and do not contribute to an offender's risk score for the current study (a list of these items is available from the authors on request).[2] The rated but not scored items were considered for several reasons. First, several of them represent items that officers felt were necessary and important to the tasks they were trying to complete with the client. In brief, they constitute items that officers felt mattered in some way. Second, because responsivity is an underdeveloped area within correctional intervention, the rated but not scored items were included to facilitate future research with the hope of developing a more advanced rubric when implementing the responsivity principle. Third, the rated but not scored items also represent items that the agency does not currently collect as a matter of course. Future research and analyses may produce evidence suggesting that some of these items should be included as part of standard data collection procedures, as well as case planning.

Rearrest. Data measuring rearrest were gathered through the use of the National Crime Information Center and Access to Law Enforcement System databases. All names and identifiers included in the sample were referenced through the FBI. Personnel at the FBI conducted record checks, including the entire criminal history for each individual in the sample. These criminal histories were then sent back to the authors, who determined whether or not an arrest occurred after the PCRA had been given. The result was a dichotomous measure indicating whether an arrest occurred or not, postadministration of the PCRA.

Vignettes. We created vignettes to ensure that trained officers could accurately score assessments prior to applying the PCRA to an actual case. The 10 vignettes covered all four risk categories (low, low/moderate, moderate, high) and were based on actual cases from the U.S. probation system. They included a presentence report, criminal history, and scripted video-recorded interview with actors. The script was based on case file information (specifically, the intake procedure to probation). In this study, we used the vignettes to assess interrater agreement on the PCRA.

Participants

To construct and validate the PCRA, we devised three sample groups: one group for the construction of the instrument and two groups for the validation of the instrument.[3] These groups were created using an existing analysis file from PACTS data that contained 185,297 offenders on probation or supervised release.[4] The construction group was created from data

[2] Because of ongoing data collection, the test items have yet to be fully analyzed. Decisions to include or omit test items will be determined by statistical significance and by the impact of a given test item on the predictive accuracy of the PCRA.

[3] Two validation samples were developed to test the robustness of the instrument.

[4] Data from the analysis file were assembled from PACTS and matched with data from the Federal Bureau of Prisons, the U.S. Sentencing Commission, and the Census Bureau. Arrest data came from the Access to Law Enforcement System and from the FBI's Computerized Criminal History database. Arrest data are current through August 13, 2009. Offenders in the analysis file began active postconviction supervision between October 1, 2004, and August 13, 2009 (see Baber, 2010). Of the 185,297 offenders in the analysis file, only 103,071 had criminal histories and other relevant items used to construct the PCRA. These "lost" cases were largely cases that were immigrants who did not have a social security number or newer cases that did not have complete data (which would have excluded them from analyses regardless because of the inability to measure a follow-up period). Based on the number of cases that remained, we are confident that our results were not influenced.

obtained from the initial case plan.[5] Using a near 50/50 randomized split, data from the first case plan were divided into two sample groups; one became the construction sample and the other became the first validation group. One validation group (Validation) was taken from the initial case plan the offender receives during the term of supervision, and the second validation group was taken from subsequent case plans (hence the name Subsequent Case Plan). Both the construction (n = 51,428) and validation (n = 51,643) groups composed offenders who started a term of supervised release or probation on or after October 1, 2005, and included all offense categories. The subsequent case plan group comprised 193,586 case plan periods.

Procedure/Analytic Plan

We used a fairly straightforward and traditional approach in the development of the PCRA. Multivariate logistic regression models were used to determine which items were superfluous.[6] As a result, we reduced the total number of items included in the multivariate model to maintain statistical significance and ensure that the direction of the relationship between predictor and recidivism was intuitive for each item. Once the multivariate model was finalized, we used bivariate cross-tabulations to assign appropriate weights.[7] This method was chosen because of its transparency and because, to date, there is little research indicating the superiority of more complex weighting structures over dichotomous coding risk factors (see D. M. Gottfredson & Snyder, 2005; S. D. Gottfredson & Gottfredson, 1979; Harcourt, 2007; Silver, Smith, & Banks, 2000).

Once the final scoring algorithm was determined, we calculated a composite score for each case in the analysis. Cutoff scores were developed by visually inspecting the data. Although the data cutoffs were fairly evident, we tested alternative cutoffs with confirmation of best fit as determined by chi-square analysis. We then conducted a final set of analyses to determine how changes or stability in risk category from the beginning to the end of supervision were correlated with change in the probability of a new arrest.

In addition to the procedures noted above regarding the testing and development of PCRA items, we conducted analyses to test interrater agreement. The last portion of the analyses tested the interrater agreement for the PCRA. Officers scored a series of randomly assigned vignettes (see above) and percentage agreement was calculated.

Results

Table 1 displays the results of a multivariate model predicting arrest during the initial case plan period using a split sample from the construction sample. As Table 1 shows, many of the variables included in the multivariate model were statistically significant at the .001 level.[8] Odds ratios in the model also appear to be consistent with existing research that supports well-accepted beliefs that alcohol and drug problems, unemployment, poor attitude (not motivated to change), criminal history, and lack of social support increase an

[5] As outlined in the *Guide to Judiciary Policy,* Volume 8, Part E, Supervision of Federal Offenders, case plans are to be submitted within 30–60 days of the start of the offender's supervision term. This plan is formally evaluated and modified during the sixth month of supervision and updated annually for the duration of the supervision term.

[6] When the outcome variable comprises only two values (e.g., arrest or no arrest), which is typical for risk classification in probation, logistic regression is usually the best approach to use. The main advantage of logistic regression is that few statistical assumptions are required for its use. In addition, it generates probability values that are constrained between 0 and 1. Logistic regression calculates the probability of an event occurring or not occurring (e.g., getting arrested or not getting arrested) and presents the results in the form of an odds ratio (Exp(B)). For the purposes of this article, the odds ratio is the number by which you multiply the odds of getting rearrested for each 1-unit increase in the independent variable (i.e., a variable in the equation). An odds ratio greater than 1 indicates that the odds of getting rearrested increase when the independent variable increases; an odds ratio less than 1 indicates that the odds of getting rearrested decrease when the independent variable increases (Menard, 2002).

[7] It is interesting to note that race was not revealed as a significant predictor when controlling for other factors in the model. Furthermore, whereas additional bivariate tests revealed significant differences between racial groups regarding the PCRA scores, AUC analysis revealed nearly identical performance of the scale when predicting recidivism.

[8] Sex was a significant predictor in the construction analyses, with a negative parameter estimate. It is interesting, however, that once the scale was constructed, further multivariate analyses that included the Risk scale and gender did not reveal gender as a significant predictor. In addition, further analyses (not shown in this article) revealed that the Risk scale performed identically for males and females when predicting recidivism.

offender's chances of rearrest. Being female appears to decrease the likelihood of rearrest, which is also consistent with much of the existing research on gender and crime (Gendreau et al., 1996).

From the multivariate analysis, variables were selected for inclusion on the risk assessment instrument. To gain a better understanding of the bivariate relationships between the significant predictors in the multivariate model, we conducted a series of cross-tabulations. In general, the bivariate cross-tabulations allowed us to assign 1 or 2 points to each of the factors. Although this approach may seem counter to prevailing wisdom on the development of weights for risk assessment, there is evidence suggesting that this approach produces an instrument that still outperforms clinical approaches to prediction (Dawes, 1979) and is more robust across time and sample variations (D. M. Gottfredson & Snyder, 2005; McEwan, Mullen, & MacKenzie, 2009).

There are 15 scored items, all of which can contribute 1 point to the composite score, with the exception of age (2 potential points) and prior arrests (3 potential points). Theoretically, the PCRA score can range from 0 to 18. An analysis of descriptive statistics indicates that there are no significant differences in the length of the prediction period (time at risk) or average risk score for the construction sample and first validation sample (6.46 and 6.43, respectively). However, there are differences in the mean risk score and length of prediction period between the subsequent case plan sample and construction sample and subsequent case plan sample and first validation sample. The difference in mean risk score is a function of the higher risk cases failing in some regard (e.g., technical violation, new criminal activity), which would have a suppressing effect on mean risk score for the samples as a function of time (6.03 for the subsequent case plan sample vs. 6.44 for samples drawn from the first case plan period). The lower mean risk score might simply be a function of lower risk offenders surviving supervision to the third and subsequent case plan periods. At any rate, there could be some debate that the difference in risk scores is not practically significant, and this argument might be valid given that all three mean scores fall into the low-risk category. The difference in prediction periods is a matter of policy, as the first case plan period is approximately 6 months long (i.e., approximately 6 months elapse between the first case plan and the second case plan). Another 12 months (approximately) will elapse before the third case plan is conducted, or 18 months after the beginning of supervision.

Figure 1 presents the distribution of risk categories by the type of sample used.[9] Across all three samples, low- and low/moderate-risk offenders accounted for at least 85% of the cases, whereas high-risk offenders accounted for only 1%. There are no statistically significant differences between the construction sample and the validation sample at an alpha level of .01. However, there is a significant difference between the second validation sample (subsequent case plan) and the construction sample, as well as between the second validation and the first validation sample. This finding is likely an artifact of lower risk cases surviving longer on supervision and therefore being overrepresented in the subsequent case plan period sample. This finding might be primarily an issue of sample size rather than holding practical significance.

Area under the curve—receiver operating characteristics (AUC-ROC) was chosen as the measure to assess predictive strength of the PCRA.[10] This measure was selected in large part because it is not impacted by base rates or sample size. Finally, the AUC-ROC is a singular measure and does not have differing calculations depending on level of measurement of the variables being evaluated, as is the case with correlation coefficients (Rice & Harris, 2005). Table 2 displays the AUC-ROC between risk scores and rearrests for each of the samples included in this study. A fourth sample (long-term follow-up) that includes initial case plan

[9] Four risk categories were developed for two reasons. First, the data best supported (via the distribution of scores) the use of four distinct risk categories. Second, the agency wanted to identify the truly high-risk cases that (theoretically) would receive the most services and interventions. Likewise, the agency desired to identify the very low-risk cases that might be denied services. Further research needs to be done to determine what specific agency responses will be associated with specific risk categories (i.e., what the actual differences will be between low vs. low/moderate, low/moderate vs. moderate, and so on).

[10] The AUC measures the probability that a score drawn at random from one sample or population (e.g., offenders with a rearrest) is higher than that drawn at random from a second sample or population (e.g., offenders with no rearrest). The AUC can range from 0.0 to 1.0, with 0.5 representing the value associated with chance prediction. Values equal to or greater than 0.70 are considered good.

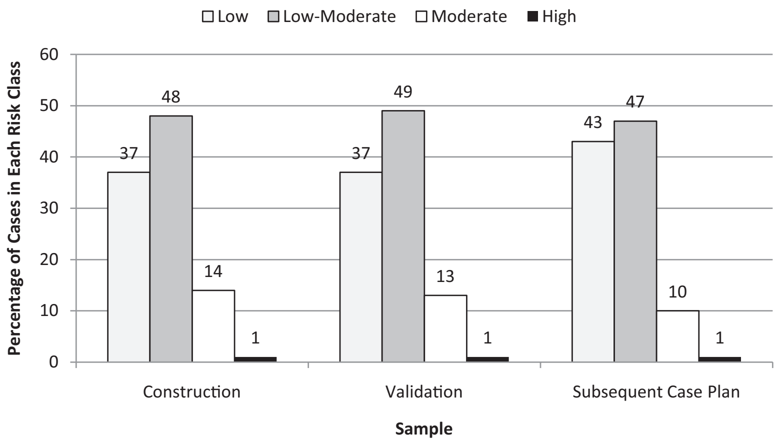

Figure 1. Percentage of cases in each risk category by sample.

data on a sample of offenders placed on supervision between September 30, 2005, and September 30, 2006, is introduced in Table 2. This sample allows for follow-up time periods that range between 3 and 4 years. Table 2 reveals that the AUC for each of the four sample groups is close to or exceeds the AUC-ROC value associated with large effect sizes (Rice & Harris, 2005). The AUC for the subsequent case plan sample rose to 0.73, and the AUC for the long-term follow-up sample rose to 0.78. Based on these results, the PCRA appears to have very good predictive validity when used to identify rearrest rates of groups of offenders included in these samples.

Table 2
Area Under the Curve—Receiver Operating Characteristics (AUC-ROC) Between Risk Score and Rearrest

Sample	AUC	Lower 95% CI	Upper 95% CI	Significance
Construction	0.71	0.70	0.72	.00
Validation	0.71	0.70	0.72	.00
Subsequent case plan	0.73	0.73	0.74	.00
Long-term follow-up	0.78	0.78	0.79	.00

Note. Analyses based on TSR versus probation supervision were estimated. AUC-ROC values for the probation subsample were 0.65 (construction), 0.64 (validation), 0.72 (subsequent case plan), and 0.76 (long-term follow-up). Although AUC-ROC values for the construction and validation samples were somewhat smaller than those generated for the overall sample, the AUC-ROC values for the subsequent case plan and long-term follow-up probation subsamples were very similar to those generated for the overall sample. From "The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)," by J. Johnson, S. VanBenschoten, C. R. Robinson, and C. T. Lowenkamp, 2011, *Federal Probation*, *75*(2), p. 19. In the public domain. Reprinted with permission.

To put the AUC values in practical terms, we calculated the failure rates by each category of risk for each sample.[11,12] These results are presented in Figure 2. With the exception of the long-term follow-up sample, the failure rates were relatively unchanged for a risk category across samples. For example, low/moderate-risk offenders failed at a rate of 13% in both the construction and initial validation samples and at 12% in the subsequent case plan sample. However, in the long-term follow-up sample, the low/moderate-risk group's failure rate increased significantly to 42%. Overall, the failure rate for the long-term follow-up group was 44%, but the failure rate was significantly higher for high-risk offenders in this same group. Moderate-risk offenders failed at a rate

[11] Rice and Harris (2005) indicate that the AUC holds the same meaning as the common language effect size indicator, that is, the probability that the PCRA score for a randomly selected recidivist is higher than the PCRA score for a randomly selected nonrecidivist. For example, using the long-term follow-up data (AUC = 0.78), if you randomly select a recidivist and a nonrecidivist, the recidivist's PCRA score should be higher than the nonrecidivist's score 78% of the time.

[12] *Failure* is defined as any new arrest during a term of supervision.

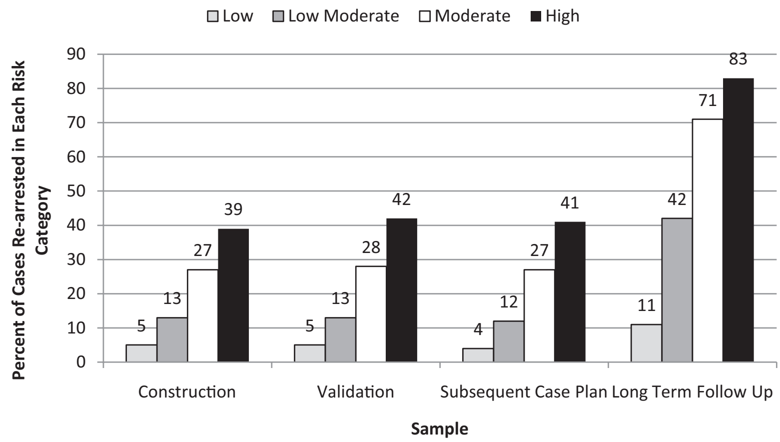

Figure 2. Rearrest rates by risk category and by sample. The outcome measure is arrest for new criminal behavior only. In the long-term follow-up, we were able to restrict our outcome to "arrest for new criminal behavior only." *Rearrest* means the offender could have been rearrested for a new offense, a technical violation, or for some other reason.

of 71% and high-risk offenders had an 83% failure rate. The uniform increase in failure rates across categories of risk and across the various samples continues to support the validity of the PCRA.

We analyzed survival for each risk category; the survival curves associated with those analyses are displayed in Figure 3. All possible data points, regardless of follow-up time, were used in the analysis.[13] The follow-up period ranged from 0 to 60 months. Survival rates for each risk category are displayed at 6 months, 12 months, 36 months, and 60 months. As Figure 3 shows, high-risk offenders have a very steep decrease in survival, as only 69% survived the first 6 months of supervision. As time passes, survival rates continue to drop rapidly for high-risk offenders, as only 46% survived at 12 months and only 17% at 36 months. After 60 months of supervision, a mere 6% of the high-risk offenders remained. In contrast to high-risk offenders, low-risk offenders have a significantly different experience on supervision. For example, whereas the survival rate for high-risk offenders was only 17% at 36 months, 90% of the low-risk offenders survived at this time period. Moreover, the survival rate for low-risk offenders decreased only 5 percentage points through 60 months to 85%.

Low/moderate-risk offenders have a survival curve that is almost precisely between the survival curves of the low- and moderate-risk cases. It is interesting that the survival curve for the moderate-risk offenders seems to follow a form that is closer to the high-risk offenders than to the lower risk offenders. Note that the survival rates continue to grow throughout the follow-up period for each group, and each curve (with the exception of low-risk offenders) shows little sign of leveling off.

The dynamic nature of third- and fourth-generation risk assessments is one of their greatest advantages. Many of the factors on the PCRA are considered to be dynamic, with some being stable and others acute (for a full discussion, see Serin, Lloyd, & Hanby, 2010; Serin, Mailloux, & Wilson, 2010). As such, we conducted analyses that took advantage of the initial and subsequent PCRA scores for a given individual. The purpose of these analyses was to determine whether changes in risk, as measured by the PCRA, are associated with failure rates that differ based on predicted failure rates according to the initial PCRA score. Table 3 lists the offenders' initial PCRA category in the left column and the last PCRA category, based on the last PCRA assessment, in the second row. Not surprisingly, Table 3 reveals that higher risk offenders fail at higher rates. Of greater interest, however, is the fact that the failure rate

[13] STATA adjusts for cases that were lost during follow-up when calculating survival tables.

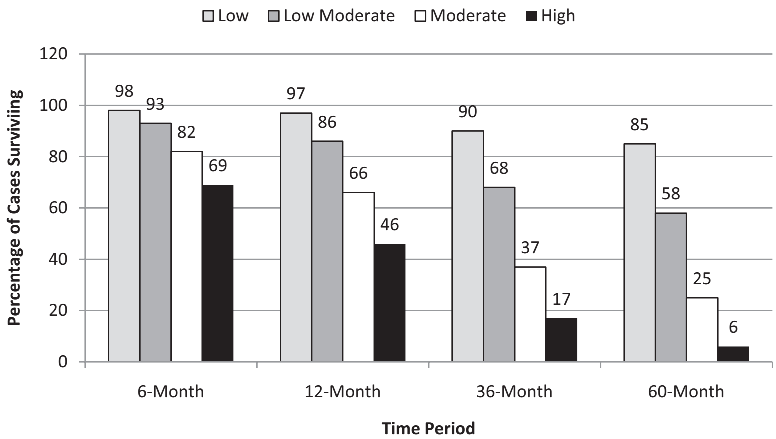

Figure 3. Survival analysis for the four risk categories: Percentage surviving at four time periods.

for a given risk category is variable and dependent on the offenders' last risk assessment category. That is, a moderate-risk offender who stays moderate risk belongs to a group of offenders that has a 38% failure rate. A moderate-risk offender who is later assessed as high risk, however, belongs to a group of offenders that has a 61% failure rate. Finally, a moderate-risk offender who is subsequently assessed as low/moderate belongs to a group of offenders that has an 18% failure rate. Similar trends are noted with the other risk categories presented in Table 3.

In the interest of further demonstrating the predictive validity of the PCRA, we conducted a preliminary prospective study. The PCRA was administered to a sample of 356 offenders, each of whom was tracked for over 1 year, using any new arrest as an outcome measure. We calculated the AUC-ROC value for this sample, which was well over the 0.70 threshold at 0.756. More prospective research needs to be done to further test the PCRA. Specifically, future prospective research may involve longer follow-up periods, more rigorous outcome measures, and larger samples, all of which will make survival analyses such as that in Figure 3 (as well as other analyses) feasible. In the meantime, however, this preliminary study further demonstrates the predictive validity of the tool.

As noted above, the last portion of our analyses involved a test of the interrater agreement regarding the PCRA (see Lowenkamp, Holsinger, Brusman-Lovins, & Latessa, 2004). Officers from three districts who had been trained to administer the PCRA were asked to score a series of vignettes that were randomly assigned. We calculated the percentage agreement across

Table 3
Changes in Failure Rates Based on First and Last Case Plan Assessment Categories

Initial case plan assessment category	Last case plan assessment category (%)			
	Low	Low/moderate	Moderate	High
Low (n = 13,589)	4	18	—	—
Low/moderate (n = 15,660)	5	16	41	—
Moderate (n = 3,581)	—	18	38	61
High (n = 233)	—	—	37	53
χ^2	237.65	396.23	162.85	10.54

Note. From "The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)," by J. Johnson, S. VanBenschoten, C. R. Robinson, and C. T. Lowenkamp, 2011, *Federal Probation*, *75*(2), p. 22. In the public domain. Reprinted with permission.

all 15 of the scored items on the PCRA, as well as the resulting classification (i.e., the percentage of officers that placed the offender in the same risk classification) and total composite scores. Table 4 presents the results of this interrater agreement analysis.

As presented in Table 4, the overall percentage agreement for all 15 items ranged from 87% to 98% across all four vignettes. Agreement was higher for the risk classification level assigned by each officer using the PCRA (ranging from 87% to 100% across the four vignettes). In addition, Table 4 presents the average composite score assigned to each vignette for all officers, as well as the standard deviations for each mean score. The standard deviation, being a measure of spread around the mean, indicates a relatively tight clustering of scores around each mean, ranging from a low of 0.56 to a high of 1.07. By all counts, the interrater agreement measured in this fashion indicates reliability in scoring across different officers who scored the same vignettes.

Discussion

As previously stated, the purpose of this article is threefold: (a) to review the research presented in Johnson et al. (2011) with a more detailed discussion, thereby introducing the validity of the PCRA to new audiences; (b) to extend the initial study by introducing new analyses and data; and (c) to define the implications for adoption and use of the PCRA and directions for future developments. This article has provided details on the methods, measures, and sample used in the development of the PCRA, which followed a rather traditional model. Our efforts were supported by a relatively large data set and fairly complete data. The sample was largely representative of the population served and allowed for one construction and two validation samples. The overall results have demonstrated that the PCRA provides adequate predictive validity both in the short term (6–12 months) and in longer follow-up periods (up to 48 months).

Multivariate analysis (see Table 1) of proposed predictors revealed that 15 factors were significantly related to the outcome of interest (new arrest). Seven additional factors tested were determined to be unrelated to the prediction of a new arrest once the effects of the other factors were controlled. One additional measure, being male, was found to be significantly related to a new arrest. Subsequent models, not reported here, indicated that adding gender to the models yielded no increase in the explanatory power of the model. In addition, we noted nonsignificant differences in the AUCs between males and females for each sample (i.e., construction, validation, subsequent case plan, and long-term follow-up). Therefore, we concluded that the instrument performs equally well for males and females, even though the failure rates for males might be slightly higher than for females with similar risk scores.

The creation of the risk score and categories allowed for the identification of four risk categories: low, low/moderate, moderate, and high. Approximately 80% of each sample was made up of low- and low/moderate-risk offenders. Much smaller percentages were identified in each sample as moderate and high risk (approximately 12% and 1%, respectively). Because the distribution of risk categories was heavily skewed toward lower risk, the validity of the instrument may be questioned. However, a current validated risk prediction instrument used in the federal system (Risk Prediction Index) yields a similarly skewed distribution. Analysis of failure rates by risk score and category using the PCRA yielded AUC-ROC values over the traditionally accepted value of 0.70 and an AUC value for the long-term follow-up over 0.78. All of the AUC-ROC values were close to or exceeded the value associated with large effect sizes. Practically speaking, the instrument provided categorizations that are associated with the group failure rates that are differentiated and meaningful for meeting the risk principle (see Table 4).

The next-to-last analysis conducted in this study related to the dynamic nature of the PCRA. Recall from Table 3 that changes in actual failure rates were associated with

Table 4
Measures of Interrater Agreement Using Four Vignettes

Vignette	*n*	Agreement (%)	Same classification (%)	Average score	*SD*
1	26	87	96	15.11	1.07
2	25	90	96	13.20	0.71
3	31	87	87	13.09	1.04
4	30	98	100	0.40	0.56

changes in risk category from the initial assessment to the last assessment. This finding is important, as it provides the opportunity to track meaningful changes in risk that occur throughout the supervision process. Moreover, Table 3 confirmed that the PCRA identifies and measures dynamic risk factors that, when changed through supervision, services, or some other unmeasured process (natural desistance), apparently lead to commensurate reductions in actual failure rates. The dynamic nature of the PCRA adds to its usefulness in developing case plans throughout the life of the supervision term.

Although this study was rather comprehensive in scope and the data set used was large and representative of the population served, there are some limitations and areas for future research that deserve mention. First, as with many studies of this type, the data used in this research were archival. The PCRA should undergo future (larger) validation research in a prospective fashion. Likewise, future prospective validation research should use varied measures of outcome (e.g., reconviction, reincarceration, severity of offense). Future research may also reveal ways in which the PCRA could be made more dynamic or lead to the development of trailer instruments that are more sensitive to change, making the assessment of offender change and progress more feasible. Similarly, the rated but not scored items need further examination, particularly because the current version of the PCRA includes only one item that measures antisocial cognitions.

Notwithstanding the limitations discussed above, some important policy implications stem from this research. First, the federal probation system now has a dynamic fourth-generation risk assessment for use with offenders under its jurisdiction. The instrument can be used to identify higher risk offenders for enhanced services (see Andrews et al., 1990), targets for change to be addressed by external service providers, and several responsivity factors. Use of the PCRA will enhance the building of case plans, in which major criminogenic targets can be identified. This will in turn enable agencies to respond to these targets (ideally with effective programming and therapeutic interventions). In addition, all users of the PCRA have to be certified via standardized training. Certified users also have to be recertified through a mandatory retraining (along with testing and recertification) every 2 years to guard against rater drift and knowledge decay. Other third-generation risk/need assessment processes do not necessarily include recertification training (although this varies across agencies, jurisdictions, and states).

Risk classification (e.g., high, medium, low/moderate, low) is not new in the field of correctional supervision and intervention. However, given the comprehensive nature of the items examined, including the responsivity considerations, use of the PCRA can increase the utility (in terms of agency response) of the high, medium, low/moderate, low classification rubric. The validity of the tool, perhaps relative to other classification tools, may provide more confidence in classification, allowing agencies to respond accordingly by streamlining resources and increasing the integrity of the implementation of the risk principle.

Case planning and intervention are areas in which the need principle comes into play. Appropriate case planning involves assessing criminogenic needs, documenting the targeting of those needs, and monitoring progress and change during and after treatment intervention. The PCRA clearly and automatically identifies the most prevalent criminogenic needs that are currently at work in the offender's life and environment. Once these criminogenic needs are identified, they can be targeted, again increasing the integrity of the application of the need principle on both the individual and agency levels.

Furthermore, the dynamic nature of the PCRA should facilitate the measurement of change at the offender level, indicating whether or not offenders are making changes in their lives and cognitions that in turn should be related to behavior (for better or worse). Relatedly, measuring change at the offender level may also aid in case processing. For example, if an offender is clearly making progress, evidence from the PCRA may allow for reductions in supervision level or even early termination (depending on the administrative restrictions of the agency).

At the agency level, in the aggregate, data from the PCRA should help determine whether or not offenders (as a group) who pass through specific programming are changing in palpable if not substantial or statistically significant ways. In short, data from the PCRA may become useful as an important part of program evaluation. Determining which programs are making the most progress with the offenders who are assigned to them is an important step toward identifying which programs are worthy of continued support.

Data analyzed in this study indicate that changes in levels of risk are associated with changes in actual failure rates. Therefore, officers need to monitor risk in a standardized way to ensure that supervision and services are having intended impacts. If intended impacts are not achieved, then officers can modify supervision services to reduce the risk of recidivism. Future research should also focus on the PCRA and its ability to predict dangerousness (e.g., violent or sex offenses). In the current research, we assessed the PCRA composite score as it relates to general propensity to engage in recidivistic behavior. Still to be done is an investigation into how well the PCRA may predict violent or sex offenses. At the same time, based on the survival analyses presented above, an examination of the moderate- and high-risk cases indicates that a “front-loading” approach may be best. In other words, given the high failure rate of both the moderate- and high-risk categories of offenders, it might be best in these cases to create policy that brings as many (relevant) services to bear right away. Perhaps doing so will stave off failure for some, although doing so is indeed dependent on validly identifying them early through the risk/need assessment.

Regarding responsivity, the PCRA will greatly enhance the agency and line-level practitioner’s ability to identify and (ideally) remove barriers to treatment. The assessment of responsivity is greatly underdeveloped in correctional practice in the United States. Often a best-case scenario will involve staff merely being able to identify what responsivity is as a construct, and perhaps name one or two prevailing responsivity considerations. Using the PCRA will help both the individual practitioner and the agency identify specific relevant responsivity considerations. Depending on agency response, putting information regarding responsivity into action—making it an active and defining part of the case plan and intervention—may help remove what otherwise would have been barriers to successful intervention. Too often offenders with criminogenic needs are placed in specific interventions designed to address that particular criminogenic need; however, if active responsivity items are in play and these are not also addressed, the treatment resources run the risk of being wasted. Use of the PCRA will increase the likelihood that these barriers will be identified, making it possible for the agency or program to address them.

References

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory—Revised*. Toronto, ON: Multi-Health Systems.

Andrews, D., & Bonta, J. (1998). *The psychology of criminal conduct* (2nd ed.). Cincinnati, OH: Anderson Publishing.

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

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

Baber, L. (2010). Results-based framework for post-conviction supervision recidivism analysis. *Federal Probation, 74*(3), 5–10.

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

Bonta, J., & Andrews, D. A. (2007). *Risk–need–responsivity model for offender assessment and rehabilitation*. Ottawa, ON: Public Safety Canada.

Bonta, J., & Wormith, S. (2007). Risk and need assessment. In G. McIvor & P. Raynor (Eds.), *Developments in social work with offenders* (pp. 131–152). London, England: Jessica Kingsley.

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. Bruce, E. Burgess, & A. Harno (Eds.), *Prediction and classification: Criminal justice decision making* (pp. 205–239). Chicago, IL: University of Chicago Press.

Connolly, M. M. (2003). *A critical examination of actuarial offender-based prediction assessments: Guidance for the next generation of assessments* (National Criminal Justice Reference Service No. NCJ 202982). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. doi:10.1037/0003-066X.34.7.571

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34,* 575–608. doi:10.1111/j.1745-9125.1996.tb01220.x

Gottfredson, D. M. (1987). Prediction and classification in criminal justice decision making. In D. M. Gottfredson & M. Tonry (Eds.), *Prediction and classification: Criminal justice decision making* (pp. 21–51). Chicago, IL: University of Chicago Press.

Gottfredson, D. M., & Snyder, H. N. (2005). *The mathematics of risk classification: Changing data into valid instruments for the juvenile courts* (National Criminal Justice Reference Service No. NCJ 209158). Washington, DC: National Center for Juvenile Justice, Office of Juvenile Justice and Delinquency Prevention.

Gottfredson, S. D., & Gottfredson, D. M. (1980). Screening for risk: A comparison of methods. *Criminal Justice and Behavior, 7,* 315–330. doi:10.1177/009385488000700306

Harcourt, B. (2007). *Against prediction*. Chicago, IL: University of Chicago Press.

Hubbard, D. J., & Pratt, T. C. (2002). A meta-analysis of the predictors of delinquency among girls. *Journal of Offender Rehabilitation, 34,* 1–13. doi:10.1300/J076v34n03_01

Johnson, J., VanBenschoten, S., Robinson, C. R., & Lowenkamp, C. T. (2011). The construction and validation of the federal Post Conviction Risk Assessment (PCRA). *Federal Probation, 75*(2), 16–29.

Lowenkamp, C., 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*(3), 34–38.

McEwan, T. E., Mullen, P. E., & MacKenzie, R. (2009). A study of the predictors of persistence in stalking situations. *Law and Human Behavior, 33,* 149–158. doi:10.1007/s10979-008-9141-0

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques*. Beverly Hills, CA: Sage.

O'Rourke, M. (2008). *Risk assessment: An overview*. Retrieved from http://www.stateclaims.ie/ClinicalIndemnityScheme/publications/RiskAssessmentOverview.pdf

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

Serin, R., Lloyd, C., & Hanby, L. (2010). Enhancing offender reentry: An integrated model for enhancing offender reentry. *European Journal of Probation, 2,* 53–75.

Serin, R., Mailloux, D., & Wilson, N. (2010). *Practice manual for use with Dynamic Risk Assessment for Offender Reentry (DRAOR) scale: Revised version*. Ottawa, ON: Carleton University.

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior, 27,* 733–764. doi:10.1177/0093854800027006004

Simourd, L., & Andrews, D. A. (1994). Correlates of delinquency: A look at gender differences. *Forum on Corrections Research, 6,* 26–31.

VanBenschoten, S. (2008). Risk/needs assessment: Is this the best we can do? *Federal Probation, 72*(2), 38–42.

Van Voorhis, P., & Brown, K. (1996). *Risk classification in the 1990s*. Washington, DC: U.S. Department of Justice, National Institute of Corrections.

Wardlaw, G., & Millier, P. J. (1978). The psychologist's role on a New Zealand prison classification committee. *International Journal of Criminology and Penology, 6,* 185–190.

Received September 27, 2011
Revision received August 3, 2012
Accepted August 10, 2012 ■