SEPTEMBER 2016

Federal PROBATION

a journal of correctional philosophy and practice

SPECIAL FOCUS ON: EXPLORING RISK OF VIOLENCE

Federal PROBATION

a journal of correctional philosophy and practice

PUBLISHED BY
The Administrative Office of the U.S. Courts

James C. Duff, *Director*

Matthew G. Rowland, *Chief*
Probation and Pretrial Services Office

Federal Probation ISSN 0014-9128 is dedicated to informing its readers about current thought, research, and practice in criminal justice, community supervision, and corrections. The journal welcomes the contributions of persons who work with or study defendants and offenders and invites authors to submit articles describing experience or significant findings regarding the prevention and control of crime and delinquency. A style sheet is available from the editor.

Federal Probation is published three times yearly—in June, September, and December. Permission to quote is granted on the condition that appropriate credit is given the author and *Federal Probation*. For information about reprinting articles, please contact the editor.

Subscriptions to *Federal Probation* are available from the Superintendent of Documents of the Government Printing Office at an annual rate of $16.50 ($22.40 foreign). Please see the subscription order form on the last page of this issue for more information.

Federal Probation can also be accessed online at no charge at www.uscourts.gov.

ADVISORY COMMITTEE

members

Dan Richard Beto
National Association of Probation Executives
Huntsville, Texas

Guy Bourgon
Public Safety Canada
Ottawa, Canada

James Byrne
University of Massachusetts, Lowell
Lowell, Massachusetts

Honorable James G. Carr
United States District Court
Toledo, OH

Alvin W. Cohn
Administration of Justice Services, Inc.
Rockville, Maryland

Ronald P. Corbett, Jr.
University of Massachusetts Lowell
Lowell, Massachusetts

Thomas Henry
Seton Hall University
South Orange, New Jersey

Edward J. Latessa
University of Cincinnati
Cincinnati, Ohio

Honorable David D. Noce
United States District Court
St. Louis, Missouri

Daniel B. Ryan
Justice Solutions Group
Lakeville, Minnesota

Faye Taxman
George Mason University
Fairfax, Virginia

EDITORIAL STAFF

Nancy Beatty Gregoire, *Executive Editor*
Ellen Wilson Fielding, *Editor*

Federal Probation
Administrative Office of the U.S. Courts
Washington, DC 20544
telephone: 202-502-1651
fax: 202-502-1677
email: Ellen_Fielding@ao.uscourts.gov

Postmaster: *Please send address changes to the editor at the address above.*

THIS ISSUE IN BRIEF

This September's issue of *Federal Probation* features a special section devoted to "Exploring Risk of Violence" in the federal probation system. Although general risk assessment instruments such as federal probation's Post Conviction Risk Assessment (PCRA) do a good job of evaluating risk of recidivism of any kind, the desire to reduce the likelihood of serious harm to the community from the violent acts of those under community supervision has led to a search for better dynamic prediction of the risk of violence. The five articles in this section discuss applying the risk principle—including risk of violence indicated by a dynamic risk assessment—to supervision, using a multi-level risk assessment in case planning, analyzing how dangerous sex offenders under federal supervision are, and exploring the idea of using sentinel event reviews in the federal system.

Among the articles outside the "special focus" topic of this issue is a rejoinder to a high-profile ProPublica article that challenged the widely used COMPAS risk assessment instrument as racially biased. In addition, this issue includes an article considering use of risk information at criminal sentencing in a state system, and finally, back in the federal system, an examination of how changes in offender risk characteristics are related to recidivism outcomes.

—Ellen Wilson Fielding
Editor, *Federal Probation*

SPECIAL FOCUS ON: Exploring Risk of Violence

DEPARTMENTS

The articles and reviews that appear in *Federal Probation* express the points of view of the persons who wrote them and not necessarily the points of view of the agencies and organizations with which these persons are affiliated. Moreover, *Federal Probation*'s publication of the articles and reviews is not to be taken as an endorsement of the material by the editors, the Administrative Office of the U.S. Courts, or the Federal Probation and Pretrial Services System.

The Real-World Application of the Risk Principle[1]: Is It Possible in the Field of Probation?

Scott W. VanBenschoten
Probation and Pretrial Services Office
Administrative Office of the U.S. Courts
John Bentley
Chief United States Probation Officer, District of South Dakota
Nancy Beatty Gregoire
Probation and Pretrial Services Office
Administrative Office of the U.S. Courts
Christopher T. Lowenkamp
Probation and Pretrial Services Office
Administrative Office of the U.S. Courts

ALTHOUGH SEVERAL CRIMINAL justice theories exist as a roadmap to effective supervision, the Risk, Needs, and Responsivity (RNR) model dominates the scholarly literature. A quick search in the Criminal Justice Abstracts Database reveals 140 peer-reviewed publications referencing the RNR model since 2000. As Andrews and Bonta (2007) note, the RNR model has been used, with increasing success, in North American and around the world. The authors further comment on the robustness of the model, but observe (2007:15) that "The greatest challenge is transferring the RNR model into 'real world' settings." Research on correctional services and the risk principle certainly supports this contention.

In a review of 38 correctional programs, researchers found only 1 program that met the criteria for varying programming intensity and duration by risk (Lowenkamp, 2004). Similarly, Lovins (2012) found that 36 out of 134 correctional treatment programs reviewed were varying program duration by risk. Finally, Lowenkamp, Pealer, Smith, and Latessa (2004) found that approximately 33 percent of supervision-based programs in Ohio were targeting high-risk offenders or varying program duration by risk or program intensity by risk. Further, only four programs were meeting all three of these criteria. Echoing Andrews and Bonta (2007), it does seem that translating the RNR model in real-world settings is a challenge and has been quite elusive.

A recent attempt to bring the RNR model to probation supervision was presented in a monograph titled *Dosage Probation* (Center for Effective Public Policy, 2014). This model drew upon the extant research on the RNR model as well as the emerging (but limited) research on dosage. One of the aspects of this model focused on setting correctional service thresholds by risk level. That is, high-risk offenders would receive 300 hours of treatment, which would require longer periods of supervision compared to moderate-risk offenders, who would receive 100 hours of treatment, thereby requiring shorter periods of supervision. This in effect would lead to varying the duration and intensity of services by risk level for offenders placed on supervision. A demonstration project of this model was developed and carried out. The authors of the evaluation found that the model had no effect on offender outcomes; however, the model (either due to complexity or some other reason) was not fully implemented (Lowenkamp, Holsinger, & Bechtel, 2016). Although it may intuitively seem easy, the struggle to properly implement the risk principle lies in the details of implementation and the availability of resources.

In the federal probation system, there has been a concerted effort to align supervision practices with the RNR model since 2009. This article examines the system's effectiveness in implementing the Risk Principle. Additionally this paper examines the adoption of a violence assessment and how this can further refine the use of this Risk Principle.

[1] The Risk Principle was introduced by Andrews, Bonta, and Hoge (1990) as a way to see the intervention of supervision through the prism of psychological principles. More specifically, they proposed three principles that make up the foundation of effective correctional practice: the Risk, Needs, and Responsivity Principles. This article is examining only the implementation of the Risk Principle, but it is important to note that this principle is only one part of an interdependent model.

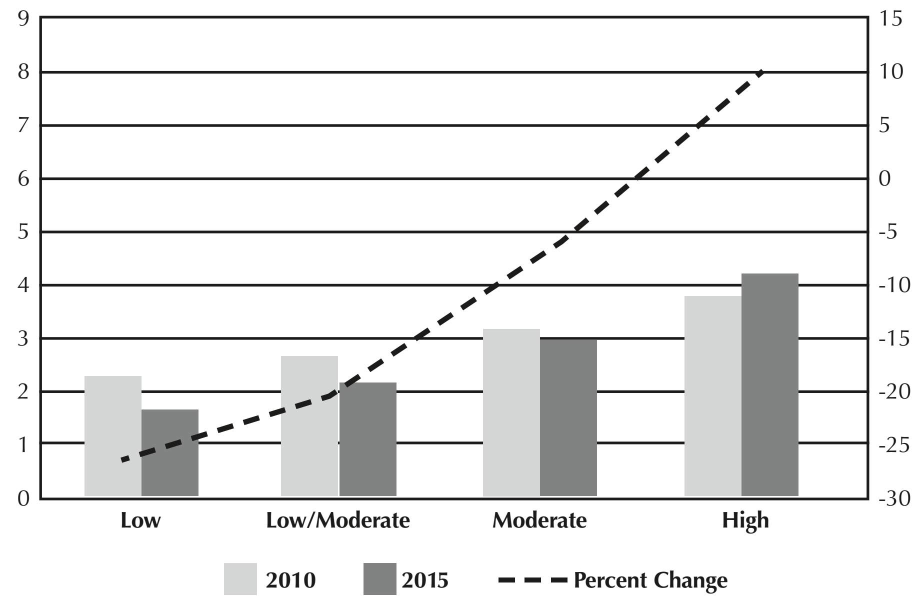
History of the Risk Principle in the Federal System

The federal probation system has a long history of using various risk prediction assessments. Individual districts used various forms of standardized risk assessment instruments throughout its history. In 1982, the Federal Judicial Center (FJC) collected survey data from across the federal system and learned that over two dozen risk prediction tools were in use (Vance, 2011). Soon after, the FJC created the RPS-80 for the entire federal probation system. The RPS-80 evolved into the RPI, which was the system's main risk assessment through 2010 (Vance, 2011).

In 2008, the Administrative Office of the United States Courts (AOUSC) embarked on the creation of a fourth-generation risk assessment instrument for federal probation. In 2010, implementation of the Post Conviction Risk Assessment (PCRA) began. As Vance (2011) recounts, over the course of 18 months, every chief probation officer, supervisory probation officer (over a supervision unit), and federal probation supervision officer was trained in the RNR model. Chief probation officers were trained through telephone calls with AOUSC staff combined with a group of pre-trained deputy chiefs. During the same time period, 94 conference calls were held where the group reviewed the risk principle and examined each district's data regarding their current application of the risk principle. Supervisors were trained in two larger regional training events that included material on the RNR model and the risk assessment instrument. Officers received training in one of dozens of regional events in a manner similar to supervisors (Vance, 2011).

Simultaneously, the AOUSC reviewed and revised its supervision policy to further comport with the RNR model. The first step of this revision expanded the pool of low-risk candidates eligible for less intense supervision by officers, and more clearly spelled out the reduced supervision requirements for those in that pool. This revision was aimed at better alignment of policy with the RNR model, but also provided an opportunity to take unnecessary tasks off the plates of officers to provide them with the time necessary to learn and implement this new assessment instrument. The rationale was that freeing officers from certain requirements for low-risk offenders would allow them to focus more on the higher-risk offenders (Vance, 2011).

FIGURE 1.
Change in the Median Number of Monthly Contacts and Percentage Change in Median Monthly Contacts from 2010 to 2015 by Risk Level



Did Policy Change Impact Officers' Behavior?

As expected, the federal probation system has seen significant change in practice, both in terms of officers' efforts and in use of treatment resources. Figure 1 illustrates this change in practice by showing the percentage change in the days between officer/offender contacts. These data were extracted from the case management system used by federal officers and show that officers are increasing the amount of time between contacts with lower-risk persons on supervision, and therefore spending fewer resources on them, while more frequently having contact with those who are at higher risk (see Cohen, Cook, & Lowenkamp, 2016).

Figure 2 displays the percentage change in daily treatment costs from 2009 to 2015 by risk level. More treatment money was allocated for higher-risk persons on supervision and fewer resources for lower-risk people on supervision.

These data show marked improvement in the federal probation systems' adherence to the risk principle. By providing a new risk instrument and supporting training on the research behind the risk principle, while simultaneously removing burdensome requirements that arguably add little to the goal of recidivism reduction, officers were able to shift resources in support of the risk principle.

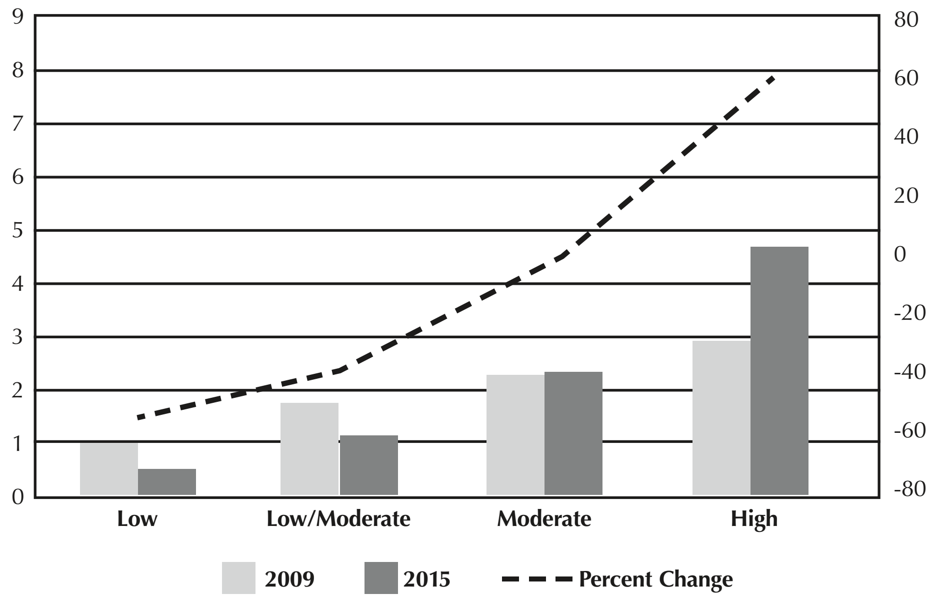
The federal probation system policy for decades has been that supervision should be individualized. There has long been an expectation that officers spend proportionately more time and energy on higher-risk persons. The new policy, though, removed specific task requirements and broadened eligibility for low-risk supervision.

Did Reduced Attention to Low-Risk Persons Put the Community at Risk?

This, of course, leads us to ask whether the reduction in supervision resources, both in officer time and in treatment dollars, had an impact on recidivism rates of low-risk persons. A recent article (Cohen, Cook, & Lowenkamp, 2016) sheds some light on this question. While not a direct test of this shift in the allocation of resources based on risk, Cohen et al. (2016) evaluated various outcomes for time periods before and after the implementation of a low-risk supervision policy that directed districts to spend fewer resources on low-risk offenders. The authors concluded that:

> In general, findings are supportive of the low-risk policy. This research shows that low and low/moderate risk offenders in the post policy group have fewer officer/offender contacts compared to their pre-policy counterparts (Cohen et al., 2016). This finding suggests that the low-risk policy is influencing officer behavior by encouraging federal officers to engage in fewer interactions with offenders on the lower end of the

FIGURE 2.
Change in Treatment Dollars Spent Per Day by Risk Level from 2009 to 2015



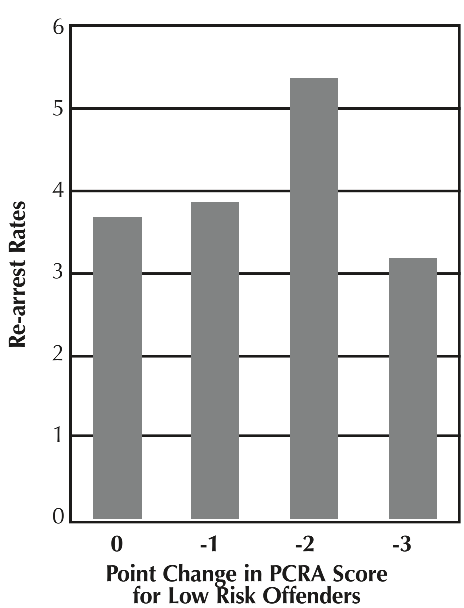
risk continuum. Importantly, the policy of supervising low-risk offenders less intensively has not compromised community safety. Post-policy low-risk offenders were no more likely to recidivate compared to their pre-policy counterparts. This finding indicates that federal officers can spend less time and resources on low-risk offenders without an accompanying rise in their recidivism rates.

Was There a Benefit to Continued Supervision at a More Intense Rate for Low-Risk Persons?

Even with this renewed commitment to the adoption of the risk principle and data to support this commitment, there were some probation officers who struggled to let go of lower-risk offenders on supervision. The officers anecdotally reported that these individuals had issues that needed attentive supervision to further lower their risk. Officers have access to community resources and may feel compelled to connect those on supervision with community resources. We wanted to explore whether there was a benefit to providing supervision resources to those individuals.

Cohen, Lowenkamp, and VanBenschoten (2016) looked at this question and determined that individuals who started in the low-risk category and whose risk score lowered even further did not have lower rates of recidivism. (See Figure 3, which displays the one-year rearrest rate for low-risk offenders by change in their second PCRA assessment.) Low-risk offenders that evinced no change in risk at their second assessment had a 3.8 percent rearrest rate. Low-risk offenders whose PCRA score dropped at their second assessment had rearrest rates of 3.9 percent, 5.2 percent, and 3.2 percent (for those low-risk offenders whose PCRA score dropped by 1, 2, and 3 points respectively).

FIGURE 3.
Twelve-Month Rearrest Rate for Low-Risk Offenders By Change in PCRA Score Time 1 to Time 2



In short, acceptance of the risk principle as applied in the federal probation system's policy has been a success. Rearrest rates of low-risk offenders have remained steady, while officers' time and treatment dollars have been protected for higher-risk cases.

Enhancing Adherence to the Risk Principle: The Identification of Violence

Although the PCRA does a solid job of predicting risk of general recidivism and revocation, it was not built to maximize predictions about the likelihood of violence. In 2014 research began to better identify which individuals on supervision are at an elevated risk of violence. This research included a large-scale data collection effort from case files of individuals who failed on supervision due to a violent act (see Lowenkamp, Johnson, Trevino & Serin in this issue). This research, in combination with research on the introduction of "due diligence" and case level assessment (see Serin, Lowenkamp, Johnson, & Trevino in this issue) provide a means for accurate and ongoing assessment of an offender's risk of violent offending. More specifically, officers will be equipped with a static estimate of violent reoffending that is based on 14 markers of risk for violence. These 14 markers, in conjunction with the PCRA score, generate AUC-ROC values of roughly 0.80 when predicting rearrest for a violent offense.

Once the violence assessment was complete, the AOUSC tasked a group of probation officers, supervisors, deputy chiefs, and chiefs to operationalize the use of the violence prediction data, in combination with the PCRA, to create supervision contact standards. This group struggled with the concept of contact standards and drew from a position paper developed in the District of South Dakota to expand the use of risk assessment into a more comprehensive supervision dosage document. Would it be possible to further enhance the adoption of the risk principle by dividing out those in the lower risk categories who have an elevated risk of violence? Would this new tool allow the system to place more individuals on low-risk supervision in order to focus additional resources on the higher-risk populations without compromising community safety?

The initial task this group faced was to take the violence tool in combination with the PCRA risk level and determine the proper amount of dosage in the three categories that

TABLE 1.
Re-arrest Rates for Any Crime and Violent Crime by PCRA/Violence Risk Category

PCRA Risk of Recidivism	PCRA Risk to Commit a Violent Act: Category 1	Category 2	Category 3
Low	L/1 (white) Any Crime = 9% Violent Crime = 1%	L/2 (white) Any Crime = 5% Violent Crime = 0%	L/3 (yellow) N/A
Low/Mod	LM/1 (green) Any Crime 23% Violent Crime = 2%	LM/2 (yellow) Any Crime = 29% Violent Crime = 8%	LM/3 (orange) Any Crime = 42% Violent Crime = 16%
Moderate	M/1 (yellow) N/A	M/2 (orange) Any Crime = 43% Violent Crime = 11%	M/3 (red) Any Crime 54% Violent Crime 21%
High	H/1 (red) N/A	H/2 (red) N/A	H/3 (red) Any Crime = 53% Violent Crime = 24%

TABLE 2.
Supervision Matrix Rehabilitation, Monitoring, Intervention Level Recommendations

PCRA Risk of Recidivism	PCRA Risk to Commit a Violent Act: Category 1	Category 2	Category 3
Low	L/1 (white) Low Risk Supervision Caseload	L/2 (white) Low Risk Supervision Caseload	L/3 (yellow) Monitoring: Elevated Restrictions: Responsive to Circumstances Interventions: Moderate
Low/Mod	LM/1 (green) Monitoring: Basic Restrictions: Responsive to Circumstances Interventions: Responsive to Circumstances	LM/2 (yellow) Monitoring: Basic Restrictions: Responsive to Circumstances Interventions: Minimal	LM/3 (orange) Monitoring: Elevated Restrictions: Intermediate Interventions: Moderate
Moderate	M/1 (yellow) Monitoring: Elevated Restrictions: Responsive to Circumstances Interventions: Moderate	M/2 (orange) Monitoring: Elevated Restrictions: Intermediate Interventions: Moderate	M/3 (red) Monitoring: Intense Restrictions: Intense Interventions: Intense
High	H/1 (red) Monitoring: Intense Restrictions: Intense Interventions: Intense	H/2 (red) Monitoring: Intense Restrictions: Intense Interventions: Intense	H/3 (red) Monitoring: Intense Restrictions: Intense Interventions: Intense

drive supervision: Monitoring, Restrictions, and Interventions (MRI). The grouping of the twelve cells in the matrix into five categories was based, in part, on rearrest rates for any crime and for a violent crime. Those numbers are presented in Table 1. The five categories of risk include the following cells from Table 1: white (cells L/1 and L/2), green (cell LM/1) yellow (cells L3, LM/2, and M/1), orange (M/2 and LM/3), and red (M/3, H/1, H/2, and H/3).

Although not binding, the advisory group provided examples of supervision levels (MRI) for each cell of Table 1. These examples of supervision levels are presented in Table 2 and direct, in a general way, districts and officers to focus on those offenders at higher risk of being arrested for any new offense and particularly those offenders at higher risk of being arrested for a violent offense. The advisory group also encouraged each district to think through what each of these levels might mean within their district and determine their local supervision standards.

When providing community-based supervision to those convicted of a federal offense, the safety of the community is paramount. Community safety is compromised by new criminal conduct committed by those under supervision and the harm caused by new offenses. Therefore, a person's risk to commit a more harmful act should be measured along with the person's risk to commit any criminal act. More resources and higher supervision levels are necessary to respond to someone who has demonstrated or has been assessed as likely to cause more serious harm should they reoffend.

As mentioned above, accurate assessment of risk to reoffend through valid actuarial instrumentation is standard practice in the federal probation system and provides the foundation to implement proven ways to reduce the likelihood of reoffending. Assessing the likely harm that might result from reoffending or from other negative behaviors is also essential to community safety and should be standard practice once valid actuarial instruments that predict harmfulness are developed and implemented. Just as marketing companies target their potential customers on both likelihood of any purchase and likelihood of an expensive purchase, so must probation officers target based on both general risk and expected severity.

As the AOUSC continues to develop national policy and procedures related to targeting supervision strategies for those at a higher risk of committing a violent offense, we wanted to consider whether our expected increased requirements for officers supervising these persons could be accomplished with current staffing levels. That is, can the current number of total contacts officers have with persons under supervision be shuffled even further away from the lower-risk persons to satisfy the increased expectations for officers supervising these newly identified violent offenders? Similarly, can existing treatment resources be re-allocated from lower-risk offenders to offenders with relatively higher risks of rearrest for a violent offense?

To answer that question, AOUSC conducted a quick analysis that included using the revised PCRA (including violence assessment) to categorize federal offenders into the five groups referenced in Tables 1 and 2. Extrapolating from data on the number of contacts officers make, we assumed for a caseload of 60 persons under supervision that an officer makes 119 contacts (including contacts at the home, the place of employment, by telephone, etc.). We then made some guesses about the number of contacts that may be appropriate in order to address the monitoring, restrictions, and interventions appropriate for a person that falls into each category. (See Figure 4.)

Using a similar process we also investigated treatment expenditures for a hypothetical but typical caseload. This process focused on determining if districts could redirect funds in an intentional way to ensure that the needs of higher-risk offenders are being addressed. We wanted to know how much treatment money would be available for the orange and red categories of offender if we shifted 90 percent of funds spent on treatment for the white category of offenders, 75 percent of the funds spent on the green category of offenders, and 25 percent of the funds spent on the yellow category of offenders. The answer to that question is contained in Figure 5 and indicates that taking 90 percent, 75 percent, and 25 percent of treatment dollars spent on lower-risk cases (White, Green, and Yellow Categories respectively) allows our system to increase treatment dollar expenditures on higher-risk cases (Orange and Red Categories) by 2.4 and 3.7 times respectively. (See Figure 5.)

Figures 4 and 5 demonstrate that with current resources, a typical officer could stretch the current risk differentiation even further, and almost double the number of contacts made related to the higher-risk cases and double or triple the amount of treatment dollars spent on the higher-risk cases. This is possible only because of the large percentage of federal cases that fall in the "green" category. It should be noted that we in no way believe that frequency of contact or shifting treatment dollars on paper is adequate to provide complete and practical application of this matrix. We also must state that in no way do we believe the number of contacts alone will increase overall effectiveness. The quality of the contact, the purpose of the contact, the skill level of the officer, all play a role in the success of supervision. Likewise the quality of treatment, the purpose of the treatment, and the skill level of facilitators in correctional treatment programs also play a significant role in determining the effectiveness of correctional efforts. Finally, we acknowledge that each district's caseload composition, contact averages, available treatment dollars, and expenditures by risk might differ from the averages we present here. Nonetheless, what we have presented above is an exercise that communicates the concepts the advisory group settled on and will hopefully lead to many additional thoughtful

FIGURE 4.
Existing Monthly and Reallocated Monthly Contact Rates by Offender Risk Category

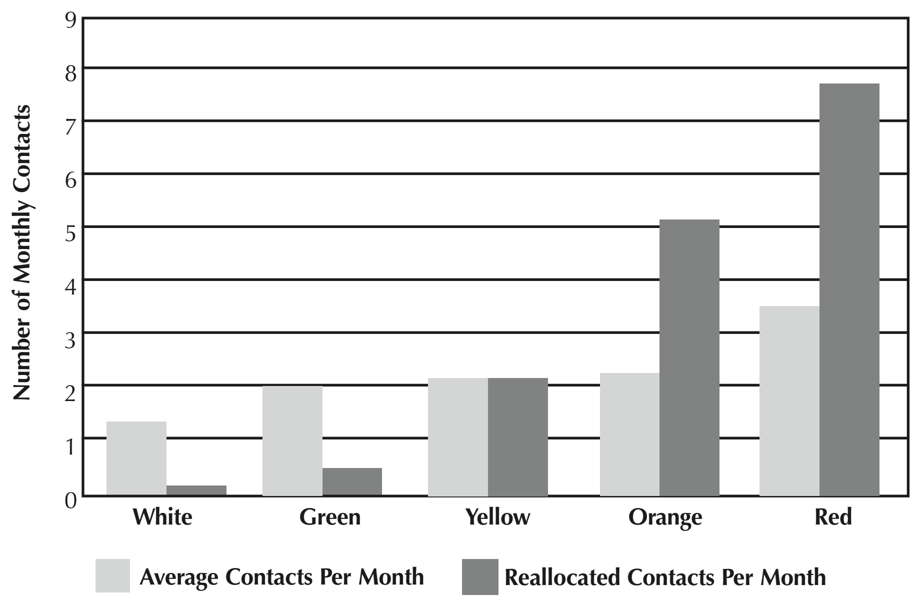


FIGURE 5.
Existing Average and Reallocated Treatment Dollars by Offender Risk Category

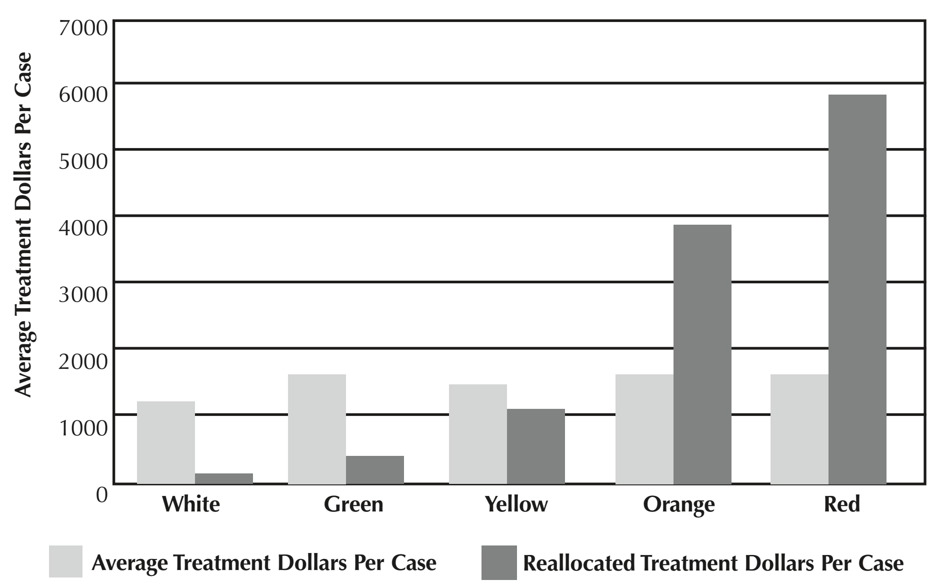


FIGURE 6.
Average Days Between Contacts and Percent Change in the District of South Dakota from Fiscal Year 2009-2016

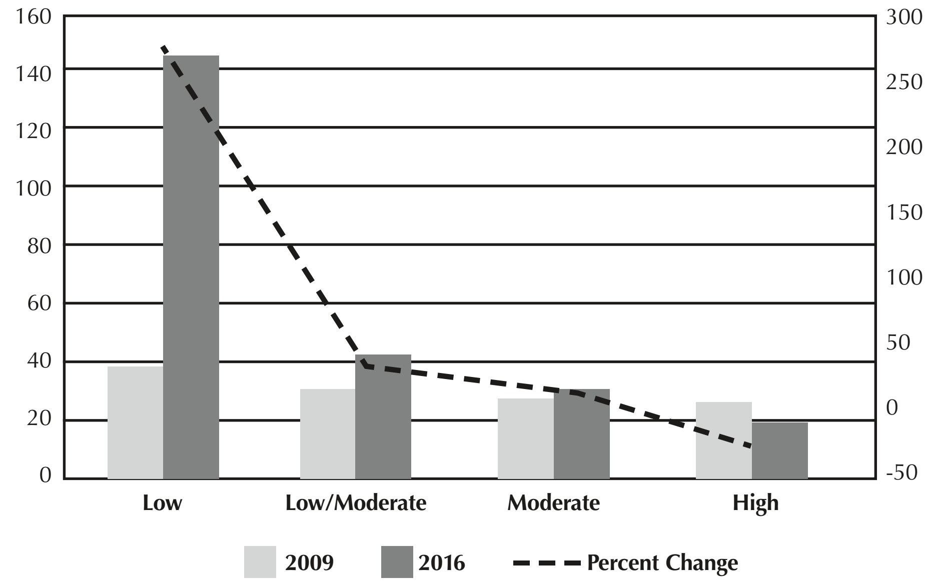
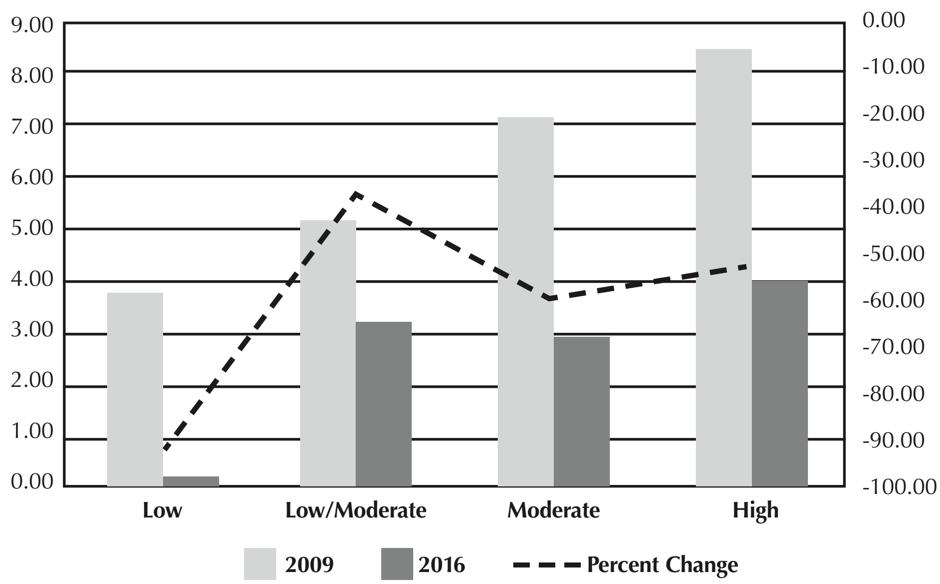


FIGURE 7.
Average Daily Treatment Costs and Percent Change in the District of South Dakota from Fiscal Year 2009-2016



conversations, in districts and on the national level, about how the federal probation system might move in this direction.

An example of the application of the advisory group's supervision process can be seen in the District of South Dakota. The experience of the District of South Dakota is presented in brief below to give staff in the field a more concrete and practical application of what has been discussed to this point.

Case Example

The United States Probation and Pretrial Services Office in the District of South Dakota (hereafter Office) has made efforts to improve its service to the public by engaging in evidence-based decisions and by aligning its resources with empirical evidence on effective practices. At a macro level, we have initiated cost-effective risk reduction and risk management strategies and practices to realize a compelling vision of enhanced community safety and greater achievement of justice. At a micro level we have engaged in a day-to-day awareness of and focus on making the best decisions. Figure 6 displays the average number of days between contacts by risk level for fiscal years 2009 and 2016, while Figure 7 displays the average daily cost of treatment by risk for fiscal years 2009 and 2016.

As Figure 6 makes clear, as we increased the number of days between contacts for low-risk offenders (by over 250 percent) from fiscal year 2009 to fiscal year 2016, we reduced the time between contacts for high-risk offenders by roughly 25 percent. Figure 7 indicates that while average treatment costs for all risk categories have decreased from fiscal year 2009 to fiscal year 2016, the greatest reductions were seen among the low-risk offenders (92 percent reduction). Smaller but meaningful reductions were also seen in the other risk categories; however, note that the average daily cost of treatment is highest among the high-risk offenders.

In summary, the data available from the District of South Dakota indicate that the risk principle is coming into focus. There is certainly some more work to be done; however, clearly there is an evident and growing differentiation in the daily cost of treatment services between low- and high-risk offenders. Further, by increasing the length of time between visits with low-risk offenders, the district has been able to increase the focus on the higher-risk offenders that cause the system and the public the greatest concern.

Conclusion

Years of effort and hard work by leaders in the federal probation and pretrial services system throughout the country have resulted in the risk principle solidly taking hold. The indicators we have, though only indicators, certainly point to a shift in the attention of probation officers to those at highest risk of recidivism. This is great news. As noted above, federal policy has long promoted individualized supervision that calls for additional resources on the higher-risk cases. Therefore it is no surprise that even the 2010 pre-RNR implementation numbers reflect a stair-step approach by officers in terms of their number of contacts with various risk categories of people. Given those numbers, the difference in these past five years is remarkable. This more extended differential between treatment of high- to low-risk persons is mirrored by the funds allocated for treatment needs. While we recognize these measures cannot capture the quality of supervision, and are therefore merely proxies for good supervision attention, they are the best indicators currently available, and demonstrate a very encouraging trend. Also supporting the risk principle is the analysis of those low-risk persons who, for whatever reason, receive a higher level of attention than their risk level would require: The additional attention is not accompanied by improved outcomes for the low-risk persons.

Although this article focuses only on risk, we realize that the risk principle's optimum value is realized only when it is embraced as part of the full risk/needs/responsivity model. We will continue to analyze the risk principle in action once the revised violence assessment is in full use. The guidance that will be shared widely with probation officers will provide a fuller view of the person's risk, and will lead to a more fine-tuned action plan for supervision. The case study from South Dakota reinforces the notion that this shift is possible without the need for additional resources. South Dakota's federal supervisee population is higher risk and more violent than most in the federal system. While all districts nationwide work toward embracing the RNR model, we hope to continue to learn from one another and to be encouraged to move forward. While we are asking federal probation offices to make these changes without additional funding, we will continue to measure and analyze the costs of success in this important endeavor. If we can demonstrate the costs of achieving the goal of fewer victims and fewer crimes in a system as diverse and large as our federal system, we will surely have advanced the conversation in an important way. We expect that the delineation of risk and accompanying suggested levels of monitoring/restrictions/interventions will lead to more consistent, targeted supervision efforts, and when addressed as a part of the federal Risk/Needs/Responsivity model, will lead ultimately to fewer victims and fewer crimes.

References

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

Andrews, D. A., & Bonta, J. (2007). The Risk-Need-Responsivity Model for offender assessment and rehabilitation. *Corrections User Report*. Ottawa, Canada: Public Safety Canada.

Andrews, D. A., & Dowden, C. (2007). The Risk-Need-Responsivity Model of assessment and human service in prevention and corrections: Crime-prevention jurisprudence. *Canadian Journal of Criminology and Criminal Justice*, 49(4): 439-464.

Center for Effective Public Policy. (2014). *Dosage probation: Rethinking the structure of probation sentences.*

Cohen, T. H., Cook, D., & Lowenkamp, C. T. (2016). The supervision of low-risk offenders: How the low-risk policy has changed federal supervision practices without compromising community safety. *Federal Probation, 80*(1)3-11.

Cohen, T. H., Lowenkamp, C. T., & Vanbenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology & Public Policy, 15*(2): 263-296.

Lovins, L. (2012). An empirical examination of variation in effective correctional program characteristics by gender. Cincinnati, Ohio: Doctoral Dissertation.

Lowenkamp, C. T. (2004). Correctional program integrity and treatment effectiveness: A multi-site, program-level analysis. Cincinnati, Ohio: Doctoral Dissertation.

Lowenkamp, C. T., Holsinger, A. M., & Bechtel, K. (2016). Evaluation of the Dosage Probation Project. Hudson, Ohio: Wisconsin DOC, unpublished manuscript.

Lowenkamp, C. T., Pealer, J., Latessa, E. J., & Smith, P. (2006). Adhering to the risk principle: Does it matter for supervision-based programs? *Federal Probation, 70*(3): 3-8.

Lowenkamp, C. T., Johnson, J. L., Trevino, P., & Serin, R. C. (2016). Enhancing community supervision through the application of dynamic risk assessment. *Federal Probation, 80(2)*.

Serin, R. C., Lowenkamp, C. T., Johnson, J. ., & Trevino, P. (2016). Using a multi-level risk assessment to inform case planning and risk management: Implications for officers. *Federal Probation, 80(2)*.

Vance, S. (2011). *An overview of the Post Conviction Risk Assessment*. Administrative Office of the U.S. Courts. http://www.uscourts.gov/statistics-reports/publications/post-conviction-risk-assessment

Using a Multi-level Risk Assessment To Inform Case Planning and Risk Management: Implications for Officers

Ralph C. Serin[1]
Christopher T. Lowenkamp[2]
James L. Johnson[2]
Patricia Trevino[2]

ONE OF THE primary goals of the federal probation and pretrial services system is to protect the community through the use of controlling and correctional strategies designed to assess and manage risk. In 2010, the Administrative Office of the U.S. Courts (AO) developed the Post-Conviction Risk Assessment (PCRA) tool as a means to assess offender risk in an effort to reduce future criminal behavior. Arguably, the best chances for reducing future criminal behavior occur when officers not only have a reliable way of identifying high-risk offenders but also can intervene in the criminogenic needs of those offenders (Andrews et al., 1990; Lowenkamp & Latessa, 2004; Bonta & Andrews, 2007; Campbell, French, & Gendreau, 2007; Johnson et al., 2011).

Clients with higher PCRA scores have poorer probation outcomes—compelling evidence of PCRA's predictive accuracy (Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011; Lowenkamp, Johnson, Holsinger, VanBenschoten, & Robinson, 2013). Half of the 18 PCRA points reflect criminal history factors, while the other half reflect viable case planning targets indicative of criminogenic needs (Bonta & Andrews, 2016). Moreover, clients with similar PCRA scores can have different point elevations across the subscales (i.e., education/employment, substance abuse, social networks, and cognitions) that identify different case planning needs for different clients. Furthermore, PCRA score changes over time are related to client outcomes; increases in PCRA scores lead to increased client failure, while decreases in PCRA scores lead to lower rates of recidivism (Cohen, Lowenkamp, & VanBenschoten, 2016; Luallen, Radakrishnan, & Rhodes, 2016). Because the PCRA has the ability to predict client outcomes for both baseline and change scores, probation officers are better equipped to identify intervention strategies for individual clients. Nonetheless, while the PCRA predicts client rearrests as well as informing case planning and risk management, this process is not completely intuitive for some officers. Therefore, the purpose of this paper is to make the process more explicit, especially regarding violent rearrest.

Revisions to the PCRA have led to the creation of PCRA 2.0, which reflects improved client normative data, clarifications of scoring rules, removal of some unscored test questions that did not substantially enhance predictive power, inclusion of static risk factor questions, and Psychological Inventory of Criminal Thinking Styles (PICTS) scales predictive of violent arrest. Despite evidence that probation officers in some jurisdictions ignore or override statistical risk assessments (Miller & Maloney, 2013), the importance of the PCRA is embedded within federal probation policy. Future training is intended to assist officers in recognizing the predictive validity PCRA 2.0 provides, while also highlighting the limitations of unstructured assessments (i.e., ignoring or overriding PCRA risk categories based on professional judgment or intuition). The expectation is that officers will incorporate PCRA 2.0 assessments into their correctional practices, thereby improving decisional accuracy, case planning, and risk management.

Increased scrutiny of sentinel events (e.g., sensational community failure—see Sheil, Doyle, & Lowenkamp, 2016, in this issue of *Federal Probation*) sparked interest within federal probation in including within the PCRA a violence risk assessment and interventions. Central to a consideration of sentinel events is the inclusion of acute dynamic risk factors that could signify the potential imminence of an event within a higher-risk group. Before including the violence assessment in PCRA, only one item was violence-specific, raising the question of whether the utility of the PCRA could be augmented through the rating of violence flags as a second level of risk assessment. The inclusion of validated violence flags is intended not only to insulate officers and the agency from undue criticism in the wake of an offender committing a serious violent offense, but also to reduce risk of harm to the community and further enhance officer safety. This risk assessment process, commonly known as due diligence in the field of risk assessment, must be credible and employ a best practice approach. The key consideration is a defensible decision process, and not merely an accurately predicted outcome.

Various sources provide important information regarding possible violence flags. First was the review of violence risk appraisal instruments (e.g., LS/CMI, Andrews, Bonta,

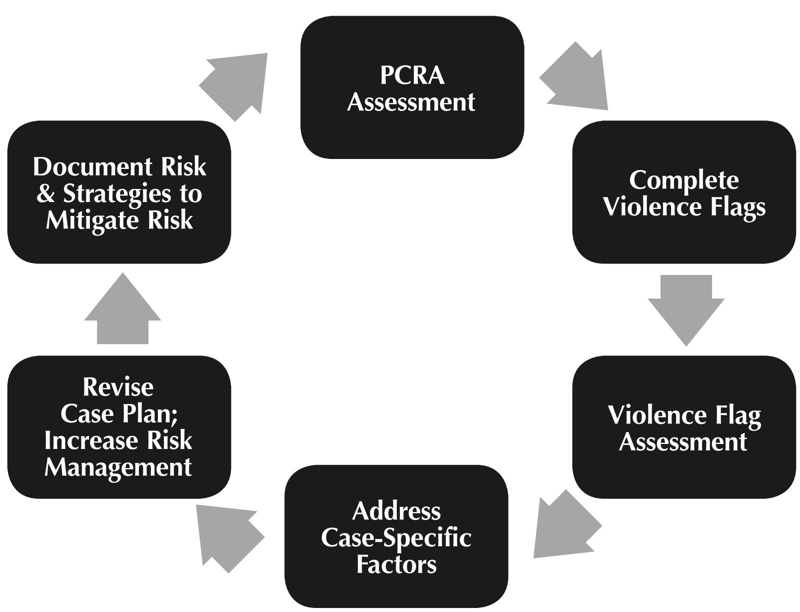
[1] Carleton University.
[2] Administrative Office of the U.S. Courts.

& Wormith, 2004; HCR-20V,[3] Douglas, Hart, Webster, & Belfrage, 2013; PCL-R, Hare, 2003; and ODARA, Hilton, Harris, Rice, Lang, Cormier, & Lines, 2004). Next came a consideration of meta-analyses and meta-reviews (Desmarais, Singh, & Johnson, in press; Yang, Wong, & Coid, 2010; Singh, 2013). Key critiques also led to potential variables for inclusion as violence flags (Douglas & Skeem, 2005; Harris, Rice, Quinsey, & Cormier, 2015; Mills, Kroner, & Morgan, 2011). Candidate variables for violence flags were shared with experienced researchers and clinical colleagues in the United States, Canada, and New Zealand for feedback. After receiving feedback, a final list of 18 factors was compiled for empirical validation. Figure 1 presents a depiction of the multi-level risk assessment model.

In composing potential violence flags, it was important to restrict the flags to factors readily available in existing case file information while avoiding duplication of factors already included in the PCRA. This, however, meant that some promising factors (e.g., diagnoses, degree of planning, hostile ideation or schema) might be excluded. It also meant, at least in the early stages of development, that the violent risk factors would be primarily static and not include acute dynamic risk factors. An important revelation in this research was the recognition that specific types of violence warrant unique predictors. For instance, meta-analytic studies suggest that predictors for non-sexual violence (e.g., hostile attitudes), intimate partner violence (e.g., violation of non-contact orders), and sexual violence (e.g., deviant sexual preference) are distinct. Although the client outcome in validating the multi-level model includes sexual crimes, given the low prevalence and base rates for hands-on sexual violence amongst federal probationers, these unique predictors were excluded. The violent rearrest behaviors of primary interest in this research were intimate partner violence, robbery, aggravated assault, and homicide/manslaughter.

Most of the criminal violence measured in this study is considered to be goal-directed or instrumental. Instrumental violence refers to violence that takes place for a clearly identifiable purpose other than as a response to provocation or frustration. Such violence typically takes place within the context of a robbery or burglary (Douglass, 2010). This means that interventions should primarily focus on criminal thinking, justifications for the use of violence, and problem solving. In cases where anger may be an issue, interventions may also include coping with anger and arousal, identification of triggers, and conflict management. This will be reviewed more fully in the discussion section.

FIGURE 1.
Multi-level Risk Assessment Model



Methods

Sample

Data were drawn from a sample of 69,311 offenders who started federal supervision at least two years prior to December 1, 2014 (the date of the record check), who had a PCRA administered within 6 months of the start of supervision, and for whom a total PCRA score was present. A sample of 25 cases from each of the 93 districts was identified, yielding a sample of 2,325 cases that were sent to the districts for data collection. A total of 1,885 records were returned, of which 1,642 provided usable or complete data. The 1,642 cases represent 48,025 male and female offenders of varied ethnicities from urban and rural locations (see Table 1).

TABLE 1.
Description of Sample

	Unweighted Sample	
	N	%
Male	1,871	80
Hispanic		
Yes	315	14
No	1,972	84
Unknown	38	2
Race		
Asian	58	3
Black	825	36
Native American/ Eskimo	95	4
White	1,313	57
Other	4	<1
Unknown	10	<1
Age	2,325	39.68 (11.98)
PCRA	2,325	6.8 (3.69)

Using presentence reports and other casework documents available in federal probation electronic records, probation officers coded violence flags for the sample of cases. As such, this was an archival study in which a coding manual with decision rules was provided to each of the districts and coders. A primary contact was assigned to address any questions regarding the coding of the violence flags.

Results

Overview of Analyses

The analysis for this study was conducted in four stages. During the first stage, potential violence flags were identified using statistical techniques. In the second stage, violence flags were validated using construction and validation samples. The third stage consisted of summing the identified violence flags to produce a violence flag score. In the fourth and final stage, violence flags and PCRA results were combined to develop a series of risk categories or bins.

Validation of Violence Flags

The final sample of 1,642 cases had fewer than 4 items missing and there was no replacement of missing items with the overall mean score during statistical analyses. A weighted sample was used in subsequent analysis. The analytic strategy involved a 50 percent random split of the sample into construction and validation samples.

The weighted sample was used to identify the strongest 10 predictors of violent arrest from the candidate violence flags. The list of 10 violence flags is presented in Table 2. In addition to the 10 violence flags, associations for the total PCRA score and the top 4 PICTS scales (in italics) are also presented in Table 2.

TABLE 2.
Association between candidate variables and violent rearrest with weighted sample

Variable	Chi-square	p-value
PCRA Category	1588.257	0.000000
Prior Violent Arrests	701.608	0.000000
Current Violent Offense	634.382	0.000000
Plans Violence	530.582	0.000000
Age at First Arrest	503.395	0.000000
PICTS - Power Orientation	*431.720*	*0.000000*
Prior Stalking	422.484	0.000000
History of Treatment Noncompliance	349.015	0.000000
Gang Member	290.739	0.000000
Ever Use a Weapon	231.363	0.000000
PICTS - Entitlement	*220.138*	*0.000000*
Current DV	187.809	0.000000
PICTS - Denial of Harm	*178.703*	*0.000000*
Prior DV Arrests	160.715	0.000000
PICTS - Self Assertion/ Deception	*150.085*	*0.000000*
Ever Victimize Stranger	68.620	0.000000

Each of the 10 factors that were present was given a value of one. The flags were then summed to produce a count of the flags present. The distribution of the flag count across the weighted sample and the failure rate associated with each score on the violence flag count is presented in Table 3.

The next strategy was to assign cases to one of three risk groups based on flag scores.

TABLE 3.
Distribution of Marker Counts for Weighted Data

Marker Count	N	%	Cum %	Failure Rate
0	12,192	25	25	0.2
1	6,538	14	39	2.7
2	6,040	13	52	6.7
3	6,138	13	64	2.1
4	4,646	10	74	12.8
5	3,847	8	82	11.1
6	2,602	5	87	7.3
7	2,485	5	93	25.6
8	1,640	3	96	18.5
9	1,154	2	98	17.2
10	501	1	100	12.4
11	231	0	100	42.4
12	11	0	100	0.0

A review of the data suggested that cutoffs of 0-3, 4-6, and greater than 7 would be appropriate. Table 4 presents outcome data for the three violence categories for comparison with the PCRA risk categories.

The results suggest that both the flags alone and the PCRA appear to usefully identify groups that are at a higher risk of committing an act of violence. Moreover, data suggest that the violence flags might function as a violence trailer to augment the PCRA, even though the original purpose of the multi-level model was to determine if the violence flags could be integrated into PCRA to provide improved prediction. Predictive validity analyses are described below.

Predictive Accuracy of Multi-Level Model

The AUC results for weighted samples are presented in Tables 5 (construction sample) and Table 6 (validation sample) for multiple violence outcomes. A review of these tables suggests acceptable predictive accuracy for both construction and validation samples and for all three client outcomes. In each situation, the inclusion of violence flags increases the predictive accuracy above that of the PCRA alone.

These findings suggest the multi-level risk assessment model has merit above and beyond either the PCRA or the violence flags alone. The increased breadth of predictors increases face validity with respect to violence risk at no decrease in predictive accuracy. In fact, accuracy is slightly increased across all comparisons. Subsequent analyses (not presented here) also indicate that the use of the violence flags in conjunction with the PCRA allows for greater accuracy in identifying offenders at increased risk of violence.

TABLE 4.
Distribution, Failure Rates, and Percentage of Violent Arrests Identified by Violence Risk Categories for Weighted Samples

Violence Risk Category	N	%	Cum %	Failure Rate	% Identified
0	25,131	52	52	1.15	9.0
1	15,186	32	84	8.18	38.4
2	7,708	16	100	22.07	52.6
Violence Risk Category	**N**	**%**	**Cum %**	**Failure Rate**	**% Identified**
Low	18,423	38	38	0.49	2.8
Low/Moderate	18,131	38	76	7.34	41.1
Moderate	8,509	18	94	11.76	31.0
High	2,962	6	100	27.38	25.0

Discussion

This research regarding the development and validation of a multi-level violence risk assessment model was initiated as a proof of concept. The goals of the research included: (1) examination of the predictive validity of the PCRA regarding violent rearrest; (2) inclusion of credible risk flags to augment validity; (3) incorporation of a due diligence approach to risk assessment in order to mitigate criticism in the event of offender failure; and (4) informing offender level case planning and risk management. Based on the findings presented, the first three goals were fully met.

In terms of case planning and risk management, the model also provides some general guidelines. The model is a sorting strategy whereby offenders with higher scores (PCRA and violence flags) are at a significantly greater risk of violent re-offending. Hence, when offenders score higher on the model, officers should be more aware of the increased likelihood of offenders engaging in violent behavior, so they can implement supervision strategies to mitigate risk and document efforts taken to manage risk. Based on the violence flags, some suggestions are presented in Table 7 for officers managing offenders with violence flags. When endorsed, the violence flags imply differential strategies to be undertaken by officers, based on overall risk level and *type* of violent offender. This approach recognizes there is heterogeneity among violent offenders, with differences in factors such as risk level, motivation for violence (goal-directed versus anger), motivation for treatment, weapon use, victim preference (stranger versus acquaintance), and degree of planning.

Case Planning and Management

The original PCRA predicted general recidivism based on scored factors related to criminal history, social networks, education/employment, drug and alcohol use, and cognitions. Overrides occurred for individuals with persistently violent histories because the PCRA did not properly assess violence. In recognition of this limitation, PCRA 2.0 was created, which incorporates a violence risk assessment. PCRA 2.0 allows for better accuracy in identifying individuals at an elevated risk for committing a violent act based on static risk factors and current PICTS scales. Use of PCRA 2.0 should result in better decision making in the case planning and risk management process, mitigate risk of harm to the community, and enhance officer safety.

Persons on community supervision for

TABLE 5.
AUCs for Prediction of Violent Rearrest With Construction Sample (*n*=1,154)

	DV	Violent No DV	Violent & DV
PCRA	0.76	0.78	0.78
Flags	0.72	0.73	0.73
Both	0.77	0.79	0.79

TABLE 6.
AUCs for Prediction of Violent Rearrest With Validation Sample (*n*=1,154)

	DV	Violent No DV	Violent & DV
PCRA	0.67	0.81	0.80
Flags	0.69	0.78	0.78
Both	0.69	0.83	0.82

TABLE 7.
Violence Flags and Differentiated Interventions

Violence Flag	Differentiated Intervention
PCRA Score	Higher scores require greater monitoring, restrictions, and interventions to mitigate risk.
Prior Violent Arrests	More likely to re-commit violent crime; target justifications for using violence to meet ends.
Current Violent Offense	More likely to re-commit violent crime; target justifications for using violence to meet ends or poor self-control (anger, impulsivity, poor problem solving).
Plans Violence	Violence is proactive not spontaneous. Target criminal thinking rather than anger.
Age at First Arrest	Earlier onset suggests longer criminal careers, requiring demonstration of change, not just verbal statements.
PICTS - Power Orientation	Violence is a choice to meet an end with rationalizations common and acceptance of responsibility lower. *Will likely reject treatment.*
Prior Stalking	More likely to re-commit domestic violence. Except in rare cases, most often knows victim.
History of Treatment Noncompliance	Use Core Correctional Practices, motivational engagement, and behavioral contracts linked to supervision requirements to increase treatment compliance.
Gang Member	Violence will be both predatory and anger-based (depending on rank in the gang). Requires monitoring of peers and victim access.
Ever Use a Weapon	If weapons taken to crime scene, risk is elevated. Violence more likely instrumental. If weapons selected by convenience, violence more likely impulsive.
PICTS - Entitlement	High levels indicative of justification for using violence, regardless of level of victim injury. *Will likely reject treatment.*
Current Domestic Violence	Violence is most likely instrumental (goal-directed to meet ends). Victim access a critical consideration.
PICTS – Denial of Harm	Rejects responsibility, justifies violence, will likely reject treatment.
PICTS – Self Assertion	Asserts will over others to achieve goals. Violence is rationalized.

a crime of violence present an elevated risk of harm to the community and may pose a greater danger to probation officers than individuals with non-violent offenses. Risk is increased even more if the individual has a recurrent pattern of violent behavior or affiliations with a gang (Battin-Pearson et al., 1998; Decker, 2000). The higher the risk an individual presents, the more intense the monitoring practices. Monitoring techniques may include, but are not limited to, increased field contacts, collateral contacts, drug testing, computer monitoring, and third-party risk assessment. Policy and procedures requiring more supervision contacts with higher-risk individuals may also be implemented. Frequency alone is not enough to deter future crime; therefore each contact must be purpose-driven and viewed as an opportunity to mitigate risk. In order to make contacts more purposeful, officers should routinely review the individual factors that led to the individual becoming high risk.

Observing current behaviors of offenders under supervision is a critical component of community corrections. However, officers should also review and investigate the circumstances surrounding prior violent offenses and consistently perform risk assessments as a part of their due diligence. A violence flag such as a history of planning violent behavior is indicative of proactive criminal thinking and may provide insight into how a person uses violence as a means to resolve conflict and control others. The prior use of weapons to commit a crime is a major public and officer safety concern, as access to firearms is empirically linked to lethal outcomes. The types of prior violent offenses and types of victims should also be carefully analyzed to properly address third-party risks. Persons under supervision for domestic violence, stalking, or threatening their victim(s) are more likely to go after the same victim. Access to victims should constantly be addressed, as it increases the likelihood of re-offense. No contact conditions and restrictions such as location monitoring and home confinement could be added by the court to address risk. When thoroughly analyzed, the totality of circumstances can aid officers in case planning and risk management. It will also contribute to increased public and officer safety.

The PICTS scales of power orientation, entitlement, denial of harm, and self-deception are used as violent flags in the multi-level assessment process. The presence of these factors merit careful consideration in case planning and risk management. Individuals with elevated scores of power orientation tend to be manipulative and intimidating and exert power over others. Individuals with the criminal thinking error of entitlement may believe they are above the law, assume ownership over others, and often systematically misidentify wants as needs. Interventions should target criminal thinking and include cognitive-based individual or group treatment, the use of core correctional practices, problem solving, impulse control, identification of triggers, assignment of homework, and enhanced coping skills.

Evidence suggests that the effectiveness of correctional interventions is enhanced when officers match proper monitoring strategies, restrictions, and interventions (Andrews & Bonta, 2010; Landenberger & Lipsey, 2005; Lowenkamp & Latessa, 2005; Lowenkamp, Latessa, & Holsinger, 2006). The multi-level assessment should make case planning more individualized, allow officers to better recognize offenders who are at a higher risk of rearrest for a violent offense, and assist in creating supervision objectives.

Conclusions

Risk recognition is the primary initial step required by officers in managing their caseload. The multi-level risk assessment model provides a new approach to assist officers to appreciate the likelihood of violent rearrest by clients. Higher scores warrant more focused and prescriptive intervention by officers. Moreover, specific elevated flags inform both case planning (intervention within sessions with the client and referrals to service providers) and risk management (frequency of contact, frequency of face-to-face meetings, behavioral contracts, assignment of homework, etc.). Finally, risk recognition increases the requirement for increased documentation, especially in terms of how the officers have addressed client risk level and how they have responded to incidents of noncompliance by the client.

References

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). User's manual for the Level of Service/Case Management Inventory (LS/CMI): An offender management system. Toronto, Canada: Multi-Health Systems.

Battin-Pearson, S. R., Thornberry, T. P., Hawkins, J. D., & Krohn, M. D. (1998). *Gang membership, delinquent peers, and delinquent behavior*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice.

Bonta, J., & Andrews, D. A. (2016). *Psychology of criminal conduct* (6th ed.). Anderson Publishing.

Cohen, T. H., Lowenkamp, C. T., & VanBenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology & Public Policy*, *15*(2), 263-296. DOI:10.1111/1745-9133.12190

Desmarais, S. L., Singh, J. P., & Johnson, K. L. (in press). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services.*

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20v3: Assessing risk for violence: User guide.* Mental Health, Law, and Policy Institute, Simon Fraser University, Burnaby, British Columbia.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014) Historical-Clinical-Risk Management-20, Version 3 (HCR-20V3): Development and overview. *International Journal of Forensic Mental Health*, *13*(2), 93-108. DOI: 10.1080/14999013.2014.906519

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347–383. doi:10.1037/1076-8971.11.3.347.

Douglas, R. L. (2010). *Instrumental and reactive violence: The role of mental health factors and maltreatment history in the manifestation of violent offending*. Dissertation Dalhousie University Halifax, Nova Scotia.

Hare, R. D. (1991; 2003). *The Hare Psychopathy Checklist – Revised*. Mental Health Systems, Toronto, Canada.

Harris, G. T., & Rice, M. E. (2015). Progress in violence risk assessment and communication: Hypothesis versus evidence. *Behavioral Sciences and the Law, 33,* 128-145.

Harris, G. T., Rice, M. E., Quinsey, V. L., & Cormier, C. A. (2015). *Violent offenders: Appraising and managing risk* (3rd ed.). ISBN: 978-1-4338-1901-8

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario Domestic Assault Risk Assessment. *Psychological Assessment, 16*, 267-275. doi.org/10.1037/10403590.16.3.267

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The construction and validation of the federal Post Conviction Risk Assessment (PCRA). *Federal Probation, 75*(2), 16-29. Washington, DC: Administrative Office of the U.S. Courts.

Luallen, J., Radakrishnan, S., and Rhodes, W.

(2016). The predictive validity of the Post-Conviction Risk Assessment among federal offenders. Criminal Justice and Behavior, doi: 10.1177/0093854816650481.

Mills, J. F., Kroner, D. G., & Morgan, R. D. (2011). *The clinician's guide to violence risk assessment.* Guilford: New York.

Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology*, 1: 451–476.

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services, 10*(1), 87-96. doi:10.1037/a0030343

Lowenkamp, C. T., & Latessa, E. J. (2005). Increasing the effectiveness of correctional programming through the risk principle: Identifying offenders for residential placement. *Criminology & Public Policy*, 4: 263–290.

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency*, 52: 77–93.

Miller, J., & Maloney, C. (2013). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice & Behavior, 40*(7), 716–736. doi:10.1177/0093854812468883

Singh, J. P. (2013). The International Risk Survey (IRiS) project: Perspectives on the practical application of violence risk assessment tools. Paper presented at the Annual Conference of the American Psychology-Law Society, Portland, OR.

Walters, G. (2015). Cognitive mediation of a crime continuity: A causal mediation analysis of the past crime-future relationship, *Crime & Delinquency, 61*(9), 1234-1256.

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*(5), 740–767. doi:10.1037/a0020473

Enhancing Community Supervision Through the Application of Dynamic Risk Assessment

Christopher T. Lowenkamp[1]
James L. Johnson[1]
Patricia Trevino[1]
Ralph C. Serin[2]

RISK FACTORS HAVE commonly been distinguished as being either static (e.g., age at first arrest, number of prior convictions) or dynamic (e.g., substance use, employment status). In the early days of risk assessment (1970s), static factors were most commonly incorporated into risk measures. They were easy to code and readily available; most importantly, these initial static risk measures demonstrated accuracy equal to or greater than unstructured assessments (Grove, Zald, Lebow, Snitz, & Nelson, 2000). Importantly, by the early 1980s, opposition to measures with exclusively static risk factors was beginning to develop, primarily because these scales could not identify intervention targets, and if scores could change, the range of potential change was greatly restricted and unidirectional (i.e., clients could only be rated worse; Bonta, 1996; Wong & Gordon, 2006). Notably, involvement in treatment could not improve scores, leading to the problematic practice of treatment completion having no impact on an individual's predicted outcome.

Andrews and Bonta (2010) presented a hierarchy of risk factors intended to identify appropriate targets for rehabilitation programs; their choice of variables was consistent with a conceptualization of dynamic risk factors as relatively slow-evolving features. Their description of these targets as *criminogenic needs* came to be considered synonymous with the concept of dynamic risk and led to the risk and need principles. Indeed these stable dynamic risks were increasingly common in risk and need measures; their inclusion was intended to inform both levels of risk and case planning requirements for clients. Clients with a greater number of stable dynamic risks (i.e., criminogenic needs) were considered higher risk, warranting more intensive intervention and level of service. Encouragingly, targeting these criminogenic needs leads to improved client outcomes (Aos, Miller, & Drake, 2006; Smith, Gendreau, & Swartz, 2009).

The PCRA is a contemporary risk and need instrument similar to other measures such as the LS/CMI, the COMPAS, and the ORAS. Validity research indicates the PCRA has comparable or superior predictive accuracy to these other instruments (Desmarais & Singh, 2013). Importantly, even though the PCRA assessment is done at baseline, at 6 months, and then yearly thereafter, change scores across time on the PCRA are related to client outcome (Cohen, Lowenkamp, & VanBenschoten, 2016; Luallen, Radakrishnan, & Rhodes, 2016). The odds of client failure can be predicted by changes from one PCRA assessment to the next. For instance, in a case where the client's PCRA score is 3 points *lower*, the probability of violent rearrest is *decreased* by 19 percent. In contrast, in a case where the client's PCRA score is 3 points *higher*, the probability of violent rearrest is *increased* by 31 percent. Clearly, change on criminogenic needs, as measure by the PCRA, is important in understanding client outcome.

Increasingly, experts in the risk assessment field have argued that accuracy regarding the timing of client outcome can be enhanced by considering changes in *acute* dynamic risk factors (Douglas & Skeem, 2005; Serin, Chadwick, & Lloyd, 2016). Specifically, the expectation is that acute risks flag imminence of problematic outcomes for clients and augment risk assessment beyond static factors. As well, elevations in acute risk should mean that clients with similar crimes and PCRA scores could be managed differently from clients without such acute risks. Several examples illustrate this viewpoint. You have a client for whom employment has been a concern in that when unemployed, the client commonly turns to criminal behavior to generate income. Hence, when that client advises you that he or she has just been fired, this should be a flag that increased monitoring (e.g., efforts to secure a new job, assistance with job search, access to and association with criminal peers, etc.) is in order. Similarly, if a client during a session reports (or you observe) increases in anger or negative emotions, this might indicate increased vulnerability to criminal thinking and criminal behavior. Such a change could warrant further scrutiny and intervention by officers.

Despite decades of risk assessment research, the field is limited in its understanding of the immediate features (whether situational or intrapersonal) that influence an individual to take criminal action (Farrington,

[1] Administrative Office of the U.S. Courts.
[2] Carleton University.

2011; Yang & Mulvey, 2012) or forgo criminal action when presented with an opportunity for crime (i.e., crime desistance; Maruna, 2010). The current research was undertaken to examine whether certain acute dynamic risks might better identify not only which clients are at risk but also *when* that risk might be most elevated for a particular client. In this manner, it is possible for officers to consider risk at the case level and intervene accordingly to mitigate it.

Fortunately, some recent research regarding acute dynamic risk is available (Serin, Chadwick, & Lloyd, 2015). Using the list of acute variables developed by Serin (2007) in the Dynamic Risk Assessment for Offender Reentry (DRAOR) measure, the present study examines if key acute risks forecast violent rearrest in a federal probation sample. The results may have implications for officer assessment and intervention strategies.

Methods

Sample

Data used for this study were assembled from federal supervision records from the Probation and Pretrial Services Office's internal case management database system (Probation and Pretrial Services Automated Case Tracking System or PACTS) and other extant data sources. The source dataset included 385,130 offenders serving either a term of probation or a term of supervised release (TSR) that commenced between October 1, 2004, and September 30, 2013. Excluded from the source dataset were offenders who were deported, serving a sentence in another jurisdiction, or otherwise unavailable for supervision.

A sample of 2,153 offenders who had been arrested for a violent offense (i.e., homicide, attempted homicide, sexual assault, robbery, and felonious assault) while under supervision was extrapolated from the source dataset. Another 1,963 cases were selected that were not arrested for a violent offense while on supervision but matched the sample of violent offenders based on supervision district, convicted offense, risk score, and year supervision began. This provided a sample of 4,116 cases.

Data collection for this study occurred over the course of two weeks in September 2014. Officers used available electronic data including presentence reports, federal Bureau of Prisons (BOP) data, and PACTS data to complete the data collection forms. Forty-seven officers ranging in experience from 5 years to 23 years collected data during the weeks of September 15-19 and September 22-26. One week prior to data collection, officers were given copies of the data collection form and the coding manual. A WebEx training was also conducted to provide an overview of the study and a detailed review of the data collection form and coding manual. The 47 officers assisted in the collection of data on 949 cases.

TABLE 1.
Distribution of Cases for Total Sample and Sample Collected in September 2014

	Total Sample		Sept, 2014 Sample	
	N	%	N	%
Homicide	696	17	258	27
Sexual assault	158	4	48	5
Robbery	533	13	151	16
Felonious assault	766	19	198	21
Comparison cases	1963	48	294	31
Total	**4116**	**100**	**949**	**100**

Experienced data quality analysts were used for quality assurance and data entry. The data quality analysts reviewed each completed data collection form for accuracy, then entered the data into a web-based version of the data collection form. The distribution of the cases for the entire sample and the cases where data were collected are listed in Table 1.

Measures

Offender data included prior criminal history, information related to imprisonment in the Federal Bureau of Prisons, current offense, needs while under supervision, and information on the violent offense committed while under supervision. The "needs while under supervision" information was collected using the Dynamic Risk Assessment for Offender Reentry (DRAOR) developed by Serin (2007) and the Two Tiered Risk Assessment (TTR) developed by Mills, Kroner, and Morgan (2011). However, the current study only uses the data on the DRAOR.

The DRAOR comprises 19 items divided into three subscales: stable factors, acute factors, and protective factors. This study used the seven acute factors: substance abuse, anger/hostility, opportunity/access to victims, negative mood, employment, interpersonal relationships, and living situation. Each item is rated using a three-point scoring format (0, 1, 2) that corresponds to anchors of "not a problem," "slight/possible problem," and "definite problem." When summed, the seven items create a score ranging from zero to 14, with higher scores indicating a greater number and/or degree of problems present for the assessment time period.

Data on acute factors were coded in 30-day increments for up to 18 months. If supervision spanned more than 18 months, then the first 6 months of supervision and the 12 months preceding the violent arrest or the end of supervision were coded. Data on violations of supervision conditions such as new arrests, job changes, travels outside jurisdiction without permission, treatment noncompliance, positive drug tests, and failure to report were also coded in 30-day increments. A total of 13,676 observational periods were coded for the 949 offenders. Due to the nature of the data collection, there were varying levels of missing data that were replaced with the most recent value recorded for a particular measure. The use of Cox Regression models produced a total of 597 cases with usable data, of which 392 cases were arrested for a violent offense while under supervision. There was a total of 7,538 observation periods associated with these 597 cases.

In addition to the DRAOR, a violence classifier was developed to capture an offender's risk for committing a violent offense. Offenders were considered at higher risk for violence if they had a PCRA score greater than eight or a PCRA score less than nine with two or more of the following factors present: gang affiliation, currently on supervision for a sex or violent offense, history of drug arrests, history of firearms arrests, or a history of arrests for violence. Finally, a dichotomous variable (early onset) was developed that had a value of zero if the offender's first arrest was at age 18 or greater and a value of one if the offender's first arrest was before the age of 18.

Analyses

Bivariate and multivariate statistics were estimated during the analysis phase of the study. Since there were different lengths of supervision, and since the violent arrest of interest in most instances stopped the collection of data, we opted to focus on survival analysis models. In addition to the DRAOR scales, the violence classifier and early onset variables were also used in the multivariate Cox Regression (survival analysis) models.

Results

The first Cox Regression model included the violence classifier, the early onset variable, and the DRAOR acute item score. The results of that model are contained in Table 2 and indicate that once the dynamic acute risk factors are taken into account, the effect of the violence classifier, a static measure, is reduced to non-significance. The measure of early onset continues to be a predictor of time to failure. The DRAOR Acute Score is a significant predictor of failure once the score reaches a value of four or greater. Note that the hazard ratios for the acute score tend to follow an upward trend indicating that, in general, as the score increases so too does the likelihood that failure occurs in the near term.

The DRAOR Acute Score was recoded into three categories (0-4, 5-10, and 11-14). These categories were then used to display the differences in survival rates based on the accumulation of acute risk factors. As indicated in Figure 1 (see last page of article), those with scores between zero and four demonstrate the highest survival rates. Those with scores between five and ten survive at a noticeably lower rate than those with lower scores. Finally, those with scores between 11 and 14 clearly have the lowest survival rates and the decrease in survival rates is, relatively, very steep.

In an effort to determine if any particular acute risk factor was a better predictor of arrest for violence than the others, a model using each of the acute risk factors as predictors, rather than the summed DRAOR Acute Score, was constructed and estimated. The results of those analyses are contained in Table 3 and indicate that three factors were significantly related to time to failure (arrest for a violent offense). Those three factors are anger/hostility, access to victims, and negative mood.

A figure displaying the survival curves for each value (0=not a problem;1=possible/slight problem; 2=definite problem) of each of the significant factors was created. These are displayed in Figures 2 through 4 (see last page of article). Figures 2 and 3 demonstrate that the survival rates drop as the ratings for anger/hostility and opportunity/victim access increase from no problem to slight/possible problem and also when an offender was ranked as having a definite problem. In Figure 4, which plots the survival curves for the different ratings of negative mood, the separation between slight/possible problem and definite problem is not as pronounced as in Figures 2 and 3. In addition, in Table 3 the hazard ratio for definite problem for negative mood is not statistically significant. It is, however, clear that as the rating for negative mood shifts from no problem to slight/possible problem, a statistically significant hazard ratio is generated.

Discussion

The findings are very encouraging and inform refinements to the risk assessment process. Despite being an archival study that may be limited due to the availability of information necessary to code acute risk, 3 of the 7 acute risks identify cases that have a greater likelihood of violent rearrest in a large sample of seriously violent clients. Problems and concerns relating to anger/hostility, victim access, and negative mood all had significant odds ratios. Specifically, the results indicate elevations on these acute risks increased the

TABLE 2.
Cox Regression Predicting Arrest Using Violence Classifier, Early Onset, and DRAOR Acute Score

Variable	Hazard Ratio	p value	95% CI Lower	95% CI Upper
Violence Classifier	1.13	0.42	0.84	1.53
Early Onset	1.45	0.00	1.14	1.84
Monthly Acute Factor				
1	1.24	0.60	0.56	2.74
2	1.67	0.15	0.84	3.33
3	2.07	0.02	1.12	3.83
4	2.62	0.00	1.48	4.66
5	6.31	0.00	3.78	10.52
6	5.33	0.00	3.27	8.68
7	5.06	0.00	2.84	9.00
8	12.16	0.00	6.83	21.63
9	10.94	0.00	6.65	17.99
10	6.88	0.00	3.93	12.03
11	9.88	0.00	5.67	17.23
12	11.71	0.00	6.79	20.18

TABLE 3.
Cox Regression Predicting Arrest for Violence Offense with Violence Classifier, Early Onset, and Each DRAOR Acute Factor

	Hazard Ratio	p value	95% CI Lower	95% CI Upper
Violence Classifier 2	1.17	0.31	0.86	1.60
Early Onset	1.29	0.04	1.01	1.63
Substance Abuse				
Slight/Possible Problem	0.84	0.24	0.62	1.12
Definite Problem	0.90	0.57	0.64	1.28
Anger/Hostility				
Slight/Possible Problem	**1.90**	**0.00**	**1.28**	**2.81**
Definite Problem	**3.08**	**0.00**	**1.81**	**5.26**
Victim Access				
Slight/Possible Problem	**1.60**	**0.01**	**1.13**	**2.26**
Definite Problem	**3.04**	**0.00**	**2.00**	**4.63**
Negative Mood				
Slight/Possible Problem	**1.41**	**0.05**	**0.99**	**2.00**
Definite Problem	**1.45**	**0.15**	**0.88**	**2.39**
Employment				
Slight/Possible Problem	1.07	0.66	0.78	1.48
Definite Problem	1.00	0.99	0.73	1.36
Interpersonal Relationships				
Slight/Possible Problem	1.03	0.84	0.74	1.44
Definite Problem	1.12	0.61	0.73	1.69
Living Situation				
Slight/Possible Problem	0.66	0.02	0.47	0.93
Definite Problem	1.17	0.45	0.77	1.78

likelihood of a violent rearrest by 26 percent, 25 percent, and 9 percent respectively. As well, overall, a higher acute risk score significantly increased the odds of violent rearrest.

Equally informative is what did *not* relate to risk of violent rearrest. Substance abuse, employment, interpersonal problems, and living situation failed to inform the likelihood of violent rearrest. Moreover, PCRA elevated score (e.g., violence classifier) did not increase the likelihood of violent rearrest.

In addition to the likelihood of violent rearrest, the current study addresses the timing of such rearrest across risk groups. The survival analyses reflect extremely steep slopes for clients with significant problems relating to acute risk and specifically for anger/hostility, victim access, and negative mood. This means that these clients fail significantly more often and more quickly. With heightened degrees of imminent risk, immediate and appropriate changes in supervision strategies can be made to address the risk to reoffend and potential risk of harm to the community.

Despite these promising findings, some caution is warranted. This was a retrospective study that relied on existing information reflected in client chronos. Replication in a prospective study is warranted. Acute risk factors can change very quickly and should be consistently addressed with higher-risk individuals in order to enhance decision making, provide adequate interventions, and improve client outcomes (Serin et al., 2016). As well, additional acute dynamic risk factors that were not included in this study may also inform the likelihood and timing of client violent rearrest. Work to expand the inventory of credible predictors should be encouraged. Finally, risk recognition through the inclusion of acute dynamic risk, while helpful for officers, is somewhat limiting without the provision of best practice approaches for officers to use when these clients and their acute risk are identified. Fortunately, this work has begun in the upcoming PCRA 2.0 training, in which officers are provided with more specific approaches to manage clients who are considered at higher risk for violence while on probation.

References

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). New Providence, NJ: LexisNexis Matthew Bender.

Aos, S., Miller, M., & Drake, E. (2006) *Evidence-based public policy options to reduce future prison construction, criminal justice costs and crime rates*. Olympia: Washington State Institute for Public Policy.

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.

Cohen, T. H., Lowenkamp, C. T., & VanBenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology & Public Policy*, *15*(2), 263-296. DOI:10.1111/1745-9133.12190

Desmarais, S. I., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. New York, NY: Council for State Governments Justice Center.

Farrington, D. P. (2011). The integrated cognitive antisocial potential (ICAP) theory. In D. P. Farrington (Ed.), *Integrated developmental and life-course theories of offending* (pp. 73-92). New Brunswick, NJ: Transaction.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19–30.

Luallen, J., Radakrishnan, S., and Rhodes, W. (2016). The Predictive validity of the Post-Conviction Risk Assessment among federal offenders. *Criminal Justice and Behavior*, DOI: 10.1177/0093854816650481.

Maruna, S. (2010). *Understanding desistance from crime*. A report for the Ministry of Justice and National Offender Management Services. Retrieved from: www.safeground.org.uk/wp-content/uploads/desistance-fact-sheet.pdf

Serin, R. C. (2007). The Dynamic Risk Assessment Scale for Offender Re-Entry (DRAOR). *Unpublished scale. Carleton University, Ottawa, Ontario*.

Serin, R. C., Chadwick, N., & Lloyd, C. D. (2016). Dynamic risk and protective factors. *Psychology, Crime & Law*, *22:1-2*, 151-170, DOI: 10.1080/1068316X.2015.1112013

Serin, R. C., Gobeil, R., Lloyd, C. D., Chadwick, N., Wardrop, K., & Hanby, L. (2016). Using dynamic risk to enhance conditional release decisions in prisoners to improve their outcomes. *Behavioral Sciences & the Law*, March 2016; *34*(2/3): 321-336.

Smith, P., Gendreau, P., & Swartz, K. (2009). Validating the principles of effective intervention: A systematic review of the contributions of meta-analysis in the field of corrections. *Victims & Offenders, 4(2),* 148-169. doi: 10.1080/15564880802612581

Wong, S. C. P., & Gordon, A. E. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology, Public Policy, and Law, 12*, 279-309. doi:10.1037/1076-8971.12.3.279

Yang, S., & Mulvey, E. P. (2012). Violence risk: Re-defining variables from the first-person perspective. *Aggression and Violent Behavior*, *17*, 198-207. doi:10.1016/j.avb.2-012.02.001

FIGURE 1.
Survival Curves by DRAOR Acute Score

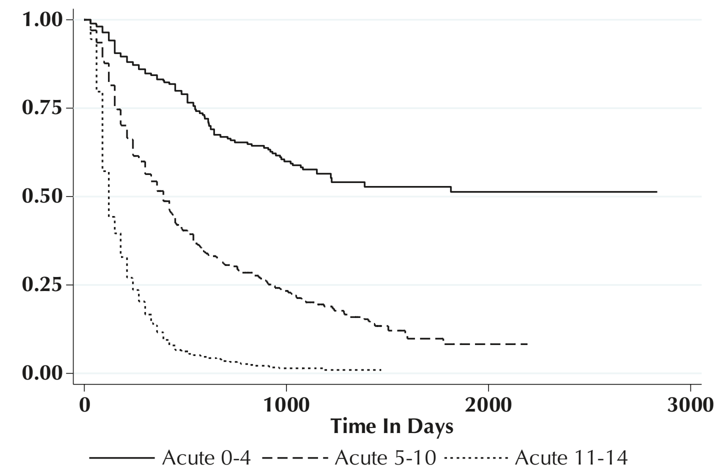


FIGURE 2.
Survival Curves by DRAOR Acute Anger/Hostility Rating

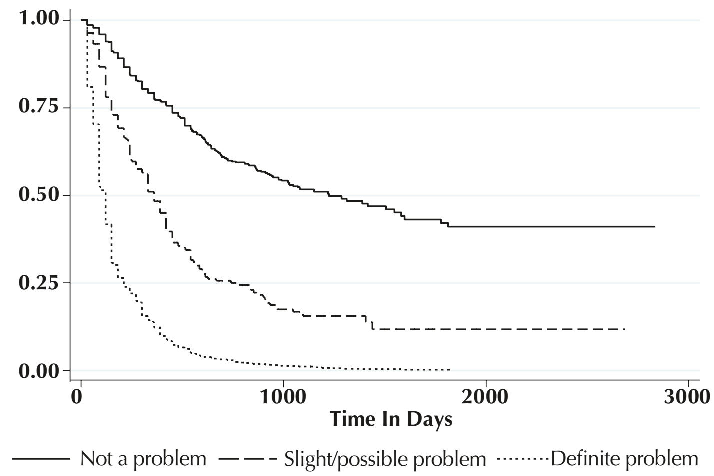


FIGURE 3.
Survival Curves by DRAOR Acute Opportunity/Victim Access Rating

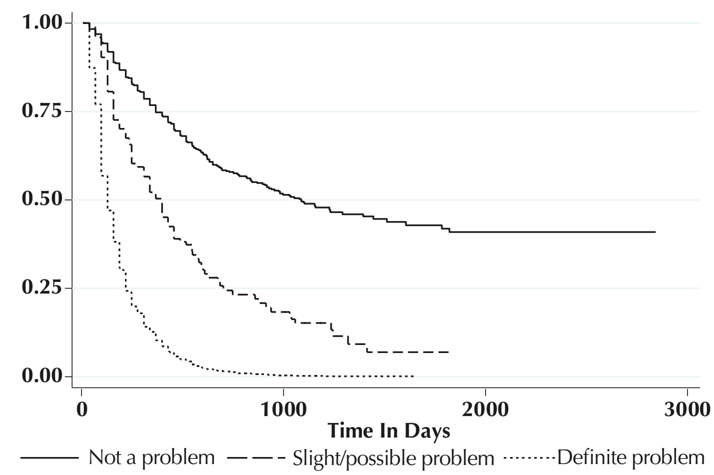
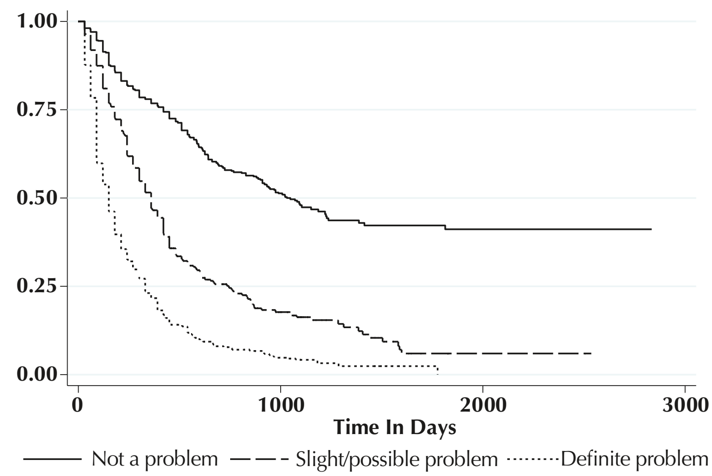


FIGURE 4.
Survival Curves by DRAOR Acute Negative Mood Rating



How Dangerous Are They? An Analysis of Sex Offenders Under Federal Post-Conviction Supervision

*Thomas H. Cohen**
Michelle C. Spidell
Probation and Pretrial Services Office
Administrative Office of the U.S. Courts

SEX OFFENSES ARE among the crimes that provoke serious public concern (Hanson & Morton-Bourgon, 2005, 2009). An especially acute concern involves the growing exploitation of children by online sex offenders who use the Internet and related digital technologies to possess, distribute, or produce child pornography or contact children for sexual purposes (Seto, Hanson, & Babchishin, 2011). Though accounting for a relatively small portion of all sex crimes against children, evidence shows substantial increases in the number of arrests involving online sexual offenses over the past ten years (Motivans & Kyckelhan, 2007; Wolak, Finkelhor, & Mitchell, 2005, 2009). Societal concern over the online sexual exploitation of children, along with evidence showing that many of these online offenders have self-reported histories of contact sexual offenses (Lam, Mitchell, & Seto, 2010; Seto et al., 2011), has produced aggressive law enforcement responses aimed at targeting sex offenders at the state and federal levels.

The federal response to the problem of sex offenders, and especially Internet child pornographers, manifests through both increased resources directed at law enforcement efforts and enhanced sentencing provisions (Faust & Motivans, 2015). Two primary federal legislative responses aimed at sex offenders are the Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today Act of 2003 (The PROTECT Act) and the Adam Walsh Child Protection and Safety Act of 2006 (The Adam Walsh Act). The PROTECT Act primarily increased mandatory minimum penalties for child pornography and sexual abuse offenders and provided federal judges with discretion to impose life supervision terms on federal sex offenders (Faust & Motivans, 2015; U.S. Sentencing Commission [USSC], 2012). The Adam Walsh Act gave the U.S. Attorney General authority to create a national registry of convicted sex offenders, authorized federal civil commitment for those certified as sexually dangerous, and permitted the imposition of search conditions for sex offenders sentenced to federal probation or supervised release (Faust & Motivans, 2015; USSC, 2012).[1] In addition to these legislative enactments, the U.S. Department of Justice has established numerous regional task forces and funded specialized units within federal law enforcement agencies to investigate and prosecute offenders engaging in Internet child sex crimes (Wolak et al., 2005).

As a result of these changes, the number of sex offenders prosecuted, incarcerated, and placed under federal post-conviction supervision has risen exponentially since the mid-1990s. (Faust & Motivans, 2015; USSC, 2012). In an examination of major trends, Faust and Motivans (2015) reported a nearly 1,400 percent increase in the number of sex offenders on post-conviction federal supervision, from 321 offenders in 1994 to 4,714 offenders in 2013, and much of this increase can be attributed to the prosecution of offenders charged with possession, receipt, distribution, or production of child pornography.[2] In addition, federal sex offenders are increasingly being sentenced to lengthy post-conviction supervision terms; for example, the United States Sentencing Commission (USSC) reported that in fiscal year 2010, the average terms of supervised release sentences imposed ranged from 220 months for offenders convicted of child pornography possession to 323 months for offenders convicted of child pornography production (USSC, 2012). By contrast, the average term of supervised release imposed on federal offenders generally in 2010 was about 43 months (USSC, 2012).

As the number of sex offenders, particularly online child pornographers, under federal post-conviction supervision has

* Thomas H. Cohen, Social Science Analyst, and Michelle Spidell, Probation Administrator, Probation and Pretrial Services Office, Administrative Office of the United States Courts, Washington, DC. The authors would like to thank Laura Baber, Trent Cornish, Christopher Lowenkamp, Stacy Merolla, and Matthew Rowland for their helpful suggestions and comments. This publication benefited from the careful editing of Ellen Fielding. Direct correspondence to Thomas H. Cohen, Administrative Office of the U.S. Courts, One Columbus Circle NE, Washington, DC 20544 (email: Thomas_cohen@ao.uscourts.gov).

[1] For additional details about the Walsh and PROTECT Acts, see 42 USC §16911 and 18 USC §2252.

[2] Faust and Motivans also noted substantial increases from the early 1990s in the number of offenders convicted of sexual abuse, sex trafficking, and violation of the Adam Walsh Act sexual registry mandates; however, by 2013, child pornography offenders accounted for the largest portion of sex offenders incarcerated within federal prisons.

increased, so too have concerns regarding whether these offenders have histories of, or are likely to engage in, offline contact sexual behavior with children. A recent meta-study of child pornography offenders conducted by Seto et al. (2011) found that about 12 percent of child pornography offenders had an official arrest or conviction record of contact sexual behavior, but 55 percent disclosed through self-reporting conducted through treatment programs, background investigations, or polygraphs that they had prior sexual contact with children.[3] A study of federal child pornography offenders conducted by the U.S. Sentencing Commission showed about 33 percent of these offenders engaging in some prior form of criminally sexually dangerous behavior (USSC, 2012).

These studies indicating that many online child pornography offenders are involved in contact offending, coupled with substantial caseload growth, raise important questions about the overall characteristics of federal sex offenders. Key questions have only begun to be explored, including what are the most common offense types (e.g., distribution of online child pornography, sexual assault) under post-conviction supervision, how many have an official arrest or conviction record of offline contact sexual behavior, what are their general recidivism risk characteristics, and how frequently do these offenders reoffend or get revoked (Bourke & Hernandez, 2009; DeLisi et al., 2016; Faust, Bickart, Renaud, & Camp, 2014; Faust & Motivans, 2015; USSC, 2012). Moreover, there have been no empirical assessments of the extent to which the current actuarial instrument used by federal probation officers to predict general recidivism—the Post-Conviction Risk Assessment (PCRA)—can predict general recidivism or revocations for federal sex offenders.

In the sections below, we discuss the federal judiciary's policy for supervising sex offenders, briefly summarize prior research on federal sex offenders, and detail the data and methods used in this study. Afterwards, principal findings will be highlighted, and we conclude by discussing policy implications and directions for future research.

Federal Policy on Supervising Sex Offenders

The Administrative Office of the U.S. Courts —Probation and Pretrial Services Office (AOUSC-PPSO) has responded to the growing number of sex offenders under federal post-conviction supervision and the concerns that many online child pornography offenders might be involved in offline contact sexual offending by issuing guidance for federal officers charged with supervising these offenders.

Under current policy, the potential threat that sex offenders pose to the community requires them to begin supervision at the "highest" levels until the officer has performed a thorough assessment using all information available. All offenders placed on supervised release after incarceration or sentenced to straight probation[4] have their risk to recidivate for any offense assessed using the PCRA.[5] The PCRA is a dynamic actuarial risk assessment instrument developed for federal probation officers that classifies offenders into the risk levels of low, low/moderate, moderate, or high (AOUSC, 2011).[6] These categories provide crucial information about an offender's likelihood of committing any offense or being revoked (AOUSC, 2011; Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011; Lowenkamp, Johnson, VanBenschoten, Robinson, & Holsinger, 2013). Importantly, however, the PCRA was not constructed to specifically measure an offender's sexual deviance or predict sexual recidivism (Lowenkamp et al., 2013).

The policies provide guidance regarding the intensity of supervision. The policies specifically state that all sex offenders should begin their supervision terms being supervised as high risk regardless of the PCRA's classification. This provides officers with the time to conduct investigations into the extent of an offender's sexually deviant background, observe their responses to treatment, and identify any protective factors, all while ensuring that the offender is being supervised at levels intensive enough to protect the community. The policies, however, guide officers to appropriately lower the supervision levels of sex offenders if, after this initial investigation, the officer determines that less intensive supervision can address risk while maintaining protection of the community.

Prior Research on Federal Sex and Child Pornography Offenders

Some recently published studies that examined recidivism of those convicted of child pornography and contact sex offenses against children in the U.S. include studies by the U.S. Sentencing Commission (2012) and Faust et al. (2014). In 2012, the USSC published a report to the U.S. Congress on the prosecution, sentencing, incarceration, and supervision of offenders convicted of federal non-production child pornography offenses. Part of this report examined the rearrest rates for 610 offenders sentenced to non-production child pornography offenses in 1999 and 2000. These offenders were tracked for an average of eight and a half years and counted as recidivists if they were arrested for any felony or misdemeanor offenses or had a technical violation leading to an arrest or revocation (USSC, 2012, pp. 295-296). The USSC reported a general recidivism rate of 30 percent and a sexual recidivism rate of 7 percent during the follow-up period. Through the presentence reports, the USSC found that about 33 percent of these offenders had a history of engaging in criminal sexually dangerous behavior (USSC, 2012).

Another study conducted by Faust et al. (2014), compared 428 offenders convicted of non-contact child pornography offenses to 210 offenders convicted of contact sex offenses involving children on several risk and recidivism-related factors. Overall, Faust et al. (2014) found that child pornography offenders had less substantial criminal histories and lower substance abuse rates than contact sex offenders; conversely, child pornography offenders tended to have higher

[3] Studies examining rates of contact sexual behavior through offender self-reporting have been criticized on the grounds that they could overinflate the contact rates because they rely on offenders participating in treatment programs "who have strong incentives to admit to sexual contacts, even if untrue, as a sign of their progress in treatment" (Seto et al., 2011: 126).

[4] Supervised release refers to offenders sentenced to a term of community supervision following a period of imprisonment within the Federal Bureau of Prisons (18 U.S.C. §3583). Probation refers to offenders sentenced to a period of supervision without any imposed incarceration sentence (18 U.S.C. §3561). Of the 135,142 federal offenders under supervision in fiscal year 2015, 86 percent were on supervised release and 13 percent were on probation.

[5] See Johnson, Lowenkamp, VanBenschoten, and Robinson (2011); Lowenkamp, Johnson, VanBenschoten, Robinson, and Holsinger (2013); and Lowenkamp, Holsinger, and Cohen (2015) for information about the construction, validation, and implementation of the PCRA in the federal supervision system.

[6] It should be noted that the PCRA is currently undergoing a revision which will involve the integration of a violence assessment into the instrument and result in offenders being placed into 12 different risk groups. At the time of this study, the revised PCRA had not yet been implemented; hence, we continue anchoring our offender population into the four risk groups discussed above.

rates of pre-incarceration employment and education levels than offenders convicted of child-related contact sex offenses. These researchers reported overall arrest rates nearly three times higher for the contact (25.7 percent) compared to the child pornography (9.1 percent) offenders. The differences in arrest rates held even when controlling for other recidivism-related characteristics such as criminal history and substance abuse.[7]

Data and Methods

Participants and Sex Offender Types

Data for this study were obtained from 94 federal judicial districts and comprised 7,416 male sex offenders released from federal prison and placed on supervision during fiscal years 2007 through 2013. In a method similar to that used by Faust & Motivans (2015), we identified sex offenders and placed them into broader categories by using the title and section of the U.S. Criminal Code associated with their instant conviction offense. The U.S. Criminal Codes was extracted from the Probation and Pretrial Services Automated Case Tracking System (PACTS), the case management system used by federal probation and pretrial officers. Through this process, we were able to categorize the 7,400 sex offenders into the following groups of sexual offenses involving either children or nonconsenting adult victims.[8]

Child pornography (N = 4,462)

18 U.S.C. § 1470: Transfer of obscene material to minors
18 U.S.C. § 2251: Sexual exploitation of children
18 U.S.C. § 2251A(a)(b): Selling or buying of children
18 U.S.C. § 2252: Certain activities relating to material involving the sexual exploitation of minors
18 U.S.C. § 2252A: Certain activities relating to material constituting or containing child pornography
18 U.S.C. § 2260(a)(b): Production of sexually explicit depictions of a minor for importation into the United States

Transportation for illegal sexual activity (N = 800)

18 U.S.C. § 1591: Sex trafficking of children or by force, fraud, or coercion
18 U.S.C. § 2422: Coercion and enticement
18 U.S.C. § 2423(a)(b): Transportation of minors
18 U.S.C. § 2425: Use of interstate facilities to transmit information about a child relating to illicit sexual activity

Sexual abuse or assault (N = 1,030)

18 U.S.C. § 2241: Aggravated sexual abuse
18 U.S.C. § 2242: Sexual abuse
18 U.S.C. § 2243: Sexual abuse of a minor or ward
18 U.S.C. § 2244: Abusive sexual contact
18 U.S.C. § 2245: Sexual abuse resulting in death

Sex Offense Registration and Notification Act (SORNA) (N = 874)

18 U.S.C §2250: Failure to register as sex offender

There was also a category of sex offenders (N= 250) that we were unable to classify according to their convicted offenses, as the statute codes in PACTS were labelled "other sex" offenses. We included these "other" offenders in the totals but excluded them from most of the analyses comparing sex offender conviction types. Last, those convicted of child pornography were further categorized by whether they had any official arrest or conviction record of contact sexual behavior prior to or concomitantly with their current offense.

Identifying Sex Offenders with an Official Arrest or Conviction Record of Contact Sexual Behavior

We further classified those on federal supervision for a sex offense according to whether they evidence any contact sexual behavior in their official records. In this study, having an official record of contact sexual behavior means that the offender was either arrested for or convicted of an offense involving contact sexual offenses (e.g., sexual assault, child molestation, child pornography production, child trafficking, etc.) before or for the current offense. We were unable to measure incidences of self-reported contact behavior that might have arisen through polygraphs or other investigative means for this study.

Being able to measure the presence of an official record of contact sexual behavior is especially important when examining Internet child pornography offenders, because research shows that offenders who commit child pornography and contact sex crimes tend to have higher risk levels and recidivism rates compared to child pornography-only offenders (Babchishin, Hanson, & VanZuylen, 2015). We used a combination of Static-99 data from the Federal Bureau of Prisons (BOP) and arrest history data to identify offenders with past or present evidence of contact criminal sexual behavior. The Static-99 is an actuarial risk prediction instrument that estimates the probability of sexual and/or violent reconviction for adult males who have already been charged with or convicted of at least one contact sexual offense against a child or nonconsenting adult (Harris, Phenix, Hanson, & Thornton, 2003). This instrument is scored on all sex offenders incarcerated within the U.S. federal prison system with an official current or prior arrest or conviction record of contact sexual offending. It is used by the BOP to screen for potential civil confinement.

The Static-99 scoring rules preclude this instrument from being used on offenders who have only been arrested for or convicted of non-violent sexual offenses including prostitution, consensual sexual activity, or online non-production child pornography (Harris et al., 2003). Hence, the Static-99's scoring rules allowed us to deduce that, if the offender was scored on this instrument, they had an official arrest or conviction record of contact sexual behavior. In addition to those with a Static-99, any offender whose criminal history indicates a prior arrest for sexual assault or sexual exploitation was classified as having an official record of contact sexual behavior.

The decision to use the Static-99 for the purpose of identifying sex offenders with an official record of contact sexual behavior necessitated that we exclude certain offenders from our analysis. Specifically, the 7,400 sex offenders were extracted from a larger database containing 9,583 offenders with an instant conviction for a sex offense between fiscal years 2005 through 2013. We excluded all female sex offenders (*n* lost = 215 offenders) and offenders sentenced to probation-only sentences (*n* lost = 522 offenders), as neither of these groups would be scored on the Static-99 by the BOP. We also removed all

[7] Other studies focused on the key trends taking place in the prosecution, incarceration, or supervision of federal sex offenders (see Faust & Motivans, 2015) or examined the frequency with which sex offenders self-reported contact sexual behavior either during polygraph (see DeLisi et al., 2016) or in treatment (see Bourke & Hernandez, 2009).

[8] It's important to note that the number of sex offenders analyzed in the current study will not approximate the numbers under active federal supervision reported by PPSO's internal systems (i.e., Decision Support Systems). This discrepancy is partially explained by the fact that DSS includes under its sex offender definition any offender with a current sex offense conviction or with a history of engaging in sexually criminal behavior.

offenders received onto federal supervision before 2007, as the BOP was not uniformly applying this instrument before that year (*n* lost = 1,304). Last, since we wanted to track offender recidivism patterns, we removed all offenders without criminal history information from the file (*n* lost = 126).

Offender Recidivism Outcomes

We defined recidivism as any arrest for new crimes (excluding arrests for technical violations of the conditions of supervision) that took place between the offender's release from federal custody date and the last date these arrest data were assembled (i.e., 3/17/2015). New arrest events encompassed the following major offense categories: arrests for any felony or misdemeanor offenses, arrests for violent nonsexual offenses (e.g., homicide and related offenses, kidnapping, robbery, and assault), and arrests for any sexual offenses violent or nonviolent (e.g., child pornography, sexual assault, and sexual exploitation).[9] We combined violent and non-violent sexual arrest activity because, as will be shown, the base rates for sexual recidivism were fairly low. We also examined the rates at which offenders were revoked during their supervision term. Revocation information was retrieved from PACTS and included any revocation that took place from the start of active supervision until the last date of revocation information retrieval (i.e., 10/30/2014).

Analytical Plan

The current study primarily uses descriptive statistics to provide an overview of the general (i.e., non-sexual) risk characteristics and recidivism rates for offenders convicted of federal sex offenses. Specifically, this study categorizes the 7,400 federal sex offenders by their instant conviction offenses, assesses how many have an official record of contact sexual behavior, details their demographic profiles, and describes their risk characteristics as measured by the PCRA. We then examine the recidivism and revocation rates within a fixed period while under post-conviction supervision. In addition, we compare the 7,416 male sex offenders with a group of 179,812 male non-sex offenders placed on post-conviction supervision during the same time period.

The final component of this study uses multivariate techniques (i.e., logistic regression) to investigate the PCRA's power at predicting general recidivism and revocation outcomes. As will be shown, the overall recidivism and revocation rates for those with instant offense convictions for sexual assault or violations of the Sex Offender Registration and Notification Act (SORNA) are significantly higher than those of the child pornography offenders. Multivariate logistic regression techniques were employed to examine whether these differences in recidivism and revocation rates still held when the PCRA was included as a statistical control. In addition, we employed an AUC-ROC (area under curve - receiver operating characteristics) analysis to assess the PCRA's ability to predict specific recidivism events including arrests for any, violent (non-sexual), or sexual offenses or revocations from supervision.

Results

Most Common Instant Conviction Sex Offenses

Table 1 examines the most common instant conviction offenses for sex offenders received into federal supervision between fiscal years 2007 through 2013 and the percentage of these offenders with an official record of arrests or convictions for contact sexual behavior. Offenders convicted of possession, receipt, distribution, or production of online child pornography accounted for the largest numbers of sex offenders under post-conviction supervision. Three-fifths (60 percent) of the 7,416 federal sex offenders had an instant offense conviction for online child pornography offenses, while the remainder were convicted of sexual abuse or assault (14 percent), SORNA violations (12 percent), or transporting minors for illegal sexual activity (hereafter illegal transportation) (11 percent). Three percent of the sex offenders in our study population were unclassifiable.

We also used the presence of a Static-99 score and criminal history data to determine the percent of sex offenders with an official record of past or present contact sexual behavior.[10] Half of the sex offenders under post-conviction supervision had an official record of engaging in contact sexual behavior, meaning that they were either scored on the Static-99 or had a prior arrest for sexual assault or exploitation. Over 90 percent of offenders convicted of sexual assault (91 percent), illegal transportation (91 percent), or SORNA (95 percent) offenses evidenced an official record of contact sexual behavior. Conversely, 24 percent of online child pornography offenders had been arrested for or convicted of contact sexual offenses.

Some caution should be used in interpreting these results on the presence of contact sexual behavior. First, the column identifying

TABLE 1.
Percent of federal sex offenders with official record of contact sexual behavior

		Percent of offenders with —		
Instant sex offense at conviction	Number	Any official contact behavior	Static-99	Prior arrest for sex assault or exploitation
All sex offenders	7,416	49.5%	43.6%	25.0%
Child pornography	4,462	23.6%	18.6%	12.0%
Other-not classifiable[a]	250	54.4%	45.6%	28.4%
Sexual assault	1,030	90.6%	86.0%	28.9%
SORNA[b]	874	94.6%	82.0%	75.1%
Transportation for illegal sexual activity	800	90.9%	85.9%	36.0%

Note: Includes federal offenders placed on supervised release between fiscal years 2007 through 2013.

Percentages will not sum to totals as offenders can have both a Static-99 and prior arrest for sexual assault or exploitation. The prior sex assault/exploitation arrest variable excludes offenses that resulted in the offender being placed on federal supervision.

[a]The non-classifiable sex offenders are excluded from subsequent analyses as a specific offense category but included in totals.

[b]Includes offenders convicted of violating the Sexual Offender Registration and Notification (SORNA) act.

[9] Prostitution offenses were excluded from the sexual recidivism events.

[10] See methods section on how we used the Static-99 to assess whether the offender had an official background of contact sexual behavior.

"any evidence of contact offending" is based on an official record of an arrest or conviction and does not include self-reported behavior. Previous research has found about half of online child pornography offenders admitting to some form of prior sexual contact with children (Seto et al., 2011). Moreover, while it might seem surprising that 14 percent of offenders convicted of sexual assault did not have a Static-99, this offense category also includes offenders convicted of consensual (e.g., statutory rape) as well as forcible sexual assault. We were unable to identify offenders convicted of consensual sexual acts through the PACTS offense coding scheme.

Table 2 shows the sex offense conviction categories included in the study. For the 36 percent of offenders convicted of sexual assault, violation of SORNA laws, or illegal transportation, no information on their contact backgrounds was integrated into the analysis because as previously shown (see Table 1), nearly all of these offenders had a record of contact sexual behavior. For most offenders convicted of sexual assault, SORNA violations, or illegal transportation, their past or present conduct would inherently involve some form of contact sexual behavior necessitating a Static-99. For those convicted of child pornography offenses, we used information from the Static-99 and criminal history data to place them into (1) an online child pornography-only group and (2) a group containing child pornography offenders with arrest or conviction records for contact sexual behavior. Table 2 also shows the offense distributions for sex offenders with PCRA assessments. Since the study includes sex offenders placed on supervised release between fiscal years 2007-2013, not all had PCRA assessments, because implementation of this risk instrument did not begin until mid-2010.

Demographic Characteristics of Federal Sex Offenders

Table 3 shows the demographic characteristics of federal sex offenders based upon their instant conviction offense. Whites accounted for 81 percent of the general sex offender population; among non-sex offenders, whites comprised 57 percent of the total population. Nearly all offenders (95 percent) convicted of child pornography offenses were white, while minorities accounted for higher portions of the non-child pornography sexual offenses. American Indians and Alaska Natives, for example, comprised 71 percent of the sexual assault offenders and African Americans accounted for 26 percent of the SORNA offenders. Sex, and especially child pornography offenders, skewed older. At the time of being placed on post-conviction supervision, the average sex offender was 45 years old, while child pornography offenders averaged 46 years in age.

PCRA Risk Characteristics of Federal Sex Offenders

Figure 1 provides information on the initial PCRA risk classifications for federal sex offenders by their instant conviction offense.[11] In general, sex offenders, with the exception of those convicted of sexual assault and SORNA laws, had lower risk levels than the non-sex offender population. For example, 12 percent of the sex offenders with PCRA assessments were classified as either moderate or high risk; in comparison, 26 percent of the non-sex offenders were grouped into the moderate- or high-risk categories. Child pornography offenders were especially likely to be considered low risk, with nearly all (97 percent) of these offenders initially being assessed in the low or low/moderate risk categories. A slightly higher percentage of child pornography offenders with official records of contact sexual behavior garnered a moderate- or high-risk PCRA classification (8 percent) compared to child pornography offenders without contact histories (2 percent). Among offenders convicted of non-child pornography offenses, almost half the SORNA (47 percent) and about a fourth of those convicted of sexual assault (27 percent) were classified moderate or high risk by the PCRA.

Table 4 depicts the average total PCRA scores as well as the average PCRA domain scores in criminal history, education/employment, substance abuse, social networks, and supervision attitudes for sex offenders and compares them to the average PCRA scores for non-sex offenders. In contrast to non-sex offenders, sex offenders averaged lower scores in the PCRA domains of criminal history, education/employment, and substance abuse; however, sex offenders manifested higher average scores in the PCRA domains of social networks and supervision attitudes. Within the specific sex offender categories, offenders convicted of child pornography scored consistently lower in most of the PCRA domains than the sexual assault or SORNA offenders. Not surprisingly, child pornography offenders with contact records of sexual offending received higher PCRA criminal history scores than their child pornography counterparts without contact records. For those offenders convicted of illegal transportation, their average PCRA scores, with the exception of criminal history, were similar to those of child pornography offenders.

Recidivism Outcomes of Federal Sex Offenders

Table 5 depicts the three-year recidivism

TABLE 2.
Instant conviction sex offense for federal sex offenders including subset with Post Conviction Risk Assessments (PCRA)

Instant sex offense at conviction	All offenders		Subset with PCRAs	
	Number	Percent	Number	Percent
All sex offenders	7,416	100%	5,284	100%
Any child pornography offense	4,462	60.2%	3,420	64.7%
No record of contact behavior	3,411	46.0%	2,651	50.2%
Official record of contact behavior	1,051	14.2%	769	14.6%
Other-not classifiable	250	3.4%	75	1.4%
Sexual assault	1,030	13.9%	548	10.4%
SORNA	874	11.8%	674	12.8%
Transportation for illegal sexual activity	800	10.8%	567	10.7%

Note: Includes federal offenders placed on supervised release between fiscal years 2007 through 2013 with and without PCRA assessments. Federal offenders began receiving PCRA assessments in mid-2010s. The non-classifiable sex offenders are excluded from subsequent analyses as a specific offense category but are included in totals.

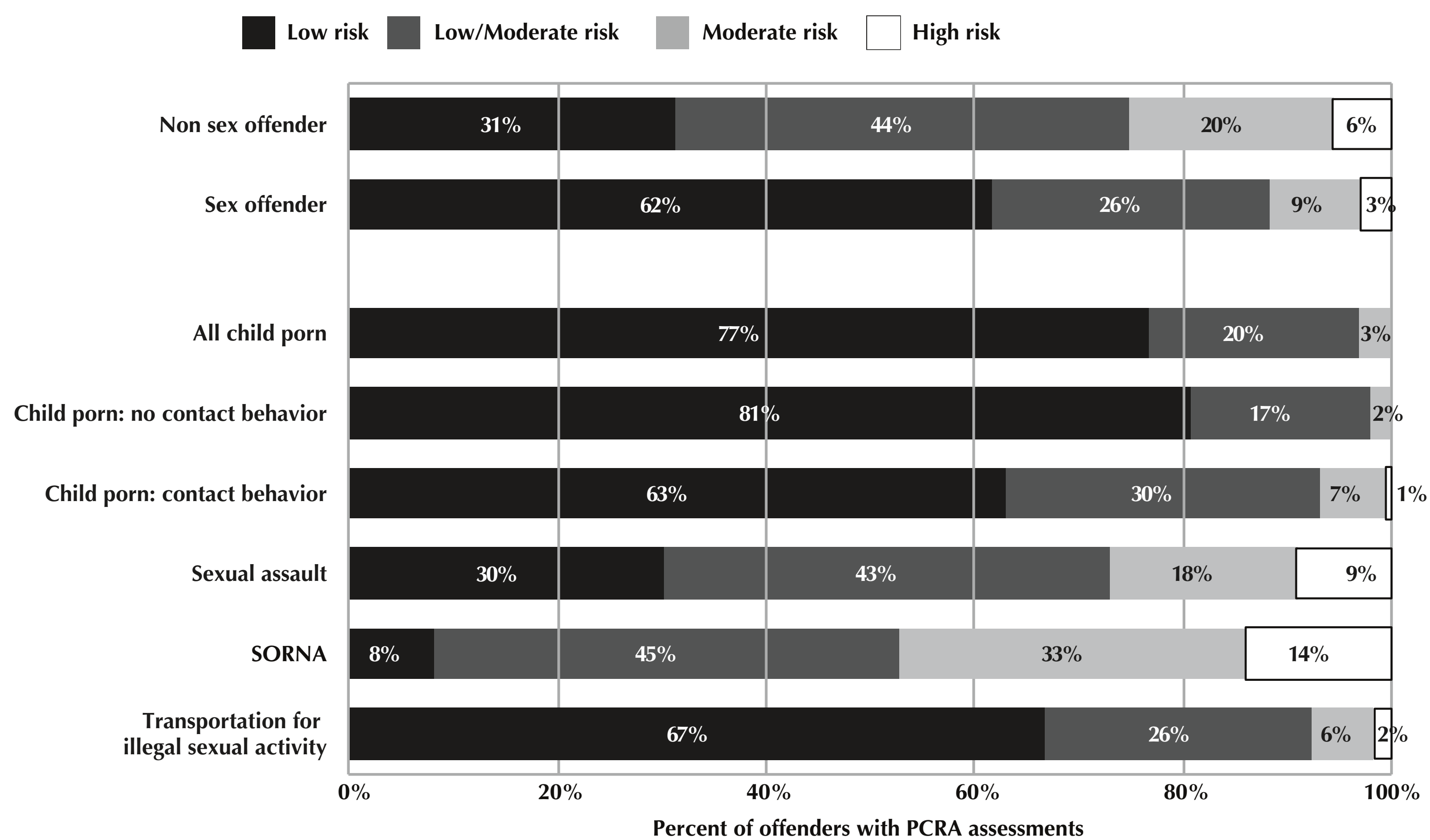
[11] Figure is limited to subsample of 5,284 offenders with PCRA assessments. Adjusted supervision levels not shown.

TABLE 3.
Demographic characteristics for federal sex offenders, by instant conviction sex offense

	Instant conviction sex offense					All offenders	
	Child pornography						
Demographic characteristics	No record of contact behavior	Official record of contact behavior	Sexual assault	SORNA	Transportation for illegal sexual activity	Sex	Non-sex
Race							
Asian	1.6%	1.2%	1.0%	0.3%	4.3%	1.5%	2.6%
Black	2.9%	3.0%	6.8%	25.5%	8.9%	6.9%	38.8%
American Indian	0.5%	0.7%	71.0%	2.8%	0.5%	10.7%	1.8%
White	95.0%	95.1%	21.2%	71.4%	87.4%	80.9%	56.8%
Hispanic ethnicity	5.8%	5.3%	5.8%	10.9%	6.4%	6.4%	25.6%
Average age (yrs.)	45.8	46.3	38.7	43.4	44.2	45.1	38.7
Number of offenders	3,411	1,051	1,030	874	800	7,416	179,812

Note: Table includes sex offenders received into federal supervision between 2007 through 2013.

FIGURE 1.
Post conviction risk assessment (PCRA) risk categories for federal sex offenders, by instant conviction sex offense



Note: Includes subset of offenders with PCRA assessments

TABLE 4.
Average Post Conviction Risk Assessment (PCRA) domain scores for federal sex offenders by instant sex conviction offense

Instant sex offense at conviction	Number of offenders	Average total PCRA score	Average PCRA domain scores: Criminal history	Education/ Employment	Substance abuse	Social networks	Supervision attitudes
All offenders							
Non-sex offender	97,537	7.17	4.59	1.12	0.28	1.08	0.11
Sex offender	5,284	5.08	2.60	0.98	0.12	1.23	0.15
Convicted sex offense							
All child pornography	3,420	3.92	1.74	0.83	0.07	1.15	0.14
No record of contact behavior	2,651	3.62	1.53	0.79	0.07	1.10	0.13
Official record of contact behavior	769	4.95	2.48	0.95	0.07	1.29	0.15
Sexual assault	548	7.44	4.28	1.30	0.31	1.38	0.17
SORNA	674	9.36	5.64	1.59	0.25	1.64	0.24
Transportation for illegal sexual activity	567	4.64	2.43	0.83	0.07	1.15	0.15

Note: Includes subset of federal sex offenders with PCRA assessments.

rates for sex offenders during their post-conviction supervision term. To be included in this table, the offender's recidivism event had to be observable for a minimum of three years and their court-ordered supervision terms had to be three years or more (Baber, 2015). Offenders, for example, were counted as having recidivated if at any time during the three years in which they were sentenced to supervised release they were either arrested or revoked. Using this approach, offenders sentenced to less than three years of supervised release or whose recidivism event could not be followed for a minimum of three years were excluded from this analysis. Recidivism events occurring after the supervision term or outside the three-year follow-up period were also omitted.

Table 5 generally shows persons convicted of sex offenses being arrested or revoked less frequently than those convicted of non-sex offenses. For example, nearly a fifth (18 percent) of sex offenders were arrested for any offense during their first three years of supervision, while about a third (31 percent) of non-sex offenders had any arrest during the same time period. The percentage of sex offenders arrested for non-sexual violent offenses (2 percent) was also lower in comparison to non-sex offenders (8 percent). Sex offenders, however, were three times more likely to be arrested for sexual offenses (3 percent) than non-sex offenders (1 percent).

Among the sex offense types, those offenders under supervision for SORNA or sexual assault were arrested or revoked at the highest rates, while those under supervision for child pornography offenses had lower recidivism rates. For example, 42 percent of the SORNA and 23 percent of the sexual assault offenders were arrested for any offense within three years of their supervision start dates, compared to 13 percent of child pornography offenders. The percentage of offenders arrested for non-sexual violent offenses was also higher for the SORNA (8 percent) and sexual assault (4 percent) offenders than for offenders on supervised

TABLE 5.
Three-year recidivism rates for federal sex offenders while under supervision, by instant conviction offense

Instant offense at conviction	Number	Recidivism outcomes: Any arrest	Major arrest[a]	Non-sexual violent arrest	Any sex arrest	Probation revocation
All offenders						
Non-sex offender	89,615	31.4%	23.0%	7.9%	0.5%	22.6%
Sex offender	3,909	17.5%	7.8%	1.8%	2.8%	19.2%
Conviction sex offense						
All child pornography	2,287	13.0%	4.9%	0.5%	2.6%	11.6%
No record of contact behavior	1,722	12.5%	4.3%	0.4%	2.2%	9.5%
Official record of contact behavior	565	14.7%	6.9%	0.7%	4.1%	18.2%
Sexual assault	605	23.1%	9.9%	3.6%	2.2%	38.5%
SORNA	299	41.8%	26.1%	8.0%	7.7%	47.2%
Transportation for illegal sexual activity	550	17.3%	7.8%	1.6%	2.4%	14.4%

Note: Sub-sample used for 3-year arrest rates is restricted to actively supervised TSR cases for which the offender was sentenced to at least 3 years of supervision.

[a]Excludes minor offenses including breaches against public peace, invasion of privacy, prostitution, obstruction of justice, liquor law violations, and traffic offenses.

release for child pornography (1 percent). There was less variation between the instant offense categories in regards to recidivism for sex offenses; child pornography offenders, for example, were arrested for sexual offenses at rates (3 percent) similar to that of sexual assault (2 percent) or illegal transportation (2 percent) offenders.

For offenders convicted of child pornography offenses, having an official record of contact sexual behavior was generally not associated with significantly higher recidivism rates. The general rearrest rates for child pornography offenders with contact sexual records (15 percent) was nearly the same as child pornography offenders without any records of contact sexual offending (13 percent). Officers, however, were more likely to revoke child pornography offenders with contact sexual records (18 percent) compared to their counterparts with no official record of contact sexual behavior (10 percent).

Predictive Efficacy of PCRA for Federal Sex Offenders

We examined whether the differences in arrest rates across the convicted sex offense categories reported in Table 5 still hold when the offender's PCRA scores are introduced as statistical controls. We conducted this examination by calculating the predicted probabilities of arrest or revocation after initial PCRA assessment for each of the major sex offense categories used in this study. These predicted probabilities were generated through a statistical technique (i.e., logistic regression) that allows us to examine the relationship between the instant conviction offenses and recidivism/revocation outcomes while holding constant the offender's PCRA scores at their means. We compare arrest/revocation probabilities for those convicted of child pornography offenses with no record of contact sexual behavior to those convicted of child pornography with an official record of contact sexual behavior, illegal transportation, sexual assault, or SORNA offenses. Significant differences between the child pornography (non-contact) and other offense types are noted by an asterisk.[12]

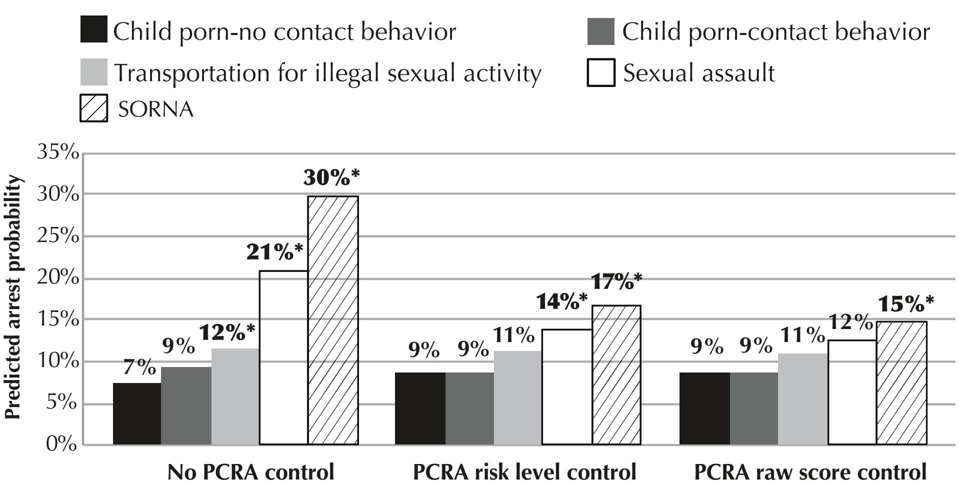
Figure 2 shows the predicted probability of any arrest across the sex offender offense categories initially without any controls and then adjusts these probabilities by statistically controlling for an offender's PCRA risk levels and raw scores. The model without any PCRA controls produces predicted arrest patterns similar to the bivariate analysis shown in Table 5. Specifically, the estimated arrest probabilities for the illegal transportation (12 percent), sexual assault (21 percent), and SORNA (30 percent) offenses are significantly higher compared to the estimated arrest probability for child pornography offenders with no record of contact sexual behavior (7 percent). Once the estimated arrest probabilities have been adjusted for an offender's PCRA risk level or raw score, they are less substantial across the sex offense categories. For example, adjusting the probability of arrest to take into account an offender's raw PCRA risk score generates estimated arrest probabilities between those convicted of child pornography offenses with no official record of contact sexual behavior (9 percent) that were not significantly different from those convicted of child pornography with official records of contact sexual behavior (9 percent), illegal transportation (11 percent), and sexual assault (12 percent). Only the SORNA offenders continued to manifest predicted arrest probabilities that are significantly higher (15 percent) than the non-contact child pornography offenders.

We also generated predicted revocation probabilities taking into account an offender's PCRA risk level or raw score, which are shown in Figure 3. In results similar to the prior analysis, we initially show significant differences in the predicted revocation probabilities between the non-contact child pornography offenders and the other sex offender types; however, when the PCRA is used to statistically control for the risk of revocation, the differences in the likelihood of revocation diminish across the sex offender categories.

Last, we focus on the PCRA's utility to predict recidivism outcomes for persons convicted of federal sex offenses through an AUC-ROC (area under curve - receiver operating characteristics) analysis and by examining the recidivism rates across the four PCRA risk categories (e.g., low, low/moderate, moderate, and high). The AUC-ROC score is frequently used to assess risk assessment instruments and is often preferred over a correlational analysis because it is not impacted by low base rates (Lowenkamp et al., 2013). Essentially, the AUC-ROC measures the probability that a score drawn at random from one sample or population (e.g., offenders with a rearrest) will be higher than that drawn at random from a second sample or population (e.g., offenders with no rearrest) (Lowenkamp et al., 2013; Rice & Harris, 2005). Values for the AUC-ROC range from .0 to 1.0, with values of .70 or greater indicating that the

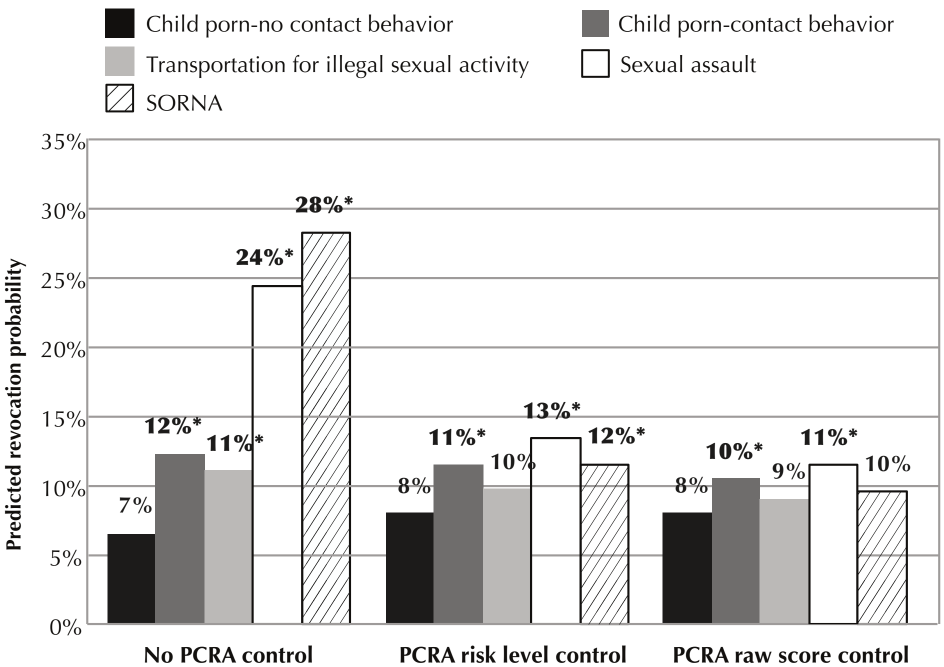
[12] Estimated arrest/revocation probabilities reported in Figures 2 and 3 will differ from percentages reported in Table 5 because these probabilities are estimated for shorter time periods (e.g., date of PCRA assessment) and are not restricted to arrests during supervision terms.

FIGURE 2.
Predicted probabilities of any arrest for federal sex offenders controlling for Post Conviction Risk Assessment (PCRA), by instant conviction offense



Note: Includes 5,284 federal sex offenders with PCRA assessments. Bold denotes significance difference between child porn offender with no contact behavior and the other sex offender offense categories. Arrest probabilities will differ from percentages reported in Table 5 as they track arrests for shorter time periods (e.g., date from PCRA assessment) and are not restricted to arrests during supervision terms.
* p <.05

FIGURE 3.
Predicted probabilities of revocation for federal sex offenders controlling for Post Conviction Risk Assessment (PCRA), by instant conviction offense



Note: Includes 5,284 federal sex offenders with PCRA assessments. Bold denotes significance difference between child porn offender with no contact behavior and the other sex offender offense categories. Revocation probabilities will differ from percentages reported in Table 5 as they track revocations for shorter time periods (e.g., date from PCRA assessment).
* p < .05

actuarial instrument does fairly well at prediction (Andrews & Bonta, 2010). Figure 4 shows the AUC-ROC scores for offenders in this study exceeding the .70 threshold for most of the recidivism outcomes, including any arrests (.72), violent arrests (.79), and probation revocations (.77). The AUC-ROC scores fell under the .70 threshold for only those outcomes associated with sexual recidivism (.63).

In addition to an AUC-ROC analysis, Figure 4 also shows the failure rates involving any arrests, non-sexual violent arrests, any sex arrests, and probation revocations by PCRA risk category for offenders with sex offense convictions. Among the non-sexual recidivism outcomes, the failure rates followed the anticipated pattern of increasing incrementally by each PCRA risk category. The recidivism rates for any arrest activity, for instance, increased from 7 percent for low-risk offenders to 15 percent for low/moderate, 33 percent for moderate, and 46 percent for high-risk offenders. Similar patterns of monotonically increasing failure rates also occurred for recidivism outcomes involving probation revocations and non-sexual violent arrests. The sexual recidivism outcome, however, manifested a weaker relationship with the PCRA risk groupings. The percentage of offenders rearrested for sexual offenses did not differ significantly for the low/moderate (4 percent), moderate (4 percent), and high (5 percent) PCRA risk categories (χ^2, 2 = .9017, p = .637).

Discussion and Conclusion

Summary of Major Findings

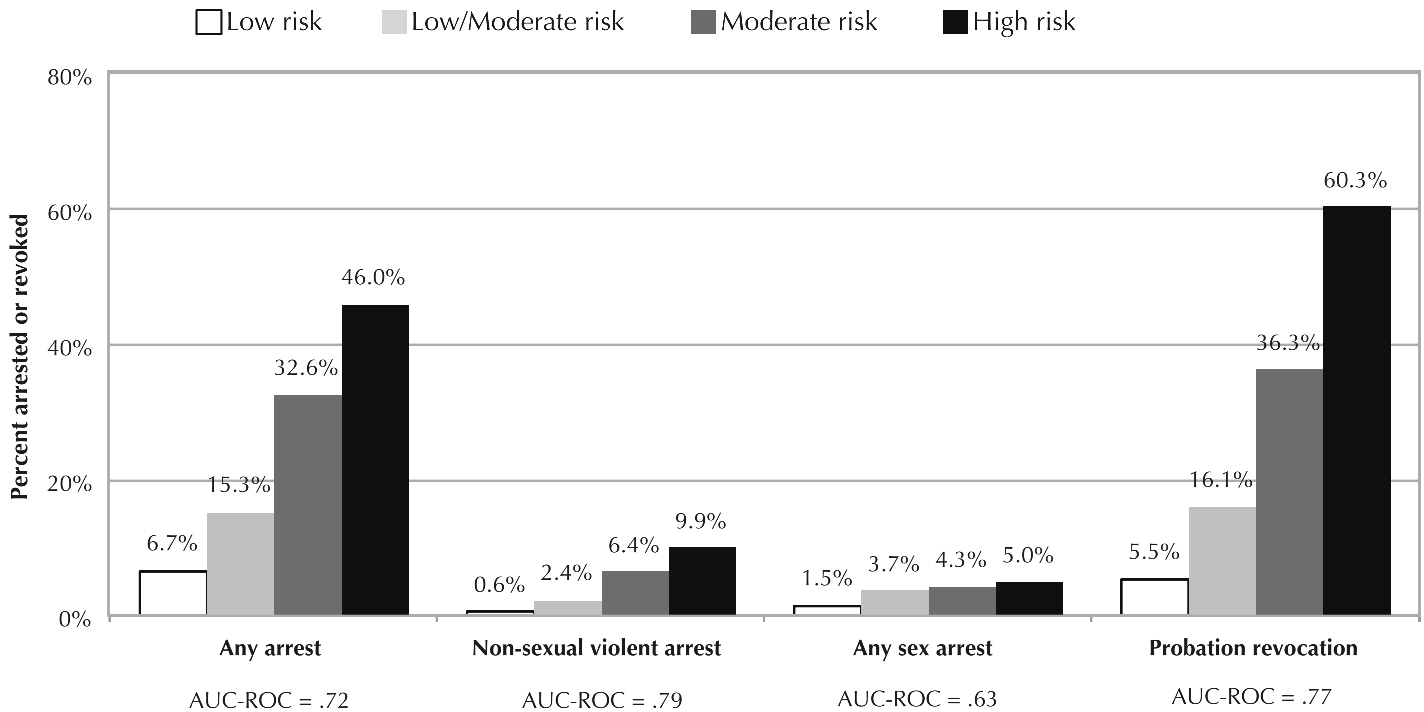
This study produced several key findings about persons convicted of sex offenses under federal post-conviction supervision. First, it shows those convicted of child pornography offenses accounting for the majority of sex offenses (60 percent), with the other offense types of sexual assault, illegal transportation, and SORNA accounting for 36 percent of federally supervised sex offenders. An examination of the official contact sexual backgrounds of these offenders shows that over 9 out of 10 of the non-child pornography offenders have an official conviction for, or arrest history of, engaging in contact sexual offenses. For those convicted of child pornography offenses, about a fourth of them had an official record of contact sexual offenses.

While those convicted of sex offenses in general scored lower on the PCRA and recidivated less frequently than those convicted of non-sex offenses, there was substantial heterogeneity in the recidivism rates and PCRA risk measures among the instant sex offense types. Specifically, those convicted of child pornography offenses had less serious criminal history backgrounds, attained higher levels of education and employment, suffered less frequently from substance abuse problems, and had stronger social support networks than those convicted of sexual assault or SORNA offenses. In fact, almost all the child pornography offenders were classified as either low or low/moderate risk by PCRA. Conversely, those convicted of sexual assault or SORNA offenses manifested general risk characteristics that were either similar to those of people convicted of non-sex offenses or, in the case of the SORNA, substantially higher.

Similar to the PCRA analysis, the recidivism patterns also varied across the conviction types. Offenders convicted of child pornography exhibited lower general and violent rearrest rates and supervision revocations compared to offenders convicted of SORNA or sexual assault. The recidivism activity for the SORNA offenders was particularly high, with about two-fifths of these offenders being rearrested within the three-year follow-up period. For the sexual recidivism outcome, however, there was less variation in arrest rates by conviction offense. It is also notable that those convicted of illegal transportation were rearrested or revoked at rates more similar to the child pornography than to the sexual assault and SORNA offenders. In a finding mirroring other studies, our analysis showed sex offenders being rearrested more frequently for non-sexual than sexual offenses (USSC, 2012; Hanson & Morton-Bourgon, 2009). Last, within the child pornography offense category, those offenders with an arrest or conviction record for contact sexual behavior evidenced only slightly higher-risk characteristics and reoffending behavior compared to child pornography offenders without any official background of contact sexual offenses.

The logistic regression and AUC-ROC analysis showed the PCRA performs well in predicting general rearrest and revocation outcomes for the 5,284 federal sex offenders with PCRA assessments analyzed in this study. Logistic regression results showed little to no significant differences among the arrest odds by the specific sex offender conviction types when controlling for the PCRA scores. Moreover, the AUC-ROC scores of .70 or above for the any arrest, violent arrest,

FIGURE 4.
Post Conviction Risk Assessment (PCRA) failure rates for any arrest, non-sexual violent arrests, any sex arrests, or probation revocations for federal sex offenders



Note: Includes 5,284 federal sex offenders with PCRA assessments. Arrest and revocation percentages will differ from percentages reported in Table 5 as they track recidivism activity for shorter time periods (e.g., date from PCRA assessment) and are not restricted to recidivism during supervision terms. Includes any arrests or revocations that occurred after the intial PCRA assessment.

and revocation outcomes, combined with the anticipated pattern of incrementally increasing failure rates for these recidivism, measures by risk category, indicate that the PCRA can be used to predict general (non-sexual) recidivism outcomes for offenders with instant convictions for sex offenses. The crucial exception is the PCRA's ability to predict sexual recidivism, as the AUC-ROC analysis and an examination of arrest patterns across the PCRA risk groups show that the PCRA is less effective at predicting this type of recidivistic behavior. This finding is not too surprising, however, because the PCRA was never constructed to predict sexual recidivism nor was it designed to measure sexual deviance (Lowenkamp et al., 2013).

Most of the findings in this paper align with prior research on federal sex offenders and are consistent with the general empirical work focusing on recidivism prediction for the sex offender population. Specifically, prior research has shown that child pornography is the most common type of sex offense within the federal system and that offenders convicted of child pornography have fewer risk characteristics and recidivate less frequently compared to contact sex offenders (Babchishin et al., 2015; Faust & Motivans, 2015; Faust et al., 2014; USSC, 2012; Babchishin, Hanson, & Herman, 2011). Although the recidivism rates reported in this paper do not exactly match those reported by the USSC, this discrepancy is attributed to the longer follow-up periods and different methodologies for measuring recidivism used by the U.S. Sentencing Commission. However, the recidivism rates for child pornography offenders reported by Faust et al. (2014) of 9 percent are fairly close to the 13 percent arrest rate reported in this study. In addition, the overall pattern of sex offenders being rearrested at higher rates for non-sexual rather than sexual offenses is consistent with the above-cited studies and other meta-analytic reviews of sex offender recidivism (Faust et al., 2014; USSC, 2012; Hanson & Morton-Bourgon, 2009).

This research is also supportive of using general risk assessments for recidivism prediction on sex offenders. Nearly all of these studies have shown that risk assessments designed to predict general or violent recidivism among the overall offender population should perform equally well in predicting these outcomes for sex offenders (Wormith, Hogg, & Guzzo, 2012; Hanson & Morton-Bourgon, 2009; Hanson & Bussiere, 1998). The prediction capacities of generalized or violent risk assessment instruments, however, are less effective in predicting sexual recidivism compared to risk assessment instruments such as the Static-99 that are specifically designed to predict sexual rearrest outcomes (Hanson & Morton-Bourgon, 2009; Hanson & Bussiere, 1998). Our research showing the PCRA's efficacy at predicting general and violent recidivism, and being less effective in predicting sexual recidivism, is consistent with these prior research efforts.

Last, in a somewhat surprising finding, this research shows that child pornography offenders with backgrounds of contact sexual offending exhibit only slightly higher risk characteristics and recidivism rates compared to child pornography offenders with no records of contact sexual offending. This finding is at odds with some studies showing offenders who commit child pornography and contact crimes having significantly higher risk levels and recidivism rates compared to child pornography-only offenders (Babchishin et al., 2015). It is interesting to note, however, that the USSC also found similar rates of general recidivism between child pornography offenders with and without histories of criminally sexual dangerous behavior (USSC, 2012). Clearly more research is needed to discern whether offenders convicted of federal child pornography offenses can be disaggregated into more useful risk typologies.

Limitations and Areas for Future Research

This current study has several limitations that could be addressed by subsequent research. First, we did not consider self-reported contact offending behavior revealed through polygraphs or other investigative techniques. Prior research has shown about half of child pornography offenders admitting to a history of contact sexual offending (Seto et al., 2011). Subsequent research could assess the frequency of self-reported contact sexual behavior identified in a sample of offenders convicted of federal child pornography offenses. Another issue is the relatively short follow-up period of three years used in the current study. Sex offender recidivism studies typically reference the need to engage in long-term follow-ups involving periods of 5 to 20 years (Hanson, Morton, & Harris, 2003). Since our study covered only three years of offender recidivism activity, subsequent work should consider extending the recidivism follow-up terms. The decision to lengthen the follow-up period, however, should be informed by the fact that even studies tracking contact sex offenders for time periods of 20 to 30 years have shown about a third of these offenders eventually being arrested for new sexual offenses (Seto et al., 2011; Hanson, Steffy, & Gauthier, 1993).

Implications for Federal Probation Officers

The policy and procedures currently in place for the investigation and supervision of sex offenders were informed by the body of empirical knowledge available at that time. In general, the findings produced by this study align with this research used to inform federal sex offender policy and hence support the general framework of federal procedures on sex offender management.

This research supports the procedural guidance advising officers to use the PCRA to assess the risk of general recidivism and criminogenic needs for sex offenders and then augment this generalized risk picture with information pertaining to an offender's sexually deviant characteristics through an extensive investigation involving polygraphs, interviews, and discussions with treatment personnel. Moreover, it advises officers to use risk instruments such as the Static-99/2002 or Stable & Acute 2007 that are constructed to predict sexual recidivism to further understand an offender's propensities toward sexual deviance. The importance of supplementing the PCRA is supported by this research showing that the PCRA does not specifically assess an offender's risk of sexual recidivism or target those behaviors related to sexual deviance.

This research also highlights areas for further examination and potential enhancements in federal sex offender policies. Currently, federal policy recommends that all sex offenders begin supervision at the "highest" risk levels and then recommends that supervision intensity be adjusted downwards if and when an investigation of the offender's background indicates they are not at risk of committing contact sex offenses. With the availability of Static-99 scores from the BOP for those sex offenders with arrests or convictions for contact sexual offenses, officers can more accurately apply the risk principle to that group of sex offenders. Utilizing the Static-99, and supplementing it with information gleaned from polygraphs and other sources, may provide officers with the details required to thoroughly understand offenders' risk to sexually recidivate and classify them into appropriate supervision levels.

This research further supports federal policy that not all sex offenders have the same risk of recidivism generally and sexual offending specifically. Among the sex offender types, those offenders under supervision for SORNA or sexual assault were arrested or revoked at the highest rates, while the child pornography offenders exhibited lower recidivism rates. Hence, this research suggests that the sexual assault and particularly the SORNA offenders are of high concern for federal probation officers. Officers should consider assessing the SORNA offenders more closely beginning with their entrance into the criminal justice system as they evidence higher rates of general and violent recidivism compared to child pornography offenders.

References

Administrative Office of the U.S. Courts. (AOUSC) (2011). *An overview of the federal Post Conviction Risk Assessment*. Washington, D.C.: Administrative Office of the U.S. Courts.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct (5th Ed.)*. Cincinnati, OH: Anderson Publishing.

Babchishin, K. M., Hanson, R. K., & Hermann, C.A. (2011). The characteristics of online sex offenders: A meta-analysis. *Sexual Abuse: A Journal of Research and Treatment, 23*(1), 92-123.

Babchishin, K. M., Hanson, R. K., & VanZuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. *Archives of Sexual Behavior, 44,* 45-66.

Baber, L. (2015). Inroads to reducing federal recidivism. *Federal Probation, 79(3),* 3-8.

Bourke, M. L., & Hernandez, A. E. (2009). The "Butner Redux": A report of the incidence of hands-on child victimization by child pornography offenders. *Journal of Family Violence, 24,* 183-191.

DeLisi, M., Caropreso, D., Drury, A., Elbert, M., Evans, T. H., & Tahja, K. (2016). The dark figure of sexual offending: New evidence from federal sex offenders. *Journal of Criminal Psychology, 6*(1), 3-15.

Faust, E., Bickart, W., Renaud, C., & Camp, S. (2014). Child pornography possessors and child contact sex offenders: A multilevel comparison of demographic characteristics and rates of recidivism. *Sexual Abuse: A Journal of Research and Treatment, 27*(5), 1-19.

Faust, E., & Motivans, M. (2015). Sex offenders in the federal correctional system: The consequence of heightened attention on increased certainty and severity of punishment. *Justice Research and Policy, 16*(1), 81-98.

Hanson, R. K., Steffy, R. A., & Gauthier, R. (1993). Long-term recidivism of child molesters. *Journal of Consulting and Clinical Psychology, 61,* 646-652.

Hanson, R. K., & Bussiere, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*(2), 348-362.

Hanson, R. K., Morton, K. E., & Harris, J. (2003). *Sexual offender recidivism risk: What we know and what we need to know*. New York, New York: The New York Academy of Sciences.

Hanson, R. K., & Morton-Bourgon, K. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*(6), 1154-1163.

Hanson, R. K., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment,* *21*(1), 1-21.

Harris, J., Phenix, R., Hanson, R. K., & Thornton, D. (2003). *Static-99 Coding Rules Revised-2003*. Ottawa, Canada: Corrections Directorate Solicitor General Canada.

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

Lam, A., Mitchell, J., & Seto, M. C. (2010). Lay perceptions of child pornography offenders. *Canadian Journal of Criminology and Criminal Justice, 52,* 173-201.

Lowenkamp, C. T., Johnson, J., VanBenschoten, S. W., Robinson, C., & Holsinger, A. M. (2013). The Federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services, 10*(1), 87-96.

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA revisited: Testing the validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services, 12*(2), 149-157.

Motivans, M., & Kyckelhan. T. (2007). *Federal prosecution of child sex exploitation offenders, 2006*. Washington, D.C.: Bureau of Justice Statistics.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC, Cohen's d, and r. *Law and Human Behavior, 29,* 615-620.

Seto, M. C., Hanson, R. K., & Babchishin, K. (2011). Contact sexual offending by men with online sexual offenses. *Sexual Abuse: A Journal of Research and Treatment, 23*(1), 124-145.

U.S. Sentencing Commission. (2012). *Report to Congress: Federal child pornography offenders*. Washington, D.C.: U.S. Sentencing Commission.

Wolak, J., Finkelhor, D., & Mitchell, J. (2005). *Child pornography possessors arrested in Internet-related crimes*. Alexandria, VA.: National Center for Missing and Exploited Children.

Wolak, J., Finkelhor, D., & Mitchell, J. (2009). *Trends in arrests of "online predators."* Durham, NH: Crimes Against Children Research Center.

Wormith, J. S., Hogg, S., & Guzzo, J. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39*(12), 1511-1538.

Imagining Sentinel Event Reviews in the U.S. Probation and Pretrial Services System

Janette Sheil
Administrative Office of the U.S. Courts
James Doyle[1]
Of counsel to the Boston firm of Bassil, Klovee & Budreau
Christopher T. Lowenkamp
Administrative Office of the U.S. Courts

DNA EXONERATIONS OF wrongfully convicted defendants have thrown a new light on the problem of error in American criminal justice. The fact that people sometimes make mistakes came as no surprise to active practitioners, but the growing list of highly publicized disasters gradually revealed a gap in our system's design. Our criminal system lacks a feature that medicine, aviation, and other high-risk fields see as critical: a way to account for the sources of the tragic outcomes that no one intended, to learn their lessons, and to use those lessons to reduce the risk of recurrence.

Corrections and probation professionals shudder at the nightmare of analogous headlines raising their own version of these questions: Why did we *release* the wrong man, so that he could inflict catastrophic harm? Or why did we keep the *right* man, but past his maximum sentence? How did we inherit this mentally ill prisoner, when we had no program of safe and useful treatment available? Why did our testing or tracking procedures fail to raise red flags? Why did we miss the red flags when they were raised?

In aviation and in medicine the recognition has grown that most catastrophes can't be understood simply by finding a frontline individual to blame. These are not single-cause events. More often, they are system errors: the outcome of normal people doing normal work in normal organizations (Dekker, 2007). As Dr. Lucien Leape (1994), one of the pioneers in medicine's patient safety movement, put it:

> While an operator error may be the proximate "cause" of the accident, the root causes were often present within the system for a long time. The operator has, in a real sense, been "set up" to fail by poor design, faulty maintenance, or erroneous management decisions.

Stopping at disciplining a "bad apple" or tinkering with an isolated procedure can leave the underlying causes of an error lying in wait for the next practitioner who comes along. With this in mind, medical reformers adopted the battle cry "Every defect a treasure" (Berwick, 1989). If we have paid the price for a mistake, they reasoned, we should learn the preventive lessons it can teach. They argue for a pivot from a focus on blame to a focus on cutting future risk. Error is an inevitable part of the human condition, and, as safety expert James Reason (2000) put it, "We cannot change the human condition, but we can change the conditions humans operate in."

Reason compares an organization to Swiss cheese: having layers of defense or protections against errors, with the holes in the block of cheese representing the weakness in those defenses. In most cases, the holes in the block do not line up, so if you look through one hole you will not see daylight on the other side. A small error may occur, but one of the layers of defense will catch it before it cascades through the system. However, in some instances, the holes become completely aligned, allowing an error to traverse the block. Reason argues that we should look at our poor outcomes to try to find ways to reduce the holes and find the weaknesses in our organizational systems. Once this is done, we can add layers to catch smaller errors (2000).

Sometimes errors are tough to identify on first glance. It is not uncommon for employees to develop work-arounds or best ways of performing a task or a series of tasks more efficiently. After time, these diversions from policy and procedures, sometimes called practical drift (Snook, 2002), become accepted practice in the organization, especially in response to a reduced workforce. In normal operations, drift may go unnoticed, but in a critical high-profile situation any deviation from policy will be scrutinized. Conducting system-wide reviews can help uncover practical drift at all levels of the organization.

So the question remains, can the criminal

[1] The views expressed by Mr. Doyle in this article are his own and not those of any firm or agency.

justice system develop this capacity for "forward-looking accountability" (Sharpe, 2003)? Can we accept error as an inevitable element of the human condition and study known errors in a disciplined and consistent way? Can we share the lessons learned from these studies to prevent future errors? Can we focus on future risks instead of on blame for the past?

What is a Sentinel Event (SE) Review?

The word "sentinel" refers to a watchman who stands guard, detecting the first sign of a looming threat and sounding a warning. A sentinel event is a significant, unexpected negative outcome—such as a wrongful conviction, the failed supervision of a dangerous probationer, or the avoidable death of a vulnerable inmate—that signals a possible weakness in the system or process. It is likely to have been the result of compound errors and may provide—if properly analyzed and addressed—important keys to strengthening the system and preventing future adverse events or outcomes.

The goal of the process is not to mobilize a *performance* review aimed at an individual whenever some front-page catastrophe occurs, but to develop a regular practice of conducting an all-stakeholders, all-ranks, non-blaming, *event* review whenever a learning opportunity arises. That opportunity can be found in every tragedy. It can also be found in many "near miss" or "good catch" situations where the ultimate disaster was averted, but only by good luck, special vigilance, or a uniquely talented individual.

In these Sentinel Event reviews, features of the system that genuinely shaped the frontline decision-making (but would be dismissed as "excuse-making" in a more typical disciplinary performance review) can be raised and analyzed for their explanatory power. The "accountability" these reviews provide can reach not only the frontline operator who was the last person in the chain of delivery (for example, the nurse who delivered the medication) but that operator's superiors and the diverse upstream and downstream actors whose budgets, policies, training, and procedures shaped the frontline operator's environment and limited his or her options.

This approach has generated important changes in the fields of aviation and hospital patient safety. It has led not only to improved safety records, but to the creation of overall "cultures of safety" in which everyone, in every rank and role, feels individual responsibility for the safety of the collective outcome, and—maybe just as importantly—takes pride in and satisfaction from their unique contributions.

A typical hospital SE review would include a team of 4-6 people, including process experts as well as others from all levels of the organization. Individuals who were involved in the event are not a part of the team, but are interviewed for information. Factors that are reviewed likely include communication (including supervisory oversight), training, environment/equipment, experience, and rules/policies/procedures (National Patient Safety Foundation, 2016). All of these areas can contribute to human error. Most hospitals will provide feedback to the persons involved and submit their review results and an action plan to the Joint Commission, which is a nonprofit organization that accredits and certifies nearly 21,000 health-care organizations and programs in the United States. The Joint Commission provides support and expertise to the hospital during its reviews, shares "lessons learned" with the medical community, and helps raise the level of transparency in the medical profession, providing a message to the public that patient safety is critical (The Joint Commission, 2016).

The military also engages in After Action Reports as standard operating procedure to discuss unintended outcomes, enabling soldiers to discover for themselves what happened, why it happened, and how to sustain strengths and improve on weaknesses. Similarly, the National Transportation Safety Board conducts approximately 2,000 aviation accidents and incidents a year and about 500 transportation accidents (NTSB, 2016) and posts the well-organized investigation reports on the Internet.

In 2014, the National Institute of Justice (NIJ) began focusing on the applicability of the Sentinel Event process to the criminal justice system with the support of then-Attorney General Eric Holder, Jr. (NIJ, 2014), who offered the following words:

> With few exceptions, justice system professionals hold themselves to high standards of integrity and are thorough and exacting in their quest for answers. If we truly hope to get to the bottom of errors and reduce the chances of repeating them, then it is time we explore a new, system-wide, way of responding…

The NIJ recognized that it is unwise to simply assume that these changes can be imported seamlessly into the unique context of criminal justice, and it has dedicated substantial resources to conducting a rigorous investigation of how the core ideas of Sentinel Event reviews can be mobilized in differing criminal justice environments (NIJ, 2014).

To test the concept, NIJ selected three jurisdictions to participate as beta sites. One of the selected sites in Milwaukee formed a group of diverse participants and analyzed the kind of event that strikes fear into any practitioner's heart: the "wrongful release," with fatal consequences, of a youthful defendant. "This was a kid who had red flags all over him," John Chisholm, the Milwaukee County district attorney, who participated in the review, later said, "Why was he still in the community?" (Starr, 2015). The usual impulse would be to hunker down under a media storm, or to blame the judge or the frontline probation officer. But after months of meetings, the all-stakeholders event review process revealed that at almost every turn, the people who made decisions about the boy had not seen his larger pattern of violent behavior because they did not have access to his complete records, or did not see them. System reforms to communications and data-sharing followed.

Is this Process a Good Fit for Federal Probation and Pretrial Services?

The Probation and Pretrial Services Office (PPSO) within the Administrative Office of the U.S. Courts (AOUSC) has a long history of providing oversight of the work of the United States courts. This function fulfills the statutory requirement of the Director of the AOUSC, or his authorized agent, to investigate the work of the probation officers and promote the efficient administration of the probation system (18 § U.S.C. 3672). Similar authorization to investigate the work of federal pretrial services rests under U.S.C. § 3153(c)(2).[2] In order to meet its statutory responsibilities, PPSO has relied in large part on its office reviews, which are cyclical on-site, broad examinations of an office's operations. In contrast, case reviews are conducted on an

[2] U.S.C. § 3153(c)(2) states that the Director of the Administrative Office of the United States Courts is authorized to issue regulations governing the release of information made confidential by 18 U.S.C. § 3153(c)(1), enacted by the Pretrial Services Act of 1982. Within these regulations, pretrial services information shall be available to the staff of the AOUSC for reviews, technical assistance, or other research related to the administration of justice.

ad hoc basis, usually looking into the supervision of an individual defendant or offender implicated in new serious criminal conduct, such as a murder or rape (Whetzel & Sheil, 2015). The number of these reviews is limited due to resource constraints. Additionally, the probation and pretrial services offices conduct their own investigations related to new criminal conduct by persons under supervision, but the scope is generally focused on the offender and not the system as a whole. Considering that in a ten-year period, from fiscal year 2005 through fiscal year 2014, there were roughly 4,000 homicides, sexual assaults, robberies, and felonious assaults committed by offenders on federal supervision,[3] the federal probation and pretrial services system could learn a considerable amount from examining more of these situations using a systematic, structured, and objective review process.

Over the last several decades evidence-based practices have taken hold in correctional systems around the country. While risk assessment has been used to identify persons on supervision who are at greater likelihood of committing an offense specifically, very little has been done to develop *systems and processes* that are keyed to reduce the risk of such an event. The Sentinel Events review process, if modeled on the same process in the medical system, promises to help us begin to understand what organizational deficiencies are occurring leading to violent offending.

As a system, we recognize that there will be mistakes, oversights, and problems despite having very capable staff; missteps by any person involved in a case are inevitable, especially in a higher-risk organization (Perrow, 1999). It takes strength of character and investment in the system to do the self-analysis required to answer the hard questions. Maybe what looked at first glance like great supervision had hidden flaws, maybe assumptions were made, or practice drift occurred? A SE review may help to draw out the systemic flaws. For example, during a post-incident case review, the review team may find out that the officer was supervising a person at a lower risk level, because the risk assessment tool was scored incorrectly. It would be easy to focus the blame on the officer's mistake. Taking a system's analysis approach would move the review beyond the officer by asking a series of "why" questions:

- If the assessment was scored wrong, why? Maybe the officer assumed he was scoring it correctly because he passed the recertification and did not feel the need to reference the scoring guide.
- Why didn't the supervisor catch it? Maybe the officer and the supervisor both were tasked with too high a caseload or too many other responsibilities.
- Did the supervisor communicate any barriers to conducting his or her work to the deputy chief? If so, did the deputy chief address the concerns?
- Was the district emphasizing the importance of risk assessment accuracy as the foundation of supervision?
- Was the national policy and training sufficient?

The potential outcome of this questioning style focuses on the agency instead of just the officer and maybe the supervisor. Officers will have the opportunity to explain—without seeming to excuse—a decision, evoking a more collaborative, "flatter," and less hierarchical approach.

The big question is how to conduct a sentinel event analysis and still hold staff accountable for performance issues. In the article *Balancing "No Blame" with Accountability in Patient Safety* (2009), the authors discuss how hand hygiene rates in the medical field barely rose past 70 percent despite aggressive efforts to change hospital practices, including policy changes, training, hand-gel dispensers in or near every patient's room, financial incentives, etc., to increase rates. The article suggests it may be easy to overlook the 30 percent as reasonable people occasionally making mistakes. However, if after system improvements are in place an individual continues to bypass the practice, negative consequences should be implemented. And of course, as James Reason acknowledges, every industry has transgressions that require discipline (1997). The idea is to create an environment where employees understand that if something happens, the leaders will look at the entire process, understanding that no one act would have been the sufficient cause of the negative outcome. Staffs also need to understand that as part of that process, they will be held accountable for their actions, especially if they have been provided with clear performance expectations or the action was egregious or deliberate. Being held accountable is understandable and acceptable if the employee knows that the agency will take ownership of system failures.

The Benefits of Sentinel Event Analysis

If our system continues to limit our examining of cases to the most egregious and/or notorious events, then our ability to identify system-wide failures will be limited. This limitation will in turn limit the number of sentinel events that can be avoided. This void can be filled by expanding the current process to include a Sentinel Event review process examining more cases, but such a change would require the assistance of each probation and pretrial services office. The local offices are in a better position to see beyond the officer, beyond the case, and beyond the supervisor. A local team could collaboratively provide constructive reflection, looking for explanations and new ideas that promote continual change, capitalizing on the talents and insights of all team members and contributors.

Even if each district did one Sentinel Event review a year and provided the resulting data to the AO, the system would benefit from a plethora of useful information about the complex network of agencies, policies and practices, and decision-making leading up to these events. Subsequent analysis could determine if the events were due to shortfalls in national policy and practice. For example, results from a series of case reviews might reveal that offenders' acute risk factors were not being assessed in an ongoing, formal, and structured way and why that is happening. Further, the application of such an assessment process will likely uncover cues that can be provided to officers to let them know when an offender's risk is rising to a potentially dangerous level.

If a sentinel event/root cause analysis were conducted in the district and involved all levels of the local hierarchy, taking advantage of the insights and knowledge of office staff in a non-blaming, forward-looking manner, officers might be more willing to talk candidly about their roles and help identify areas for improvement. Inevitably, SE reviews will identify a lot of quality work. Managers can capitalize on these insights to praise officers and develop others.

From the officer's perspective, the office's adoption of the SE process can help reduce work-related pressures associated with supervising higher-risk offenders. Since 2012, federal probation managers have been adjusting caseloads to allocate more time, attention, and resources on higher-risk offenders to

[3] This number represents 2.5 percent of the total population of federal offenders entering post-conviction supervision during that same time period.

better align with the risk, needs, and responsivity principle (Cohen, Cook, & Lowenkamp, 2016). Probation officers around the country are beginning to express increased stress levels as a result. In the article "'It's relentless': The impact of working primarily with high-risk offenders" (2016), the authors interviewed county juvenile officers about their high-risk caseloads. One officer stated:

> …you're going to be left with domestic violence cases who are manipulative, aggressive and controlling, you're going to be left with sex offenders who just, the nature of the work can just be distressing, and violent offenders who are quite possibly going to be kind of aggressive towards you. Plus underlying all that is the terrifying thought that one of them is going to go and do something really serious and you're going to have a big case review and investigation into how good or bad you are as a probation officer.
>
> Plus, just the thought of one of your cases committing a really serious offence and harming somebody is just horrible.

In the current federal probation and pretrial services review process, talented officers who have done exceptional work with a defendant or offender may feel as though they are being attacked; the process makes them feel like a "second victim" (Dekker, 2015). In an interview with an officer after an AO case review on one of his supervisees, he said the whole process felt like *he* was under investigation. Although the AO administrators explained that the process was intended to bring about improvement, he was nervous that he had missed something, even though he felt like he had really worked hard with the person from day one. He was worried that because of public and political pressure, he was going to be the scapegoat, so he was reluctant to expand upon his answers. It was a very stressful time and made him rethink why he wanted to be a probation officer.

Likewise, probation and pretrial services officers who supervise high-risk cases that have not been under the limelight are feeling the pressures of the *potential* for media attention on their performance, because they hear about situations from colleagues across the country. Chiefs are reporting that it is difficult to convince officers to apply for promotional opportunities. Making the move to a Sentinel Event process can help reduce these types of pressures. As stated, it is bad enough knowing your case could cause serious harm, without the stress of a "big case review and investigation" that feels like someone is looking for a scapegoat. Reducing the fear of misdirected consequences related to making occasional human mistakes allows the officer to focus more energy on working with the individuals under supervision. Additionally, potential applicants may be encouraged to work for an organization that is viewed as a progressive, learning organization (Senge, 2006).

Conducting SE reviews will likely build future leaders who have the desire to ask the hard questions, delving deep into the interrelated operational and administrative actions of the office that contribute to the success (or failure) of individuals involved in the justice system. Gaining these types of skills is huge for a system that struggles, along with the rest of the civilian federal government workforce, with the retirement of large numbers of experienced employees (General Accounting Office, 2014). According to the AO personnel data, in the next five years, 45 percent of chief probation and pretrial services officers, 33 percent of deputy chiefs, and 33 percent of supervisors will be retiring, leaving a significant need for opportunities to build capable leaders—leaders willing to accept feedback at all levels, providing a safe, trusting environment that encourages officers to talk about deficiencies and offer suggestions for strategic improvements that align with the agency's mission to become outcome-driven.

From a national perspective, the contributions of information from districts on just a handful of SE reviews would provide a unique view into the interworkings of probation and pretrial offices in relation to the entire system. This concept of learning from situations is not new. Researchers are acknowledging that just looking at the data points without the human element shows an incomplete story. In fact, some have begun discussing the limitations of big data and have introduced the term "thick data." Wang (2013) describes these two concepts this way:

> Big Data reveals insights with a particular range of data points, while Thick Data reveals the social context of and connections between data points. Big Data delivers numbers; thick data delivers stories. Big data relies on machine learning; thick data relies on human learning.

Or perhaps in a more familiar context, Ulmer (2012), discussing the state of the research and new directions in sentencing research, stated:

> As the discussions of recent literature and desirable new research directions show, the study of sentencing in the past decade has been highly focused on quantitative measurement and modeling. As I said earlier, this is not a problem in itself. However, if we do not match that focus on modeling with a parallel focus on the in situ decisions and activities of courtroom workgroup participants, and how these are shaped by their surrounding court community contexts, our understanding of sentencing will be truncated.

Both Wang and Ulmer are making the point that big data and quantitative studies using available datasets are limited in their ability to help us develop a true understanding of how and why events occur as they do. We would argue that Sentinel Events review would provide us with the "thick data" to supplement our big data and begin to develop a thorough and explanatory reason as to why these sentinel events occur and how to best reduce the likelihood of these events going forward.

Can We Do This?

Conceptualizing the Sentinel Event or systems analysis approach may be difficult for the U.S. probation and pretrial services system due to concerns about time pressures, legal concerns, and confidentiality, but consider the similar stakes at play in the medical, aviation, energy, and transportation industries. These industries have forged the way for the past 20-plus years to provide us with an evidence-based approach that offers a substantial opportunity to learn and help us grow as a system to better help those under our charge and the community. Since probation and pretrial services offices already conduct post-incident case reviews locally, albeit not consistently and not necessarily with a systems lens, adding a non-blaming team approach on a small cohort of Sentinel Event cases may be an acceptable time commitment. The overall value of these types of reviews may far outweigh the allocation of resources.

Before the federal probation and pretrial system embarks on the Sentinel Event analysis track, input and support has to come from the chief probation and pretrial services officers and their staffs to take advantage of this learning opportunity. If chiefs engage in the process, it has to be with interest and commitment to help protect the community and improve our work, not just because the AO is asking. A working group is the logical venue to establish short- and long-range strategic goals for engaging in this process. The group will be charged with tasks such as exploring the research, defining a sentinel event, and making recommendations for a path forward.

Conclusion

We have a choice to work together on a shared goal to improve the U.S. probation and pretrial services system at all levels, capitalizing on less than optimal situations. By getting away from the "single-minded focus," we can draw out insights from all layers of the organization. Jeffrey Thomason, chief of the Idaho U.S. probation office, has experience with these types of review both in and outside of the federal probation arena and sums it up well.

> In the probation system, we tend to look at failure from the perspective of the failed. The high-risk individual who revokes with a new offense inside of a year on supervision is performing to type and may not raise an eyebrow. However, when that new offense causes significant damage and results in attention both from within and outside the organization, our tendency is to circle the wagons. Across our system, we have a large enough number of these cases in the aggregate that conducting a robust post-incident review has the potential to greatly improve our case management, and hopefully, prevent even one of these cases from occurring in the future.

The idea is to have a coordinated effort to learn as much as we can in the interest of improving the system, the experience for the person under supervision, and most of all, the community. With the chiefs at the helm, encouraging a synergistic, action-oriented process, the U.S. probation and pretrial services system can prepare for the future of corrections.

References

Berwick, D. (1989). Continuous improvement as an ideal in healthcare. *New England Journal of Medicine*, 320, 53-54.

Cohen, T., Cook, D., & Lowenkamp, C. (2016). The supervision of low-risk federal offenders: How the low-risk policy has changed federal supervision practices without compromising community safety. *Federal Probation*, *80*(1), 3-11.

Dekker, S. (2007). *Just culture: Balancing safety and accountability*. Farnham, UK: Ashgate Publishing.

Dekker, S., & Breakey, H. (2016). 'Just culture:' Improving safety by achieving substantive procedural and restorative justice. Brisbane, Australia. Griffith University, *Safety Services*, 85, 187-193.

General Accounting Office. (2014). Federal workforce: Recent trends in federal civilian employment and compensation, Washington, D.C. Retrieved from http://www.gao.gov/products/GAO-14-215

The Joint Commission. (2016). CAMH Update 2. Oakbrook Terrace, IL. Retrieved from: https://www.jointcommission.org/assets/1/6/CAMH_24_SE_all_CURRENT.pdf.

Leape, L. (1994). Error in medicine. *Journal of the American Medical Association,* 272, 1851-1857.

National Institute of Justice (NIJ). (2014). *Special report: Mending justice: Sentinel Event reviews*. Washington, D.C.

National Patient Safety Foundation. (2016). *RCA2: Improving root cause analyses and actions to prevent harm*, V.2. Boston, MA. Retrieved from http://www.npsf.org/?page=RCA2

National Transportation Safety Board (2016), The investigative process. Retrieved from http://www.ntsb.gov/investigations/process/Pages/default.aspx.

Perrow, C. (1999). *Normal accidents*. Princeton, N.J.: Princeton University Press.

Phillips, J., Westaby, C., & Fowler, A. (2016). 'It's relentless': The impact of working primarily with high-risk offenders. *Probation Journal*, *63*(2), 182-192.

Reason, J. (1997). Engineering a just culture. In *Managing the risks of organizational accidents.* Hampshire, United Kingdom: Ashgate, 205-12.

Reason, J. (2000). Human error: Models and management. *The BMJ, 320*, 768–70.

Senge, P. M. (1990, revised 2006) *The fifth discipline: The art & practice of the learning organization*. New York: Doubleday.

Sharpe, V. A. (2003). Promoting patient safety: An ethical basis for policy deliberation. *Hastings Center Report Special Supplement*, *33*(5), July/August.

Snook, S. A. (2002). *Friendly fire: The accidental shootdown of U.S. Black Hawks over northern Iraq.* Princeton, New Jersey: Princeton University Press.

Starr, D. (2015). A new way to reform the judicial system. The New Yorker (online edition). Retrieved from: http://www.newyorker.com/news/news-desk/the-root-of-the-problem .

Wachter, R. M., & Pronovost, P. J. (2009). Balancing "no blame" with accountability in patient safety. *The New England Journal of Medicine*, 361, 1401-1406.

Wang, T. (2013). Big Data needs Thick Data. *Ethnography Matters* (online). N.P., Retrieved from http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data.

Whetzel, J., & Sheil, J. (2015). Accountability and collaboration: Strengthening our system through office reviews. *Federal Probation*, *79*(3), 9-13.

Ulmer, J. T. (2012). "Recent developments and new directions in sentencing research." *Justice Quarterly, 29*(1), 1-40.

False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."

Anthony W. Flores
California State University, Bakersfield
Kristin Bechtel
Crime and Justice Institute at CRJ
Christopher T. Lowenkamp
Administrative Office of the United States Courts
Probation and Pretrial Services Office

The validity and intellectual honesty of conducting and reporting analysis are critical, since the ramifications of published data, accurate or misleading, may have consequences for years to come.
—Marco and Larkin, 2000, p. 692

PROPUBLICA RECENTLY RELEASED a much-heralded investigative report claiming that a risk assessment tool (known as the COMPAS) used in criminal justice is biased against black defendants.[12] The report heavily implied that such bias is inherent in all actuarial risk assessment instruments (ARAIs).

We think ProPublica's report was based on faulty statistics and data analysis, and that the report failed to show that the COMPAS itself is racially biased, let alone that other risk instruments are biased. Not only do ProPublica's results contradict several comprehensive existing studies concluding that actuarial risk can be predicted free of racial and/or gender bias, a correct analysis of the underlying data (which we provide below) sharply undermines ProPublica's approach.

Our reasons for writing are simple. It might be that the existing justice system is biased against poor minorities due to a wide variety of reasons (including economic factors, policing patterns, prosecutorial behavior, and judicial biases), and therefore, regardless of the degree of bias, risk assessment tools informed by objective data can help *reduce* racial bias from its current level. It would be a shame if policymakers mistakenly thought that risk assessment tools were somehow worse than the status quo. Because we are at a time in history when there appears to be bipartisan political support for criminal justice reform, one poorly executed study that makes such absolute claims of bias should not go unchallenged. The gravity of this study's erroneous conclusions is exacerbated by the large-market outlet in which it was published (ProPublica).

Before we expand further into our criticisms of the ProPublica piece, we describe some context and characteristics of the American criminal justice system and risk assessments.

Mass Incarceration and ARAIs

The United States is clearly the worldwide leader in imprisonment. The prison population in the United States has declined by small percentages in recent years and at year-end 2014 the prison population was the smallest it had been since 2004. Yet, we still incarcerated 1,561,500 individuals in federal and state correctional facilities (Carson, 2015). By sheer numbers, or rates per 100,000 inhabitants, the United States incarcerates more people than just about any country in the world that reports reliable incarceration statistics (Wagner & Walsh, 2016).

Further, it appears that there is a fair amount of racial disproportion when comparing the composition of the general population with the composition of the prison population. The 2014 United States Census population projection estimates that, across the U.S., the racial breakdown of the 318 million residents comprised 62.1 percent white, 13.2 percent black or African American, and 17.4 percent Hispanic. In comparison, 37 percent of the prison population was categorized as black, 32 percent was categorized as white, and 22 percent as Hispanic (Carson, 2015). Carson (2015:15) states that, "As a percentage of residents of all ages at yearend 2014, 2.7 percent of black males (or 2,724 per 100,000 black male residents) and 1.1 percent of Hispanic males (1,090 per 100,000 Hispanic males) were serving sentences of at least 1 year in prison, compared to less than 0.5 percent of white males (465 per 100,000 white male residents)."

Aside from the negative effects caused by

[1] The authors wish to thank James Bonta, Francis Cullen, Edward Latessa, John Monahan, Ralph Serin, Jennifer Skeem, and Stuart Buck for their thoughtful comments and suggestions.

[2] The main article and an accompanying analysis report were authored by the same individuals, albeit with a different order of authorship. The main ProPublica article is cited as Angwin, Larson, Mattu, and Kirchner (2016) or Angwin et al. (2016). The analysis report is cited as Larson et al. (2016).

imprisonment, there is a massive financial cost that extends beyond official correctional budgets. A recent report by The Vera Institute of Justice (Henrichson & Delaney, 2012) indicated that the cost of prison operations (including such things as pension and insurance contributions, capital costs, legal fees, and administrative fees) in 40 states participating in their study was 39.5 billion (with a b) dollars per year. The financial and human costs, and perhaps absurdity, of these practices have become so obvious that there has been bipartisan support for efforts to develop solutions to reduce the amount of money spent on incarceration and the number of lives negatively impacted by incarceration (Skeem & Lowenkamp, 2016b).

An example of one such effort has been the investigation of the use of ARAIs to partially inform decisions related to sentencing and other correctional decisions. Whether it is appropriate to use ARAIs in criminal justice settings is a popular debate. However, as Imrey and Dawid (2015:18)[3] note, the debates and "… considerations [of using ARAIs in such settings] are properly functions of social policy, not statistical inference." That is, there might be much to debate about how and why we would *use* valid ARAIs. The issue that is no longer up for debate is that ARAIs predict outcomes more strongly and accurately than professional judgment alone. Several studies and meta-analyses have reached similar conclusions indicating that actuarial risk assessments are superior to unstructured professional judgment in terms of predicting the likelihood of both general recidivism and even specific recidivism outcomes (Grove, Zald, Lebow, Snitz, & Nelson, 2000), including future sex offending (Hanson & Morton-Bourgon, 2009). Noteworthy research on the predictive accuracy of risk assessments can be attributed to Meehl (1954) and Grove et al. (2000), including the oft-cited and comprehensive review of risk assessments from Andrews, Bonta, and Wormith (2006).

Given that this research often goes unrecognized by those concluding that ARAIs cannot be relied upon to predict outcomes, it is relevant to clarify what the potential consequences are for ignoring (presumably unintentionally) a vast body of research on the performance of ARAIs. Specifically, the implications could be as serious as dismissing the use of risk assessments outright. This type of abrupt response and return to subjective judgment would be unethical, and one poorly informed statement should not replace over 60 years of research in which consistent findings are produced in support of ARAIs.

ARAIs are intended to inform objective decision-making, so proper administration of the instrument and clear guidance on what information risk assessments are capable of reliably providing for a target population are relevant points of discussion. What is equally important is that the development of these tools be rigorous and that subsequent tests of their performance in predicting recidivism include independent evaluations. Finally, critiques of risk assessments, including questions about racial bias, should be properly conducted and described. Thankfully, there are empirical standards for testing whether assessments are biased—standards that were not discussed or applied in the ProPublica pieces.

One of the more common concerns that arise in the discourse on the use of risk assessment in correctional and sentencing contexts is racial bias. Given the racial disproportionality already seen in prison populations (and at other points in the criminal justice process), racial bias is a salient issue for the use of ARAIs or any other method to structure decision-making. But concerns that the use of ARAIs would increase racial disproportionality were drawn from *hypothetical* or *theoretical* linkages and limited empirical evidence between certain risk factors and race (see Skeem & Lowenkamp, 2016a). It is unfortunate that this concern—race-based bias in risk assessment—is threatening to stall sentencing and correctional reform, especially when it is likely, given the racial disproportionality in the correctional system, that minorities could benefit most from unwinding mass incarceration. Still, these concerns over bias *are legitimate*. At the same time, these concerns can and should be *properly* investigated.

In their attempt to investigate test bias of the Northpointe COMPAS across different categories of race, the ProPublica authors constructed four multivariate models. Two models predicted the likelihood that the defendant was classified as high-risk and two estimated the effect of race on the relationship between the COMPAS score and recidivism (any arrest and arrest for a violent offense). The authors conclude that

> The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.
>
> When a full range of crimes were taken into account—including misdemeanors such as driving with an expired license—the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.
>
> We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.
>
> The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants.

We appreciate that Angwin et al. (2016) made their data available for subsequent analyses by other researchers but take issue with how they analyzed the data and, consequently, their conclusions. Before we proceed further, we want to make it clear that we are not supporting or endorsing the idea of using risk assessment at sentencing (although we do support its use at certain decision points in the correctional system) nor are we advocating for the Northpointe COMPAS. We also are not making any blanket statements about race, test bias, and all ARAIs. With the previous qualifications, we present five concerns that we have with the analyses (Larson, Mattu, Kirchner, & Angwin, 2016) and the accompanying article by Angwin et al. (2016).

Criticisms of Angwin et al. (2016)

First, Angwin et al. (2016) conducted a study on a sample of pretrial defendants to determine if an instrument (the COMPAS) was biased when that instrument was not designed for use on pretrial defendants. Specifically, the COMPAS scales were developed upon and for individuals on post-disposition supervision. Further, the original sample for the ProPublica study also comprised probation and parole clients; however, Larson et al. (2016) excluded these relevant subjects from the study but failed to provide a detailed and acceptable reason for doing so. The sample they used (and shared for subsequent analysis) included only pretrial defendants, i.e., offenders who have not been convicted of the offenses for which they are being detained. This is a relevant distinction, as ARAIs that are intended to predict

[3] Also see Dawid, 2014, and Harris, Lowenkamp, & Hilton, 2015.

general and violent recidivism are typically developed and administered to probationers and parolees. Pretrial ARAIs are intended to predict different outcomes, such as failure to appear, for defendants. However, Larson et al. (2016) removed failure to appear arrests as an outcome measure for their analysis of pretrial defendants.

Additional clarification should be offered related to the COMPAS scales and their use in Broward County, Florida. The COMPAS does have a scale to examine pretrial failure outcomes, and Broward County does administer the pretrial, general recidivism, and violent recidivism scales to pretrial defendants; however, the general and violent recidivism scales are only appropriate for those on post-disposition supervision, when recidivism data would be collected within a specified time frame. The COMPAS validation study that the ProPublica authors cite to justify their definition and interpretation of their measures of recidivism (i.e., Brennan, Dieterich, & Ehret, 2009, p. 25) actually indicates that the COMPAS recidivism scales are intended to predict new offenses with probationer samples. There is no mention that the COMPAS recidivism scales are intended to predict recidivism for pretrial defendants (See page 25 from Brennan, Dieterich, & Ehret, 2009). Note, the purpose of the current study is not to address Broward County's use of the COMPAS scales with pretrial defendants, but we would strongly urge that examinations into the performance of an ARAI begin with a solid understanding of the tool's purpose, target population, and intended outcome(s).

Second, the authors force a dichotomy on the COMPAS. The COMPAS was not made to make absolute predictions about success or failure. Instead, it was designed to inform probabilities of reoffending across three categories of risk (low, medium, and high). Further, in their false positive/false negative analysis the authors collapsed all the moderate and high-risk defendants in the "high" category. The standard for this is to put all the moderate and high-risk defendants in a category and then reverse that and put low and moderate into a collapsed "low" category to observe if there are statistical changes as a result. See Singh (2013) for a methodological primer regarding performance indicators for ARAIs.

Third, the authors equate racial differences in mean scores on a risk assessment instrument (which would be highlighted by their model referenced in number 2 above) with test bias. This is not true—not true at all. See the Standards for Educational and Psychological Testing (discussed in the following point).

Fourth, well-established and accepted standards exist to test for bias in risk assessment. Larson et al. (2016) and Angwin et al. (2016) do not mention—or appear to be aware—that such standards exist. The analysis conducted in the ProPublica article fails to actually test for bias within these standards, which is critical given that this is the main focus of the report. Skeem and Lowenkamp (2016a) cover this issue extensively in their evaluation of the federal Post Conviction Risk Assessment (PCRA) and properly test for predictive bias within the guidelines from Standards for Educational and Psychological Testing. (For more information, see American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Fifth, Larson et al. (2016) overstate the effect of their results and fail to cite the limitations of their study. It is well known—and commonly taught in introductory statistics courses—that even trivial differences can attain statistical significance in large sample sizes. To address this, researchers have several options to select from, including pulling a random but smaller sample of cases from the original sample to conduct the analysis or setting the values to test for significance higher (e.g., $p \leq 05$, $p \leq .01$, $p \leq .001$) as sample size increases. Larson et al. (2016) take the opposite approach: Even though the interaction terms in their two Cox Regression models do not reach statistical significance *by conventional standards* applied with relatively small samples, the authors interpret a significant difference when $p = 0.0578$. Larson et al. (2016) should have considered, with a sample of over 10,000 defendants, a more appropriate significance value (p) of .001. A preferable option would be to focus on effect sizes (with confidence intervals), which convey how large and meaningful a difference is, rather than merely whether it reaches "statistical significance."

We would like to explore one final thought in this section. Some readers might be wondering why anyone should care about our concerns. Discussions about ARAIs, statistics, methods, and test bias may seem complex and uninteresting (we find them rather fascinating). We are at a unique time in history. We are being presented with the chance of a generation—and perhaps a lifetime—to reform sentencing and unwind mass incarceration in a scientific way, and that opportunity is slipping away because of misinformation and misunderstanding about ARAIs. Poorly conducted research or misleading statements can lead to confusion and/or paralysis for those charged with making policy. The quote from a subsequent ProPublica article makes this point (Kirchner, 2016). Relying on the research of Angwin et al. (2016), the chair of the federal defenders legislative committee, David Patton, when being interviewed by Lauren Kirchner, posed the question "Will it be possible to validly measure those things [risk factors] for somebody who is institutionalized?" and stated "We just don't know that such a tool can be developed, or if it can, whether it will exhibit similar racial biases of current tools." In response to these issues, we analyzed a reduced set of data used by Larson et al. (2016); based on our findings, we conclude that the Larson et al. (2016) analysis was misguided and the subsequent conclusions offered by Angwin et al. (2016) are faulty. Below is a description of the methods employed to test for race-based bias with the COMPAS.

Methods

To properly test the COMPAS for race-based bias, we downloaded the dataset comprising only the sample of pretrial defendants (as the probation and parolee data were excluded) and syntax that Larson et al. (2016) used in their analyses. Two separate files were available for analysis. One file contained the information needed to test the relationship between the Northpointe COMPAS and arrest for any crime. The second file contained the information needed to test the relationship between the Northpointe COMPAS and arrest for a violent crime. We made all variable transformations in R using the same syntax as Larson et al. (2016).

We departed from their analysis in the following ways: First, we kept for analysis only those defendants whose race was either black or white. This was done as Larson et al. (2016) only mention bias between black and white defendants and doing so simplifies the analysis and subsequent discussion. This process reduced our sample sizes to 5,278 for the "any arrest" analysis file and 3,967 for the "arrest for a violent offense" analysis file with a two-year follow-up to measure recidivism.

Second, rather than analyze group mean differences to determine if bias exists, we used a framework that tests for bias in the degree of prediction as a function of race and functional form of prediction (i.e., slope and intercept) as a function of race. This framework is based on methods of testing for bias developed, recognized, and used in other professions. The framework is reviewed and applied to a risk assessment by Skeem and Lowenkamp (2016a, 2016b) and Skeem, Monahan, and Lowenkamp (2016).

Third, the Northpointe COMPAS decile score was used in all analyses rather than the category ratings (low, medium, high). It should be noted that the results of the analyses were not dependent on the scale of the risk score. The same results were obtained when we used the decile score or the category ratings, and the decile scores provided a more refined or precise estimate than the risk categories (e.g., low, medium, high).

To test for bias in the degree of prediction as a function of race, we calculated AUC-ROC values for the overall sample and for each race.[4] The AUC-ROC values for white and black defendants were then compared using z-tests. We calculated and analyzed AUC-ROC values using any arrest as the outcome measure and then using arrest for a violent offense as the outcome measure.

To test for bias in form as a function of race, we calculated a series of logistic regression models predicting each of the outcomes (first for any arrest and then for arrest for a violent offense). To test for bias in form, we inspected interaction terms between the race and the Northpointe decile score for each outcome of interest. In addition, the magnitude and statistical significance of the coefficient for race was inspected for each outcome.

Results

Initial analyses involved an examination of general recidivism base rates for the sample and then across race. Results of these analyses are presented in Table 1, which shows the base rate of failure (general rearrest) as 47 percent for all defendants, 39 percent for White defendants, and 52 percent for Black defendants. It is important to note that the general recidivism base rate for Black defendants is significantly higher than it is for White defendants specifically, and the overall sample generally. Racial differences in failure rates across race describe the behavior of defendants and the criminal justice system, not assessment bias. Results also indicate that failure rates seem to monotonically increase with the risk categorizations of the COMPAS in that 29-35 percent of low-risk defendants were rearrested (White and Black respectively), 53-56 percent of medium-risk defendants were rearrested (respectively), and 73-75 percent of high-risk defendants were rearrested (also respectively). Note here that while the base rate of general recidivism differed significantly for White and Black arrestees (with Black defendants evidencing higher rearrest rates), the general recidivism failure rates for White and Black defendants are somewhat similar across low-, medium-, and high-risk categorizations.

To explore the predictive fairness of the COMPAS, we first examined whether the degree of the relationship between COMPAS scores and general recidivism varied due to race. Analyses of the degree of accuracy involved AUC-ROC analyses, which are appropriate for accomplishing this goal because they identify the chance (or probability) that a randomly selected arrestee will have a higher COMPAS score than will a randomly selected non-arrestee. AUC-ROC values range from zero to one, with .5 indicating mere chance prediction (or “fifty-fifty”), 1 indicating perfect prediction, and AUC-ROC values of .56, .64, and .71 signifying small, medium, and large predictive benchmark effects, respectively (Rice & Harris, 2005). As an interpretive example, an AUC-ROC value of .71 would translate to a randomly selected arrestee scoring higher on the COMPAS than would a randomly selected non-arrestee 71 percent of the time. If the COMPAS is differentially accurate in its degree of recidivism prediction across race, corresponding z-tests derived from AUC-ROC values for White and Black arrestees will be significantly different from one another. The following analyses are those that comport with accepted standards for determining if a particular test is biased against a particular group.

Degree of Relationship

In accordance with standard practices in testing for bias on education and psychological tests, the AUC-ROC values were generated and compared for the entire sample and for each group of race. AUC-ROC analyses presented in Table 1 show a moderate to strong degree of predictive accuracy for all defendants, as well as across defendant race. The COMPAS demonstrated a strong degree of accuracy in prediction for all defendants, with an AUC of .71. The AUC estimate for White defendants was .69 and .70 for Black defendants, with no significant difference between values by race. This simple lack of difference in predictive utility for the COMPAS by race contradicts the conclusions reached by Larson et al. (2016).

Table 1 also presents DIF-R values for the sample and across race to investigate the dispersion of recidivism base rates across risk categorizations of the COMPAS (as opposed to COMPAS decile score accuracy, which was examined above using AUC-ROC analyses). The values of the dispersion index for risk (or DIF-R) range from one to infinity, with larger values indicating greater accuracy, across and within each risk category as a function of base rate dispersion (Silver, Smith, & Banks, 2000). Results of the DIF-R analyses support the COMPAS risk categorizations as unique from one another and meaningful. The calculated DIF-R values in Table 1 are consistent with those found in other risk assessment studies.

TABLE 1.
Failure Rates, AUC-ROC, DIF-R for General Recidivism

	All	White	Black
Low	32	29	35
Medium	55	53	56
High	75	73	75
Base Rate*	47	39	52
AUC	0.71	0.69	0.70
DIF-R	0.73	0.65	0.70

*= $\chi^2(2) = 88.85$; $p < 0.001$

Table 2 shows the degree of prediction for the COMPAS and violent recidivism. Analyses performed were identical to those just presented above in Table 1, save for the different outcome. Failure rates for violent recidivism were 17 percent for the sample, 12 percent for White defendants, and 21 percent for Black defendants. Violent recidivism failure rates across risk categories increased with risk categorization successively, although Black defendants were arrested for a violent offense at a much higher rate than White defendants across all three categories of risk. Again, note that different (higher) violent arrest rates for Black defendants than White defendants is not an indicator of assessment bias. As noted above for general recidivism in Table 1, AUC-ROC analyses show moderate to strong and statistically similar predictive accuracy for both Black and White defendants. Further, DIF-R values for violent arrest evidence acceptable base-rate dispersion for the sample and across race, with slightly better

[4] We chose AUC-ROC as it is recognized as a standard measure in assessing diagnostic accuracy of risk assessments and has properties that make it not affected by base rate or sample size (Rice & Harris, 2005).

TABLE 2.
Failure Rates, AUC-ROC, DIF-R for Violent Recidivism

	All	White	Black
Low	11	9	13
Medium	26	22	27
High	45	38	47
Base Rate*	17	12	21
AUC	0.71	0.68	0.70
DIF-R	0.63	0.47	0.64

*= $\chi^2(2) = 49.41$; $p < 0.001$

risk category dispersion for Black defendants.

The above examination of failure rates, degree of predictive accuracy, and base rate dispersion across race fails to support the conclusions of racial bias made by Angwin et al. (2016) and, instead, finds a degree of prediction that is remarkably consistent for both Black and White defendants.

We made the argument above that Angwin et al.'s false positive/false negative analysis of the COMPAS was flawed and present a reanalysis in Tables 3 and 4. When dealing with a risk assessment that provides more than two categories, it is recommended that tests based on a 2×2 contingency table (e.g., positive predictive value, negative predictive value, false positive rate, false negative rate) be run using a specific binning strategy. That is, a decision has to be made on how to create two groups from a three (or more) category risk assessment. Singh et al. (2011) recommend first binning the low cases as the "low-risk group" and comparing them to the moderate and high-risk offenders binned together as the "high-risk group." This would be considered a "rule-in" test. The second binning process involves combining the low and moderate-risk offenders into the "low-risk group" and comparing them to the high-risk offenders (high-risk group). This would be considered a "rule-out" test.

When this process is followed, note that the false positive rates decrease substantially when binning the low and moderate risk cases together and treating them as the "low-risk" group (or the group that would be expected to succeed). In contrast, false negative rates go up in both groups. These two reversals—a decrease in false positive rates and an increase in false negative rates—might be preferred by some, as it limits the number of individuals that are identified as "high-risk." For others with a low tolerance for recidivism and victimization, the binning process where moderate and high-risk were combined to form the "high-risk" group would be preferred. Regardless, what should be taken away from these tables is the fact that when recommended practices are followed for calculating performance indicators of predictive instruments, a somewhat different pattern of results and conclusions is drawn.

Form of Relationship

To further investigate Angwin et al.'s rather serious claims of racial bias, subsequent analyses, suggested by accepted testing standards, center on the form of the relationship between recidivism and COMPAS score. More specifically, if the algorithm upon which the COMPAS is based was to perform similarly across race, then the mathematical regression slope and intercept of that relationship should also be similar across racial subgroups (Aguinis, Culpepper, & Pierce, 2010). Put more simply, we are examining the functional form (slope and intercept) of the relationship between the COMPAS and recidivism to see whether an average COMPAS decile score of x corresponds to an average arrest rate of y across race, which is the standard for examining predictive bias.

To investigate the form of the relationship between the COMPAS and recidivism across race, we estimated four logistic regression models for each of the two outcomes (general and violent recidivism) that were then compared to determine whether slope and intercept differences exist between White and Black defendants. Table 5 presents the results of these analyses, showing that Model One predicts arrest with age, gender, and

TABLE 3.
Performance Indicators Low vs. Moderate/High

White				Black			
		Actual				Actual	
		NR	R			NR	R
Predicted	NR	999	408	Predicted	NR	873	473
	R	282	414		R	641	1188
FN	0.50			FN	0.28		
FP	0.22			FP	0.42		
Sensitivity	0.50			Sensitivity	0.72		
Specificity	0.78			Specificity	0.58		
PPV	0.59			PPV	0.65		
NPV	0.71			NPV	0.65		

FN = False negative rate; FP = False positive rate; PPV = Positive predictive value; NPV = Negative predictive value; NR = Not recidivist; R = Recidivist

TABLE 4.
Performance Indicators Low/Moderate vs. High

White				Black			
		Actual				Actual	
		NR	R			NR	R
Predicted	NR	1220	660	Predicted	NR	1303	1027
	R	61	162		R	211	634
FN	0.80			FN	0.62		
FP	0.05			FP	0.14		
Sensitivity	0.20			Sensitivity	0.38		
Specificity	0.95			Specificity	0.86		
PPV	0.73			PPV	0.75		
NPV	0.65			NPV	0.56		

FN = False negative rate; FP = False positive rate; PPV = Positive predictive value; NPV = Negative predictive value; NR = Not recidivist; R = Recidivist

TABLE 5.
Logistic Regression Models Predicting Two-Year General Recidivism (N = 5278)

	Model 1	Model 2	Model 3	Model 4
Age	0.97*	0.98*	0.99*	0.99*
Female	0.58*	0.60*	0.61*	0.61*
Black	1.45*	--	1.09	1.12
NPC Decile	--	1.30*	1.30*	1.30*
NPC Decile X Black	--	--	--	0.99
Constant	2.29*	0.42*	0.40*	0.39*
Chi Square	297.68	804.42	806.13	806.19
LL	-3500.37	-3247.00	-3246.14	-3246.11
Pseudo-R^2	0.04	0.10	0.11	0.11

Note: The two dashes '- -' in the table above indicate that the variable was not included in the model.

TABLE 6.
Logistic Regression Models Predicting Two-Year Violent Recidivism (N = 3967)

	Model 1	Model 2	Model 3	Model 4
Age	0.96*	0.99	1.00	1.00
Female	0.47*	0.57*	0.57*	0.57
Black	1.57*	--	1.24	1.21
NPC Decile	--	1.32*	1.30*	1.30*
NPC Decile X Black	--	--	--	1.01
Constant	0.66	0.09*	0.08*	0.08*
Chi Square	183.53	345.34	350.49	350.52
LL	-1725.6	-1644.70	-1642.12	-1642.11
Pseudo-R^2	0.05	0.09	0.10	0.10

Note: The two dashes '- -' in the table above indicate that the variable was not included in the model.

race; Model Two predicts arrest with age, gender, and COMPAS decile score; Model Three predicts arrest with age, gender, race, and COMPAS decile score; and Model Four predicts arrest with all of the above variables, including an interaction term for race and COMPAS decile score.

Comparisons across these four models presented in Table 5 reveal two important findings relevant to an investigation of racial bias in assessment. First, an examination of Models Three and Four indicates that the addition of the interaction term between the COMPAS and race is not significant and does not improve the prediction of general recidivism for the model overall. So, the slope of the relationship between the COMPAS and general recidivism is similar for both Black and White defendants, and race does not moderate the utility of the COMPAS to predict general recidivism. Second, a comparison of Models Two and Three shows that there are no significant racial differences in the intercept (or constant) for the relationship between the COMPAS and general recidivism. Taken together, these findings suggest that there are no significant differences in the functional form of the relationship between the COMPAS and general recidivism for White and Black defendants. A given COMPAS score translates into roughly the same likelihood of recidivism, whether a defendant is Black or White.

Similar analyses were conducted for the relationship between the COMPAS and violent recidivism and these results are presented in Table 6. As above, there is no observed evidence of assessment bias in these analyses. Specifically, the relationship between race and violent recidivism becomes insignificant once the COMPAS decile score is introduced into the logistic equation. Furthermore, the interaction term between race and COMPAS decile score in Model Four is also insignificant. As above, these findings indicate no difference in the form of the relationship between the COMPAS and violent recidivism for White and Black defendants.

As a final analysis of predictive fairness by race for the COMPAS, we calculated predicted probabilities of any arrest (general recidivism) based on regression Model Four in Table 5, grouped together those probabilities for each COMPAS decile score, and then displayed the grouped probabilities across race in Figure 1. Examination of this figure shows that the slope of the relationship between the COMPAS and general recidivism does not differ by race, although Black defendants do have higher predicted (and observed) arrest rates. Similarly, we then calculated the predicted probabilities of violent arrest based on Model Four of Table 6, grouped together those probabilities for each COMPAS decile score, and then displayed the grouped probabilities across race in Figure 2. As was observed in Figure 1, the slope of the relationship between COMPAS score and violent arrest does not differ across race (again, although Black defendants have higher predicted violent arrest rates). Taken together, these two figures further support parity in the form of the relationship between the COMPAS and rearrest (general and violent).

Finally, Figures 3 and 4 visually summarize this study's findings. The bar chart in Figure 3 shows recidivism rates for any arrest by COMPAS risk category (low, medium, and high) and across race. The figure also displays a graphed line showing the percentage of Black defendants in each risk category. The graphed line shows that the percentage of Black defendants increases along with risk categorization, meaning there are more high-risk Black defendants than there are medium risk, and more medium-risk Black defendants than there are low risk. Overall, this means that Black defendants tend to score higher on the COMPAS than White defendants. Alone, this might suggest bias. However, examination of the bar chart shows that subsequent arrest rates increase along with risk categorization for both White and Black defendants and that Black defendants have higher recidivism rates than White defendants across all three

FIGURE 1.
Predicted Probability of Any Arrest by Race

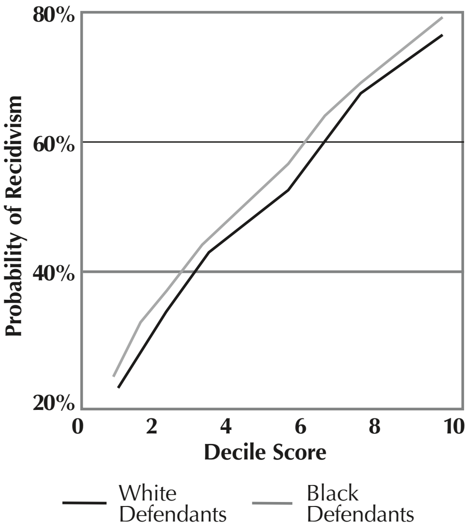
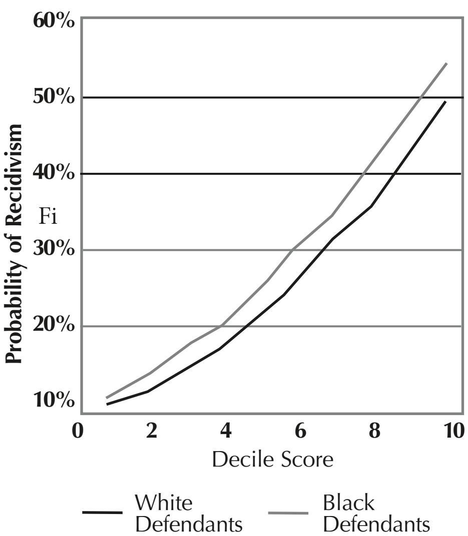


FIGURE 2.
Predicted Probability of Arrest for Violent Offense by Race



categories of risk.

Taken together, the two aspects of this figure show us that, despite the conclusions of Angwin et al. (2016), racial differences in mean risk scores are less indicative of test bias than of true differences in the likelihood of recidivism. The same pattern of findings also holds for violent arrest shown in Figure 4.

Discussion

A recent ProPublica.org article by Angwin et al. (2016) investigated the presence of racial bias in one of the more popular and commonly used

FIGURE 3.
Recidivism Rates by Race and Percent Black in Each Risk Category—Any Arrest

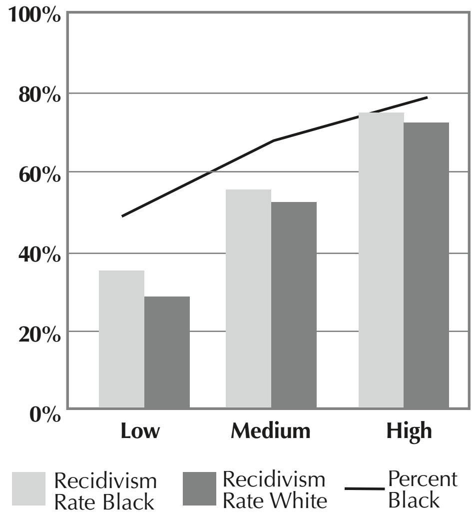
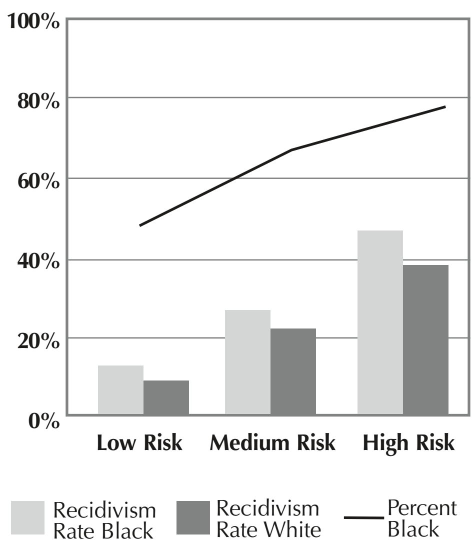


FIGURE 4.
Recidivism Rates by Race and Percent Black in Each Risk Category—Arrest Violent Offense



actuarial risk assessment instruments, namely the COMPAS. The authors' conclusions are rather obvious given the title of their article: "*There's software used across the country to predict future criminals. And it's biased against Blacks.*" However, upon analyzing the same data, we came to a quite different conclusion. This section summarizes the findings of our analyses and then offers insight as to how Angwin et al. (2016) obtained different results. Ultimately, we challenge their understanding of the COMPAS and how it is to be both scored and used, their understanding of research methods and statistics, and, perhaps, their adherence to their own code of ethics.

Our initial analyses looked at the observed recidivism rates for Black and White defendants for any arrest (general recidivism) and for a violent arrest (violent recidivism). Results indicated that Black defendants were significantly more likely to be arrested for any arrest and for violent arrest. In addition, low-, medium-, and high-risk Black defendants were also rearrested more than their low-, medium-, and high-risk White defendant counterparts (for both any arrest and for violent arrest). Our second set of analyses focused on the degree of accuracy for the COMPAS in predicting any arrest and violent arrest. Our results found the COMPAS to be a good predictor of both types of arrest and, more importantly, to predict outcome equally well (i.e., of moderate degree) across both races. Furthermore, logistic regression analyses conducted to estimate the form of the relationship between the COMPAS and outcome (any arrest and violent arrest) revealed no differences in the slope and intercept, indicating that the COMPAS predicts recidivism in a very similar way for both groups of defendants. Most important, the interaction term between race and COMPAS decile score is not significant and adds no predictive power to the models overall (see Tables 5 and 6). Stated differently, the COMPAS does not predict outcome differently across groups of Black and White defendants—a given COMPAS score translates into roughly the same likelihood of recidivism, regardless of race. This may be seen visually in Figures 1 and 2. Higher mean risk scores do not indicate bias if they correspond with higher arrest rates.

In all instances, we failed to find evidence of predictive bias by race in the COMPAS. Interestingly, these findings are remarkably consistent with existing literature that has also tested for bias in other ARAIs (see Skeem & Lowenkamp, 2016a, 2016b). Had Angwin et al. (2016) conducted a thorough review of rigorous research on test bias, they undoubtedly would have discovered the existence of standards for educational and psychological testing put forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). Because they failed to do so, they also failed to test for bias within these existing standards. Given the gravity of their conclusion for criminal justice policy, this failure is neither acceptable nor excusable.

However, failing to perform an exhaustive (or even cursory) literature review that might have informed their "study" is just the beginning of Angwin et al.'s (2016) shortcomings.

In addition to applying the COMPAS to an incorrect population (which in and of itself is sufficient grounds to discredit their study), Larson et al. (2016) imposed a false dichotomy on the COMPAS by reducing the risk categorizations into just two groups defined by the binary categorization of recidivist or non-recidivist. While it is problematic that they collapsed medium- and high-risk defendants into one category that was then compared against the low-risk defendants, more problematic is their interpretation of what information COMPAS scores provide (Singh, 2013). Just as medicine uses actuaries to inform patient prognoses and the auto insurance industry uses actuaries to inform probabilities of risky driving behavior, the COMPAS is based on an actuary designed to inform the probability of recidivism across its three stated risk categories. To expect the COMPAS to do otherwise would be analogous to expecting an insurance agent to make absolute determinations of who will be involved in an accident and who won't. Actuaries just don't work that way. This error discredits their main finding that Black defendants were more likely to be incorrectly identified as recidivists (false positives) while White defendants were more likely to be misclassified as non-recidivists (false negatives). Furthermore, our reanalysis of false positives and false negatives also calls into question the validity of their conclusions regarding this method of analysis when an assessment tool comprises more than just two categories (see Tables 3 and 4).

Another of their main conclusions stems from a Cox regression model predicting general recidivism with a number of variables, including an interaction term between race and COMPAS score. In this analysis, they observed a p value of .0578 for the interaction term and then concluded that race moderated the relationship between outcome and COMPAS score. This erroneous conclusion further demonstrates the carelessness in their approach, as .0578 is less than .05—particularly with a sample size of 10,000—only in the world of "data torturing" (see Mills, 1993), where authors are outright looking for something of significance to make their point.

An additional statistical oddity of Larson et al. (2016) concerns the ordering of variables in their general recidivism logistic regression model, in which they predict the COMPAS score with recidivism and a number of other demographic variables. Because assessment scores occur before recidivism, it appears as though they have their independent and dependent variables confused. We're not sure of the logic behind predicting an assessment score with recidivism but we do believe that this analysis is responsible for their conclusion that, somehow, higher average COMPAS scores for Black defendants indicate bias. Given the higher observed recidivism rates for Black defendants, and given the demonstrated validity of the COMPAS, it is nothing short of logical that these defendants evidence higher COMPAS scores (after all, isn't that precisely what the COMPAS is measuring?).

In summary, this research sought to reanalyze the study by Larson et al. (2016), using accepted methods to assess the presence of test bias. Using these accepted methods, we found no evidence of racial bias. Our analysis of Larson et al.'s (2016) data yielded no evidence of racial bias in the COMPAS' prediction of recidivism—in keeping with results for other risk assessment instruments (Skeem & Lowenkamp, in press; 2016a).

We would be remiss if we failed to report the limitations of our re-analysis of the ProPublica analysis. First, we did not completely replicate the ProPublica study, as we excluded those defendants whose race was not white or black. We also did not estimate the survival analysis models. Second, the outcome measure is limited to new arrest. The limitations (as well as strengths) for this measure have been well documented (see Maltz, 1984). Third, the extent to which the findings of this study are generalizable to other samples, jurisdictions, and other instruments is unknown. Finally, while this article was sent out to numerous colleagues for review and input, it was not a blind review and this research is yet to be published in a peer-reviewed journal.

Conclusion

It is noteworthy that the ProPublica code of ethics advises investigative journalists that "when in doubt, ask" numerous times. We feel that Larson et al.'s (2016) omissions and mistakes could have been avoided had they just asked. Perhaps they might have even asked…a criminologist? We certainly respect the mission of ProPublica, which is to "practice and promote investigative journalism in the public interest." However, we also feel that the journalists at ProPublica strayed from their own code of ethics in that they did not present the facts accurately, their presentation of the existing literature was incomplete, and they failed to "ask." We believe the result demonstrates that they are better equipped to report the research news, rather than to make the research news.

We hope that this rejoinder and its consistency with the existing literature provides some comfort (in the form of evidence) to policymakers who have been exposed to misleading information about the reliability, validity, and fairness of ARAIs. At the very least, this article highlights an accepted and legitimate approach that agencies and jurisdictions can use to determine if the ARAI they use, or are considering using, is in fact subject to predictive bias towards a particular group of people. Clearly, ARAIs hold considerable promise for criminal justice reform in that they are capable of better informing what were previously subjective and indefensible criminal justice decisions (Andrews, Bonta, & Wormith, 2006).

References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648-680.

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

Andrews, D., Bonta, J., & Wormith, S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52(1),* 7-27.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Brennan, T., Dietrich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment. *Criminal Justice and Behavior, 36(1),* 21-40.

Carson, E. A. (2015). Prisoners in 2014. Washington, DC: Bureau of Justice Statistics. Retrieved 10/10/15 from: http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5387

Dawid, A. P. (2015). Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application, 2,* 273-303.

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

Hanson, R. K., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21.

Harris, G. T., Lowenkamp, C. T., Hilton, N. Z. (2015). Estimate for risk assessment precision: Implications for individual risk communication. *Behavioral Sciences and the Law, 33,* 111-127.

Henrichson, C., & Delaney, R. (2012). The price of prisons: What incarceration costs the taxpayer. Vera Institute of Justice. Retrieved from: http://www.vera.org/pubs/special/price-prisons-what-incarceration-costs-taxpayers

Imrey, P. B., & Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy, 2*(1), 1-18.

Kirchner, L. (2016). The Senate's popular sentencing reform bill would sort prisoners by risk score. Retrieved from: https://www.propublica.org/article/senates-popular-sentencing-reform-bill-would-sort-prisoners-by-risk-score

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Retrieved from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Marco, C.A. & Larkin, G. L. (2000). Research ethics: Ethical issues for data reporting and the quest for authenticity. *Academic Emergency Medicine, 7*(6), 691-694.

Maltz, M. D. (1984). *Recidivism*. Originally published by Academic Press, Inc., Orlando, FL.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Mills, J. L. (1993). Data torturing. *The New England Journal of Medicine, 329*(16), 1196-1199.

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*(5), 15-20.

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.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Criminal Justice and Behavior*, *31(1),* 8-22.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violent risk assessment tools: A systematic metaregression analysis of 68 studies involving 25,890 participants. *Clinical Psychology Review, 31(3),* 499-513.

Skeem, J. L., & Lowenkamp, C. T. (2016a). Risk, race, and recidivism: Predictive bias and disparate impact. Manuscript under review. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339

Skeem, J., & Lowenkamp, C. T. (2016b). Race and actuarial risk assessment: The Level of Service Inventory Revised. University of California, Berkeley: Unpublished Manuscript.

Skeem, J. L., Monahan, J., & Lowenkamp, C. T. (in press). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior*

United States Census Bureau (2014). United States Quick Facts. Retrieved from: https://www.census.gov/quickfacts/table/PST045215/00

Wagner, P., & Walsh, A. (2016). States of incarceration: The global context. Prison Policy Initiative. Retrieved from http://www.prisonpolicy.org/global/2016.html

Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis

R. Barry Ruback
Penn State University
Cynthia A. Kempinen
Pennsylvania Commission on Sentencing
Leigh A. Tinik
Pennsylvania Commission on Sentencing
Lauren K. Knoth
Penn State University

RESEARCH IS RELEVANT to policy if it assesses the effects of different policy options using measures that are important to policy makers (Ruback & Innes, 1988).[1] But policy-relevant research by itself does not necessarily lead to policy change. Unless policy makers understand both the findings of the substantive research and how it can be implemented in the policy context, the research is unlikely to be used. Thus, for implementing policy and for understanding the implementation process, it is important to study how policy-relevant research is best communicated to policy makers.

One policy decision currently facing many states is the type of information that should be available for sentencing, treatment, and release decisions in criminal justice. In particular, at criminal sentencing the information judges are given can include or not include an actuarial instrument predicting the likelihood of recidivism. This study experimentally examined the communication of such recidivism risk information to judges, attorneys, and probation officers in order to determine how best to communicate both risk scores and the meaning of those risk scores.

[1] Points of view expressed here do not necessarily represent those of the Pennsylvania Commission on Sentencing.

Actuarial Predictions of Recidivism

Most actuarial risk instruments classify individuals into risk categories, each of which has an associated probability of recidivism based on the proportion of individuals who recidivated (Scurich & John, 2012). The assumption is that an individual in a particular risk group has a probability of recidivating similar to the overall group. Critics of actuarial risk assessments have suggested that the application of group-level probabilities to assess an individual's likelihood of recidivating is inappropriate and fails to meet any standards of precision or certainty (Hart & Cook, 2013). But there have been responses to these criticisms, and the debate over the accuracy and use of risk assessments is ongoing (Mossman, 2015; Harris, Lowenkamp, & Hilton, 2015).

In general, actuarially based predictions are more accurate than clinical judgments because humans are subject to numerous errors and biases (Kahneman, 2011; Meehl, 1986). Moreover, because humans can be tired or bored or distracted, they may make different decisions at different times about the same problem. This inconsistency further lowers the validity of their predictions. Similar criticisms have been made of structured professional judgment (SPJ) models, with some authors concerned about the subjectivity introduced by clinicians during completion of the assessment (Hilton, Harris, & Rice, 2006). The superiority of actuarial predictions over clinical judgments has been argued for 60 years (Meehl, 1954), with studies comparing the accuracy of actuarial risk assessment instruments, SPJ tools, and unstructured clinical judgment finding that pure actuarial models perform as well as or better than SPJ tools or clinical judgments (Harris & Rice, 2015; Campbell et al., 2007; Hanson & Morton-Bourgon, 2004).

Actuarial risk assessment is presumed to have several advantages in sentencing, including improving decision making, limiting discretion, increasing accountability, and better predicting future risks. Because of these advantages, the trend is for jurisdictions to use actuarial instruments in sentencing. In recent years, states have begun to require that actuarial risk scales be incorporated into criminal justice decisions (Monahan & Skeem, 2013). In particular, seven states are developing or have developed statistical models of recidivism for use at sentencing (Hannah-Moffat, 2013). The assumption behind these laws is that judges will be able to make more accurate predictions of future offending if they are given actuarial models than if they rely only on their own knowledge and experience.

In practice, however, actuarial models will be better only if judges, attorneys, probation officers, and others concerned with sentencing understand the statistical information given to them. Although there has been research on how best to convey actuarial risk information to mental health practitioners and (regarding weather) to the general public (Monahan & Steadman, 1996), there has been little work on how risk information should be conveyed to practitioners in the criminal justice system (Buchanan, 2013). States vary in how risk information is presented at sentencing (e.g., length of report, specific scores versus summary levels of risk), but there are "no evidence-based practices to guide decisions" (Casey, Warren, & Elek, 2011, p. 54) about which methods are best. More generally, whether risk assessment tools actually affect and improve sentencing needs to be tested (Skeem, 2013).

The study presented here is an experimental investigation of the communication of statistical information about recidivism risk in sentencing. We were interested in knowing whether the statistical information affects beliefs about risk and, if so, whether these effects are consistent across crimes and cases.

Background for the Study

Act 95 of 2010 (42 PA.C.S. §2154.7) mandated that the Pennsylvania Commission on Sentencing adopt an empirically based risk assessment instrument to be used by judges at sentencing that takes into account an offender's risk of re-offense and threat to public safety and that can be used to help determine whether the offender should be considered for alternative sentencing programs.

Over the past four years, the research staff of the Pennsylvania Sentencing Commission (PCS) have developed an actuarial instrument, based on the procedure outlined by Gottfredson and Snyder (2005), for offenders at Levels 3 and 4 of the Guidelines.[2] The staff focused on these levels because of the wide variety of offense seriousness encompassed in these levels and the variety of possible sentences (including incarceration, probation, and alternative sentencing) that are available under the Guidelines. The researchers developed the model using a random sample of half of the Level 3 and 4 offenders sentenced during the three-year period 2004-2006. Predictors included information in the PCS database, as well as prior criminal history information available from the Pennsylvania State Police. The dependent variable was a rearrest for any crime within three years after sentencing (for those on probation) or after release from incarceration[3] as evidenced in the Pennsylvania State Police database.

The final scale, which was developed to predict recidivism, not reduce recidivism (see Monahan & Skeem, 2013, for the distinction), was a weighted measure of eight factors: age, gender, county, total prior arrests, prior property arrests, prior drug arrests, offense gravity score (the PCS measure of offense seriousness), and whether the current crime was a property offense.[4] Information was included in the scale if it was available statewide to probation and court staff at the time of sentencing (such information as prior drug use, criminal attitudes, and psychopathy was not), if it was reliable, and if it was predictive of subsequent arrest, the latter two being standards proposed by Gottfredson and Moriarty (2006). What these conditions meant was that no dynamic factors and no validated scales (e.g., LSI-R) could be used in the risk scale. Scores on the constructed scale could range from 0 to 14. The scale was validated on the remaining half of the PCS data for the 2004-2006 period and revalidated on PCS data from the years 2007-2008.[5]

Although risk assessment instruments have been used by practitioners in criminal justice, especially for prison classification and parole release decisions, there is little research on how risk information can best be presented to nonspecialists, particularly individuals without statistical training, such as lawyers and judges. This lack of research is problematic because how information is presented affects the way it is used in decisions (Sanfey & Hastie, 1998).

Three examples illustrate the effects of presentation. First, estimates involving frequencies (e.g., 1 in 10) lead to greater perceived risk than the equivalent percentage (e.g., 10 percent), probably because frequencies are easier than percentages to visualize (Slovic, Monahan, & MacGregor, 2000). Second, one study found that clinical psychologists who work in forensic settings were less likely to make release decisions when violent behavior was described in vivid rather than pallid terms (Monahan, Heilbrun, Silver, Nabors, Bone, & Slovic, 2002). Third, there are differences in perceived risk depending on whether information is presented as the probability of an event occurring (e.g., violence) rather than the probability of no event (e.g., no violence). Thus, as compared to statistics framed in a negative fashion (e.g., 74 percent likely to be nonviolent), statistics framed in a positive fashion (e.g., 26 percent likely to be violent) lead to more commitment decisions (Scurich & John, 2011), an effect that occurs because people have a strong aversion to loss (Kahneman, 2011).

The present study addressed four issues. First, does the risk information affect judgments? Second, if so, does the way it is presented make a difference in the size of the effect? We expected the risk information to affect participants' judgments, but we did not have hypotheses about the effects of type of presentation. Third, does the risk information have the same effect across all types of crime, or does it vary by the type of crime?

Fourth, do decision makers in criminal justice have a preference regarding the presentation of risk information? Do they want just the score, information about the items on which the offender had a score, or information about all of the items, regardless of whether or not the item was a risk? And, regarding the meaning of risk score in terms of predicted risk of recidivism, do they prefer that the information be presented in a table or a graph? Based on a study by Scurich, Monahan, and John (2012), we expected participants to prefer more rather than less information, but based on the absence of empirical data, we did not make a hypothesis regarding a preference for type of presentation.

These questions were tested in an experimental framework using case information from six actual cases. Two examples of each

[2] The Guidelines have five sentencing levels, with 1 representing the least serious offenders and 5 representing the most serious offenders. Since this study, the Commission has been working on the development of a risk assessment instrument for all five sentencing levels.

[3] For those sentenced to state prison we used the release date from the Pennsylvania Department of Corrections. For those sentenced to county jail, we estimated the release date using the minimum sentenced imposed.

[4] Subsequent to this study, the Commission decided to eliminate county from the scale, although it is included in the statistical models as a control.

[5] Details about the validation of the risk scale are available in "Interim Report 7: Validation of Risk Scale" available online: http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment/phase-i-reports/interim-report-7-validation-of-risk-scale/view

of three crimes (Burglary, Theft, Drugs) were presented to judges, district attorneys, public defenders, probation officers, and other criminal justice personnel from four counties in Pennsylvania.

Method

Before implementing the risk model the staff had developed, the Pennsylvania Sentencing Commission wanted to test different ways of presenting risk information, in order to determine which method is best understood by those individuals who will incorporate the risk information into the sentencing decision (judges and probation officers), as well as the attorneys (public defenders and assistant district attorneys) who are responsible for making legal arguments about the appropriateness of a criminal sentence. In this study, participants saw one of the six presentation styles of the risk information and were asked to make judgments about each of six cases using this risk information. The results were subsequently presented to focus groups in the four counties in which the study was conducted: Allegheny, Blair, Philadelphia, and Westmoreland. Discussions with a subsample of respondents at these subsequent focus groups were used to better understand the patterns identified in analysis of the survey. The study was approved by the University IRB.

Sample

The target frame consisted of 63 criminal court judges, 449 district attorneys and assistant district attorneys, 248 public defenders, 230 probation officers, and 10 others who worked in one of four counties in Pennsylvania: Allegheny (Pittsburgh), Blair (Altoona), Philadelphia, and Westmoreland (Greensburg). We used a stratified random assignment procedure in order to ensure that across occupations and counties there were approximately an equal number of participants assigned to each of the conditions. That is, within the 20 cells (5 occupations × 4 counties), participants were randomly assigned to the 12 different conditions that are described below.

Emails were sent to the 1000 individuals identified by agency representatives in the four counties. Of these, 38 were returned because the individuals were no longer at the agency or because the address was rejected. We received usable responses from 200 individuals, 21 percent of the 962 individuals who received an email.

The final sample of 200 individuals comprised 79 from Allegheny County (26 percent response rate), 19 from Blair County (53 percent response rate), 75 from Philadelphia County (15 percent response rate), and 27 from Westmoreland County (25 percent response rate).[6] There were 57 district attorneys (13 percent response rate), 24 judges (38 percent response rate), 73 probation officers (32 percent response rate), 30 public defenders (13 percent response rate), and 16 individuals in other positions. Of the 200 individuals, 34 (17 percent) had attended an earlier session at which the risk scale had been presented and discussed.

Procedure

The initial email about the survey was sent out on July 8, 2013. Subsequent reminders were sent out on July 24 and August 9. Data collection was closed on September 1, 2013. Participants received an email from someone in their office that they would be sent a survey in which they would be presented with six cases, including risk information about the offenders. The email from the Research Director of the Pennsylvania Commission on Sentencing is available in Appendix A. Each participant received an email directing him or her to a site in Survey Monkey containing one of the 12 versions of the survey. The case information, based on actual presentence investigation reports, consisted of the types of information typically used by judges: (a) demographic information about the offender (age, sex, race, date of birth); (b) information from the Sentencing Guidelines (offense, Offense Gravity Score, Prior Record Score, guideline recommendation); (c) prior record (juvenile, adult, detainers or charges pending); (d) social history of the offender (marital history, education, employment history, mental health, drug and alcohol history). The information from the presentence investigation reports was condensed into one-page single-spaced summaries for each of the six cases.

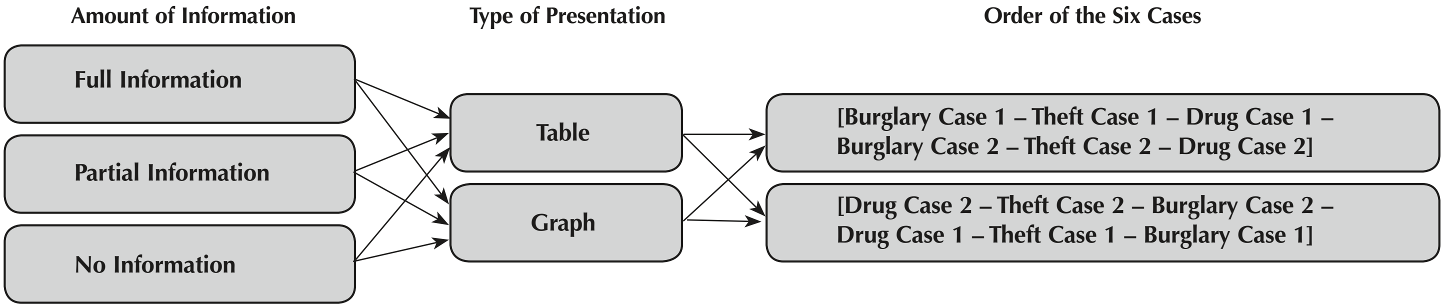
Participants received one of six presentation methods in a 3 × 2 (Amount of Risk Scale Information × Presentation of Recidivism Risk) between-subjects x 3 x 2 (Type of Crime × Cases) within-subjects (repeated measures) design. This mixed-design (between and within subjects) was analyzed using analysis of variance models. There were two between-subjects variables that related to the risk information: amount of information presented about the risk scale (three levels) and presentation of recidivism risk (two levels). The offender's risk score on the overall scale was presented in one of three ways (see Appendix B): (a) the score alone without any further information about the eight factors or the offender's points for each of the eight factors (Risk Score Only); (b) the total score and the number of points for each of the risk factors on which the offender received points (Partial Scale Information); and (c) the total score, the number of possible points for each of the eight categories, and the number of actual points received for each of the eight categories (Full Scale Information). We included the Risk Score Only condition because one of the criticisms of risk assessment scales is that sometimes the people who use them receive only a score, without understanding how that score was arrived at (Hannah-Moffat, 2013).

The risk of recidivism for the offender's risk score was presented in one of two ways, a bar graph or a table (see Appendix C for greyscale version of survey). Each format (graph and table) presented the likelihood of being arrested within three years of release for each of the risk scores (0-14). For the bar graph, the specific offender's likelihood of being arrested within three years of release was highlighted in yellow, and the likelihood of offenders with other risk scores being arrested was shown in blue. For the table, the offender's likelihood of being arrested within three years of release was shown by a number (percentage arrested) and was highlighted in yellow while the recidivism likelihood for offenders with other scores was presented but not highlighted.

After reading summary information about a real case, participants were asked to indicate their judgment of the likelihood that the offender would be arrested within three years of release using a ten-point scale. The scale ranged from 0 to 100 percent and was divided into 10 percent increments (0 - <10 percent, 10 - <20 percent, 20-<30 percent, 30-<40 percent, 40-<50 percent, 50-<60 percent, 60-<70 percent, 70-<80 percent, 80- <90 percent, 90 – 100 percent). For the analyses, this 10-point scale was scored, respectively, as 5 percent, 15 percent, 25 percent, 35 percent, 45 percent, 55 percent, 65 percent, 75 percent, 85 percent, and 95 percent.

[6] An additional 46 individuals started but did not complete the survey (20 individuals dropped out during the first case, an additional 14 dropped out during the second case, an additional 6 dropped out during the third case, and a final 6 dropped out during the fourth case). The responses of all 46 of these individuals were excluded from subsequent analyses.

FIGURE 1.
Design of the Experiment: Three Amounts of Information, Two Types of Presentation, Two Orders of Six Cases



Note. Participants in the study received one of three amounts of information (full, partial, or none) in one of two formats (table or graph). The six cases they reviewed were presented in one of two orders.

After making this risk judgment, participants then saw the risk scale information (Risk Score Only, Partial Scale Information, or Full Scale Information) and the presentation of the recidivism risk (Bar Graph or Table). Following exposure to the risk and contextual information, respondents were again asked to indicate the likelihood that the offender would be arrested within three years. They then indicated the type of sentence they would impose (e.g., prison, jail, probation) and the length of the sentence they would impose. Participants made these same judgments about risk and sentence for each of the six cases.[7]

The six cases were presented in one of two counterbalanced orders: (a) Burglary-1, Theft-1, Drug-1, Burglary-2, Theft-2, Drug-2 or (b) Drug-2, Theft-2, Burglary-2, Drug-1, Theft-1, Burglary-1. In sum, what varied among participants was (a) the presentation of the risk scale (each participant saw one of three ways), (b) the presentation of the recidivism risk (each participant saw one of two ways), and (c) the order in which the six cases were presented (each participant saw one of two ways). Thus, there were 12 different surveys (3 presentations of risk scale x 2 presentations of recidivism information x 2 orders of the six cases). A diagram of the study design is presented in Figure 1. Participants received one of these 12 surveys and made judgments about six cases.

In addition to representing different crimes, the cases represented different actuarial risk levels, as shown by the scores on the 14-point risk scale we created and the associated risk of recidivism within three years (see Table 1). Participants read and made judgments on all six cases.

With regard to the order of presentation of the six cases, across the analyses there were 6 significant effects (of 48 tests involving order).[8] Although this number is above chance, there was no systematic pattern of effects across the three repeated measures analyses of variance that were conducted (one analysis for each set of two cases within a crime type). Thus, we collapsed the other variables across order and do not discuss order further. However, that there were any significant effects indicates that we were correct in counterbalancing order across participants, since punishment judgments can be affected by order (e.g., Pepitone & DiNubile, 1976).

The final set of questions asked participants to make judgments about the six different ways of presenting risk information. Respondents were presented with each of the six methods of presenting risk scale and recidivism information. For each method of presentation, respondents were asked how satisfied they were with the level of detail included in that particular presentation of risk information and how easy it was to understand and interpret each presentation of risk information. For measures of satisfaction, respondents were given a 5-point scale ranging from very dissatisfied to very satisfied. For measures of understanding, respondents were given a 5-point scale ranging from very difficult to very easy. Finally, respondents were asked to rank each of the six methods of presentation in order from most favorite to least favorite. The survey took about 35 minutes to complete.

Results

The results are presented in terms of the four issues that were the focus of the investigation: (a) whether the risk information affects practitioners' judgments; (b) whether the way risk information is presented affects those judgments; (c) whether the effects of actuarial risk information are consistent across crimes and cases; and (d) whether criminal justice practitioners have a preference regarding the presentation of risk information.

Effect of the Risk Information

The effect of the risk information was assessed in two different ways. First, we examined, across all cases and all respondents, the difference between respondents' initial judgment of risk (after the case information) and the final judgment of risk (after the risk information had been presented). Second, we examined the effects of risk information by type of crime and by presentation of the risk information.

Pre/Post difference. Each of the 200 respondents was asked to make pre/post risk judgments about six cases, for a total of 1200 difference scores. Responses were excluded for cases in which the respondent did not provide both a pre and post estimate (36 instances),

[7] The length of the presentence investigations ranged from 14 to 31 pages, about 2-6 pages of which were the state's guidelines forms for the case. Because there is no uniform method for reporting a presentence investigation, there are dramatic differences across jurisdictions. In the one-page summaries, we included those important static factors that were consistently recorded across all counties. Because presentence investigations are conducted in only about one-quarter of criminal cases in Pennsylvania, detailed social histories and validated needs and risk scales are not available statewide and were not used in either the PCS risk scale or in this study. In that these six cases were based on cases with presentence investigation, they may not be representative of cases in general.

[8] For each of the three crimes, there are 4 between-subjects effects involving order, 4 within-subjects effects involving case and order, 4 within-subjects effects involving risk judgments and order, and 4 within-subjects effects involving case, risk judgments, and order. Across the three crimes, there are 48 effects involving order.

TABLE 1.
Descriptive Information, Risk Estimates, and Proposed Sentence for the Six Cases

Case Type	Actuarial Risk Level	Recidivism Rate for Risk Level	Pre Risk Estimate[a]	Post Risk Estimate[a]	t-test[b]	% Prison Sentence Imposed	*M* Length of Prison Sentence Imposed (months)[c]	Actual Prison Sentence (months)	
								Minimum	Maximum
Burglary Case 1	10	69%	72.0%	73.5%	1.73	78%	24.6	48	120
Burglary Case 2	7	47%	63.7%	61.2%	3.51***	53%	15.4	12	36
Theft Case 1	4	26%	34.9%	31.4%	4.89***	28%	11.5	9	24
Theft Case 2	10	69%	75.8%	76.7%	1.27	76%	23.1	15	36
Drug Delivery Case 1	5	33%	47.6%	44.6%	3.39***	34%	12.8	24	48
Drug Delivery Case 2	4	26%	63.7%	50.1%	11.66***	29%	12.4	12	24

Note. Actuarial Risk Level and the Recidivism Rate for Risk Level came from analyses conducted by research staff at the Pennsylvania Commission on Sentencing. The Actual Prison Sentence came from records of the Pennsylvania Commission on Sentencing.

[a]The Pre-Risk and Post-Risk estimates, based on the 10-point scale completed by respondents (0 - <10%, 10 - <20%, 20-<30%, 30-<40%, 40-<50%, 50-<60%, 60-<70%, 70-<80%, 80- <90%, 90 – 100%), were scored, respectively, as 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95%.

[b]Paired sample t-tests were used to test for significant changes in the pre and post risk estimates.

[c]Mean Length of Prison Sentence Imposed in months was converted from the 8-point scale completed by respondents: 1 = 0 months; 2 = < 6 months; 3 = 6-12 months; 4 = 12-18 months; 5 = 18-24 months; 6 = 24-30 months; 7 = 30-36 months; 8 = > 36 months.

*** p < .001

resulting in a final sample of 1164 scores. Only 13 percent of the respondents made no changes in any of the six cases after seeing the risk information. The distribution of respondents by number of changes across the six cases was as follows: 1–18 percent, 2–17 percent, 3–15 percent, 4–17 percent, 5–13 percent, 6–6 percent. On average, respondents changed their risk judgments on 2.61 cases (SE = .18). Of the 1164 possible pre/post judgments, there was a change in 521 (45 percent). Change was most likely for the two drug crimes and somewhat less likely for the two burglaries and the two thefts (see Table 2).

Over all participants, crimes, and cases, there was a significant mean pre/post difference in respondents' judgments, $M = .34$, 95 percent CI [.26, .43], indicating that overall the risk information decreased respondents' judgments of risk. However, these changes were not consistent across respondents or crimes. As can be seen in Table 1, there were significant differences between the pre and post risk judgments on four of the six crimes. For all four of these crimes, the post-information mean was smaller and closer to the actuarially determined rate of recidivism than was the pre-information mean, indicating that in general respondents' judgments were influenced in the direction indicated by the actuarial information. Notably, there was no significant difference between the pre-information mean and the post-information mean for offenders with the highest risk score.

TABLE 2.
Descriptive Information: Changes in Pre and Post Information Risk Estimates

	Total Responses	Estimates Changed Pre/Post		Estimates Unchanged Pre/Post	
		N	%	N	%
Burglary Case 1	195	77	39.49	118	60.51
Burglary Case 2	193	82	42.49	111	57.51
Theft Case 1	194	80	41.24	114	58.76
Theft Case 2	193	63	32.64	130	67.36
Drug Case 1	193	96	49.74	97	50.26
Drug Case 2	196	123	62.76	73	37.24
Total	**1164**	**521**	**44.76**	**643**	**55.24**

Effects of presentation type and type of crime on ratings of risk. Aside from knowing whether the actuarial risk information affected ratings of risk overall, we tested whether the effect of the risk information differed depending on how it was presented in terms of the amount of risk information (none, partial, full) and the way the recidivism information was presented (table or graph). Analyses were conducted within each of the three types of crimes using a doubly repeated measures (2 risk judgments for each of 2 cases) analysis of variance. The between-subjects factors were amount of risk information (none, partial, full) and the type of presentation of the meaning of the risk score for recidivism (table or graph). No significant effects were found for the amount of information provided about the risk scale, and this variable is consequently not discussed further.

For Burglary, there were no significant between-subjects effects. Regarding within-subjects effects, there was a significant effect for the two burglary cases, $F(1, 194) = 52.47$, $p = .000$, $\eta_p^2 = .21$, indicating that there was a significant difference in perceived risk in the two cases ($M_1 = 7.25$, SE = .13 vs. $M_2 = 6.23$, SE = .13). The effects for the two burglary cases were conditioned by two significant interactions. First there was a significant Burglary Cases x Risk Judgments interaction, $F(1, 194) = 17.66$, $p = .000$, $\eta_p^2 = .08$ (see Table 1), such that for the first case there was a slight increase from pre to post ($M_{pre} = 7.17$, SE = .14 and $M_{post} = 7.34$, SD = .13), whereas for the second case there was a slight decrease ($M_{pre} = 6.37$, SE = .15 and $M_{post} = 6.09$, SE = .13). Second, there was a significant interaction of Burglary Case x Risk Presentation Type, $F(1, 194) = 4.29$, $p < .04$, $\eta_p^2 = .02$. For the first case, there was greater judged risk when recidivism information was presented using a graph rather than a table ($M_{graph} = 7.42$, SE = .18 and $M_{table} = 7.09$, SE = .19). In contrast, for the second case, there was more judged risk when recidivism information was presented using a table rather than a graph ($M_{graph} = 6.10$, SE = .19 and $M_{table} = 6.36$, SE = .19).

For Theft, there were no significant between-subjects effects. Regarding within-subjects effects, there was a significant effect for the two theft cases, $F(1, 190) = 634.81$, $p = .000$, $\eta_p^2 = .77$, indicating that there was a significant difference in perceived risk in the two cases ($M_1 = 3.33$, SE = .13 vs. $M_2 = 7.61$, SE = .13). There was a significant effect for risk scores, $F(1, 190) = 9.22$, $p = .003$, $\eta_p^2 = .05$, such that the pre score was higher than the post score ($M_{pre} = 5.54$, SE = .11 vs. $M_{post} = 5.40$, SE = .09). Both of the main effects were conditioned by a significant Theft Cases x Risk Judgments interaction, $F(1, 190) = 18.00$, $p = .000$, $\eta_p^2 = .09$ (see Table 1), such that for the first case there was a slight decrease from pre to post ($M_{pre} = 3.51$, SE = .15 to $M_{post} = 3.15$, SE = .13), whereas for the second case there was a slight increase ($M_{pre} = 7.57$, SE = .14 to $M_{post} = 7.65$, SE = .12).

For the Drug crimes, there were no significant between-subjects effects. Regarding within-subjects effects, there was a significant effect for the two drug cases, $F(1, 192) = 42.75$, $p = .000$, $\eta_p^2 = .18$, indicating that there was a significant difference in perceived risk in the two cases ($M_1 = 4.58$, SE = .15 vs. $M_2 = 5.70$, SE = .15). There was also a significant effect for the two risk judgments, $F(1, 192) = 97.50$, $p = .000$, $\eta_p^2 = .34$, such that the pre score was higher than the post score ($M_{pre} = 5.57$, SE = .13 vs. $M_{post} = 4.72$, SE = .13). Both of the main effects were conditioned by a significant Drug Cases x Risk Judgments interaction, $F(1, 192) = 61.50$, $p = .000$, $\eta_p^2 = .24$ (see Table 1), such that for the first case there was a decrease from pre to post ($M_{pre} = 4.75$, SE = .17 to $M_{post} = 4.42$, SE = .14), whereas for the second case there was a large decrease ($M_{pre} = 6.39$, SE = .16 and $M_{post} = 5.02$, SE = .16).

In general, then, there were differences between cases within crimes and differences between pre and post risk judgments. But the amount of risk information did not affect any of the post risk judgments, and the presentation of the recidivism information affected only the post judgments for the two burglary crimes.

Relationship to actual sentences. One of the fears of providing actuarial risk information at sentencing is that there will be an increase in punishment severity (Hannah-Moffat, 2013). To test that notion, we examined the percentage of individuals who said they would incarcerate the individual. In actuality, all six individuals had been incarcerated in state prison. As can be seen in Table 1, among respondents the incarceration rates for the six cases ranged from 28 percent to 78 percent, and two of the mean incarceration sentences were for less time than was actually imposed. Thus, these data suggest that actuarial risk information does not necessarily increase punishment severity, an initial conclusion that warrants further research.

Preference for How Risk Information is Presented

At the end of the survey, respondents were shown all six combinations of risk information and the meaning of the risk information used in this study. They were then asked to rank the six combinations in terms of their preference for how the information should be presented at sentencing. As shown in Table 3, respondents showed two clear preferences: (a) a preference for more information about the risk scale: full information over partial information over no information, and (b) a preference for the graph over a table, within each one of those information levels. Ratings of understanding were related only to the level of information about the risk scale: full information over partial information over no information. Ratings of satisfaction followed the same pattern regarding level of information, although within the full information and partial information categories, respondents said they were more satisfied with the table than with the graph.

Discussion

This study was designed to test the impact of actuarial risk information on decision makers' judgments of risk and to examine whether these effects were consistent across crime types and across cases within crime types.

TABLE 3.
Ratings of Six Different Ways of Presenting the Risk Information and the Meaning of the Risk Information

Amount of Risk Information	Presentation of Recidivism Information	Mean Rank		Satisfaction		Understanding	
		Mean	SE	Mean	SE	Mean	SE
Full	Graph	4.88[f]	(.104)	3.46[d]	(.067)	3.78[c]	(.053)
Full	Table	4.46[e]	(.101)	3.54[e]	(.070)	3.71[c]	(.062)
Partial	Graph	3.69[d]	(.094)	2.98[b]	(.075)	3.44[b]	(.068)
Partial	Table	3.26[c]	(.096)	3.13[c]	(.074)	3.47[b]	(.064)
None	Graph	2.68[b]	(.104)	2.48[a]	(.076)	2.96[a]	(.081)
None	Table	2.35[a]	(.117)	2.55[a]	(.080)	3.10[a]	(.080)
	F	94.88***	$\eta_p^2 = .348$	69.33***	$\eta_p^2 = .270$	37.45***	$\eta_p^2 = .169$
	N	179		188		185	

Note. For rankings, higher numbers indicate greater preference. For ratings of satisfaction and understanding, higher numbers indicate, respectively, higher satisfaction and greater understanding. Within a column, means with different superscripts are significantly different according to a post-hoc Newman-Keuls test ($p < .05$).

*** $p < .001$

Effect of Risk Score Information

Even though the risk score information significantly affected mean risk judgments overall and in four of the six cases, only 45 percent of all possible decisions were affected by the risk information. Moreover, about a third of the sample changed no judgments or only one risk judgment, and almost half changed two or fewer of the six judgments. Given the fear that actuarial risk information would overwhelm all other information in risk judgments, it is somewhat surprising that the risk score information did not have stronger effects on the post-risk-presentation ratings.

There are four possible reasons why the risk information may not have had a stronger effect. First, the participants knew the study was a simulation and they may not have taken the study seriously. As with all simulations, this possibility cannot be discounted. Second, it is possible that the participants did not understand the risk information and therefore were not influenced by it. This possibility is unlikely, however, in that the average movement was in the direction toward that indicated by the actuarial information. Third, the participants likely considered themselves to be experts, and, as such, they would be likely to discount other information. This explanation would be an example of resistance to using actuarial information (Elstein, 1976). In this study, the respondents may have thought that they were considering cases that were exceptions to the general information presented in the actuarial recidivism scale (i.e., what Meehl calls the "broken leg" problem). Fourth, in the presentence report the participants already had all of the information that was used in the actuarial scale presented to them (i.e., age, gender, prior record, offense severity, county), and they may have believed that the scale thus added no new information.

One of the concerns about the use of risk scales, which have the appearance of scientific validity, is that it would be too determinative of the final outcome (Hannah-Moffat, 2013). Our results suggest that is not the case. For the cases we used, although overall the risk information tended to reduce respondents' judgments of risk, the resulting reductions were small in magnitude (an average change of 4.2 percent across the six cases, ranging from .9 percent change to 13.6 percent change). This differential effect across cases suggests that respondents were using the information appropriately for individual cases, rather than being overwhelmed by the actuarial information.

The inconsistent effects of the risk information across the three crimes and the two cases within each crime type also suggest that it is incorrect to say that actuarial risk information has a single effect. Rather, pending further research, it appears that decision makers consider it differently for different crimes and different cases.

Preference for and Effect of Presentation of Information

One of Hannah-Moffat's (2013) fears was that judges would receive only summary scores of the risk scale and would therefore not understand the underlying basis for the score. The results of this study suggest, consistent with Hannah-Moffat's concern, that judges, attorneys, and other criminal justice personnel prefer full information about the risk scale. Moreover, the possible concern about full information (i.e., that it would confuse people) was not borne out, in that there was no difference in the post-risk judgments with respect to the amount of risk scale information presented (none, partial, full), although low statistical power is a possible explanation for the absence of difference.

There was only one significant effect of the type of presentation (table or graph) of recidivism information (on the two burglary crimes, but not on the two theft crimes or the two drug crimes). Our conclusion would be that although respondents had a clear preference for full information about the risk scale and a preference for the graph over the table, in general these factors have little effect on judgments of risk.

Subsequent Focus Groups

About six months after the survey, we presented the results of the survey to focus groups of 15-20 individuals in each of the four counties and asked participants to comment on the findings. At all four focus groups, attendees agreed with the finding that the full risk information should be presented, rather than only a partial amount of information or only the summary risk score. Similarly, the general sense of participants was a slight preference for a graph over a table.

The real question to focus group participants was how the risk scale should be used at sentencing. Three options were presented: (a) as simply another piece of information to be used at sentencing, (b) as information incorporated into the Sentencing Guidelines, or (c) as explicitly mitigating or aggravating information. For the most part, participants suggested that the scale should be just another piece of information that judges consider at sentencing.

In part, this preference for a limited use of the actuarial risk scale was because of a concern, voiced at all four focus groups, about biases that may be built in to the recidivism scale (e.g., arrest bias against certain groups). Although judges, district attorneys, and probation officers are almost certainly aware of these possible biases inherent in the data (see Hannah-Moffat, 2013), the focus group discussions suggest that it might be worthwhile to remind decision makers about the limitations of actuarial scales.

Policy-Relevant Research

This study was initiated by the Commission in order to help them determine how best to communicate risk information to judges, attorneys, and probation officers. Our research suggests that these decision makers want complete information, that they generally understand the concepts and findings, and that they are not overwhelmed by the improved accuracy of actuarial over clinical predictions. Our finding that the actuarial risk information tended to lower respondents' judgments of risk may be unique to the cases we used in the study and must be replicated with other cases and other crimes. In addition, the next step is to test how the presentation of actuarial risk information is used in actual decisions.

References

Buchanan, A. (2013). Violence risk assessment in clinical settings: Being sure about being sure. *Behavioral Sciences & the Law*, *31*(1), 74-80.

Campbell, M. A., French, S., & Gendreau, P. (2007). *Assessing the utility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders*. Ottawa, ON: Public Safety Canada.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011). Using offender risk and needs assessment information at sentencing: Guidance for courts from a national working group. Williamsburg, VA: National Center for State Courts. http://www.ncsc.org/~/media/Files/PDF/Services%20and%20Experts/Areas%20of%20expertise/Sentencing%20Probation/RNA%20Guide%20Final.ashx

Elstein, A. S. (1976). Clinical judgment: Psychological research and medical practice. *Science*, *194*, 696-700. doi: 10.1126/science.982034

Gottfredson, D. M., & Snyder, H. N. (2005). The mathematics of risk classification: Changing data into valid instruments for juvenile courts. NCJ 209158. Washington, DC: Office of Juvenile Justice and Delinquency Prevention. https://www.ncjrs.gov/pdffiles1/ojjdp/209158.pdf

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency*, *51*, 178-200. doi:10.1177/0011128705281748

Hannah-Moffat, K. (2013). Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, *30*, 270-296. doi:10.1080/07418825.2012.682603

Hanson, R. K., & Morton-Bourgon, K. (2004). Predictors of sexual recidivism: An updated meta-analysis 2004-02. *Ottawa, Canada: Public Safety and Emergency Preparedness Canada*.

Harris, G. T., Lowenkamp, C. T., & Hilton, N. Z. (2015). Evidence for risk estimate precision: implications for individual risk communication. *Behavioral Sciences & the Law*, *33*, 111-127.

Harris, G. T., & Rice, M. E. (2015). Progress in violence risk assessment and communication: Hypothesis versus evidence. *Behavioral Sciences & the Law*, *33*(1), 128-145.

Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences and the Law*, *31*, 81-102. doi: 10.1002/bsl.2049

Hilton, N. Z., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist*, *34*(3), 400-409.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*, 370-375. doi:10.1207/s15327752jpa5003_6

Monahan, J., Heilbrun, K., Silver, E., Nabors, E., Bone, J., & Slovic, P. (2002). Communicating violence risk: Frequency formats, vivid outcomes, and forensic settings. *International Journal of Forensic Mental Health*, *1*, 121-126. doi:10.1080/14999013.2002.10471167

Monahan, J., & Skeem, J. L. (2013). Risk redux: The resurgence of risk assessment in criminal sanctioning. Presentation at the annual meeting of the National Association of Sentencing Commissions, Minneapolis.

Monahan, J., & Steadman, H. J. (1996). Violent storms and violent people: How meteorology can inform risk communication in mental health law. *American Psychologist*, *51*, 931-938. doi:10.1037/0003-066X.51.9.931

Mossman, D. (2015). From group data to useful probabilities: The relevance of actuarial risk assessment in individual instances. *Journal of the American Academy of Psychiatry and the Law Online*, *43*(1), 93-102.

Pepitone, A., & DiNubile, M. (1976). Contrast effects in judgments of crime severity and the punishment of criminal violators. *Journal of Personality and Social Psychology*, *33*, 448-459. doi:10.1037/0022-3514.33.4.448

Ruback, R. B., & Innes, C. A. (1988). The relevance and irrelevance of psychological research: The example of prison crowding. *American Psychologist*, *43*, 683-693.

Sanfey, A., & Hastie, R. (1998). Does evidence presentation format affect judgment? An experimental evaluation of displays of data for judgments. *Psychological Science*, *9*, 99-103. doi:10.1111/1467-9280.00018

Scurich, N., & John, R. S. (2011). Prescriptive approaches to communicating the risk of violence in actuarial risk assessment. *Psychology, Public Policy, and Law*, *18*, 50-78. doi:10.1037/a0024592

Scurich, N., & John, R. S. (2012). A Bayesian approach to the group versus individual prediction controversy in actuarial risk assessment. *Law and Human Behavior*, *36*, 237-246. doi:10.1037/h0093973

Scurich, N., Monahan, J., & John, R. S. (2012). Innumeracy and unpacking: Bridging the nomothetic/idiographic divide in violence risk assessment. *Law and Human Behavior*, *36*, 548-554. doi:10.1037/h0093994

Skeem, J. (2013). Risk technology in sentencing: Testing the promises and perils (Commentary on Hannah-Moffat, 2011). *Justice Quarterly*, *30*, 297-303. doi:10.1080/07418825.2012.687513

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, *24*, 271-296. doi:10.1023/A:1005595519944

Appendix A: Survey recruitment email issued by the Research Director of the Pennsylvania Commission on Sentencing to potential respondents.

Act 95 of 2010 mandated the Pennsylvania Commission on Sentencing to develop a risk assessment instrument to assist the court at sentencing. To address this new mandate, the Commission undertook a Risk Assessment study to determine what factors best predict which offenders will be rearrested for a new crime. This study involved over 18,000 offenders from Levels 3 and 4 of the sentencing guidelines who were sentenced during 2004-2006. Eight factors were found to be the best predictors of rearrest: gender, age, county, number of prior adult arrests, prior property arrest, prior drug arrest, property offender, and offense gravity score. These eight factors were used to develop a Risk Assessment Scale to identify offenders at low risk of rearrest. The Scale resulted in risk scores ranging from 0 to 14, with low risk being defined by the Commission as a score of 0 to 4.

The Pennsylvania Commission on Sentencing is conducting this survey to determine how best to present risk information. This survey will take about 30 minutes to complete. You will be presented with 6 case scenarios and corresponding risk information and asked to answer 4 questions per scenario. After completing the 6 scenarios, you will be asked to compare different presentations of risk information. Your participation is voluntary. You can stop at any time and you do not have to answer any questions you do not want to answer.

Your participation is confidential. The survey does not ask for any information that would identify you or allow us to link you to your responses. In the event of any publication or presentation resulting from the research, information will be presented only in large categories of people. Please contact *** with any questions about this survey. We thank you in advance for your participation.

Appendix B. Amount of Information About the Risk Scale Presented with the Risk Score

OPTION 1
Risk Score only

This offender has a risk score of 10.

The Commission has determined risk scores 0-4 to be low risk.

OPTION 2
Risk Score with partial information

This offender has a risk score of 10.

The Commission has determined risk scores 0-4 to be low risk. Below is the calculation of the offender's risk score based on the 8 identified risk factors. Displayed is the number of actual points received by the offender. The offender's total risk score is the sum of points received across all 8 risk factors. The sum ranges from 0-14.

	Risk *Actual Points*
Gender	
Male	1
Age	
30-49	1
County	
Semi-urban	1
Number Prior Adult Arrests	
13+	4
Prior Property Arrest	
Yes	1
Prior Drug Arrest	
Yes	1
Current Property Conviction	
Yes	1
Offense Gravity Score [OGS]	
4+	0
Total Risk Score	10

OPTION 3
Risk Score with full information

This offender has a risk score of 10.

The Commission has determined risk scores 0-4 to be low risk. Below is the calculation of the offender's risk score based on the 8 identified risk factors. Displayed is the number of possible points for each risk factor and the number of actual points received by the offender. The offender's total risk score is the sum of points received across all 8 risk factors. The sum ranges from 0-14.

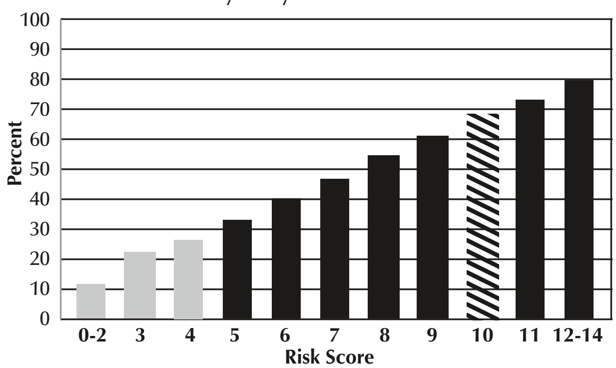
Risk Scale	*Possible Points*	*Actual Points*
Gender		
Male	1	1
Female	0	
Age		
Less than 24	3	
24-29	2	
30-49	1	1
50+	0	
County		
Rural	0	
Semi-urban	1	1
Urban	2	
Number Prior Adult Arrests		
0	0	
1	1	
2-4	2	
5-12	3	
13+	4	4
Prior Property Arrest		
No	0	
Yes	1	1
Prior Drug Arrest		
No	0	
Yes	1	1
Current Property Conviction		
No	0	
Yes	1	1
Offense Gravity Score [OGS]		
1-3	1	
4+	0	0
Total Risk Score		**10**

Appendix C. Types of Presentation of Recidivism Rates for Different Risk Scores

OPTION 1
Graph

The graph below depicts the offender's likelihood of being arrested within 3 years of release from incarceration or imposition of probation/county IP (striped bar) compared to other offenders with different risk scores. The low risk scores are highlighted in grey.

Percentage of Offenders Arrested within 3 Years of Release from Incarceration or Imposition of Probation/County IP by Risk Score



OPTION 2
Table

The table below displays the offender's likelihood of being arrested within 3 years of release from incarceration or imposition of probation/county IP (highlighted in dark grey) compared to other offenders with different risk scores. The low risk scores are highlighted in light grey.

Percentage of Offenders Arrested within 3 Years of Release from Incarceration or Imposition of Probation/County IP by Risk Score

Risk Score	Percent Arrested
0-2	12
3	23
4	26
5	33
6	40
7	47
8	55
9	61
10	69
11	73
12-14	80

Examining Changes in Offender Risk Characteristics and Recidivism Outcomes: A Research Summary

Thomas H. Cohen [1]
Christopher T. Lowenkamp
Scott W. VanBenschoten
Probation and Pretrial Services Office
Administrative Office of the U.S. Courts

THE POST CONVICTION Risk Assessment (PCRA) is a correctional assessment tool used by federal probation officers that identifies offenders most likely to commit new crimes and the criminogenic characteristics that, if changed, could reduce the likelihood of recidivism. Implementation of the PCRA allows federal probation officers to measure whether the criminogenic factors of offenders are changing over time and the relationship of these changes to subsequent reoffending behavior. We explored how changes in offender risk influence the likelihood of recidivism (i.e., arrests for either felony or misdemeanor offenses within one year after the second PCRA assessment) by tracking a sample of 64,716 offenders placed on federal supervision. The study found that many offenders initially classified at the highest risk levels moved to a lower risk category in their second assessment and that offenders tended to improve the most in the PCRA risk domains of employment and substance abuse.

The study also found that high, moderate, and low-moderate risk offenders witnessing decreases in either their risk classifications (i.e., going from high to moderate risk) or overall PCRA scores (i.e., going from 18 to 15 points) were less likely to recidivate compared to their counterparts whose risk levels or scores remained unchanged or increased. Conversely, increases in offender risk were associated with higher rates of arrests irrespective of whether the increase in risk involved higher risk levels or overall PCRA scores. For the most part, offenders with decreasing scores in any of the dynamic risk domains were consistently less likely to be rearrested. Finally, offenders in the lowest risk category saw no recidivism reduction if either their overall score or the score of any of their risk domains decreased.

This is a synopsis of key findings from our study examining federally supervised offenders with multiple PCRA assessments, which was published in the journal *Criminology and Public Policy* (Cohen et al., 2016). The PCRA is a dynamic fourth-generation risk assessment tool that predicts an offender's likelihood of recidivism at multiple time points. This instrument identifies offenders who are most likely to recidivate, ascertains crime-supporting characteristics that will benefit from supervision intervention, and provides information on barriers to successful offender re-integration and/or treatment (AOUSC, 2011).

With the implementation of the PCRA, we can for the first time investigate how much the risk levels of offenders are decreasing between assessments, which risk domains are most likely to get better, and whether offenders with declining risk levels are being arrested less frequently compared to their counterparts with stable or increasing risk levels. These issues are explored in this study using a sample of federally supervised offenders with multiple PCRA assessments. Before discussing this study's findings and implications, we briefly provide an overview of the PCRA risk tool, discuss previous research on the PCRA's capacity to assess change in offender recidivism risk, and detail the methodological approaches utilized in this study.

Using the PCRA to Examine Changes in Offender Risk

The PCRA is a dynamic risk assessment instrument that was developed for United States probation officers (Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011; Lowenkamp, Johnson, VanBenschoten, Robinson, & Holsinger, 2013). The instrument uses five general domains that have been shown to be both theoretically and statistically predictive of offender recidivism: criminal history, education/employment, substance abuse, social networks, and cognitions (i.e., attitudes towards supervision) (Johnson et al., 2011; Lowenkamp et al., 2013). The PCRA has been shown to be highly predictive of whether an offender will reoffend after the

[1] Thomas H. Cohen, Social Science Analyst, Christopher T. Lowenkamp, Social Science Analyst, and Scott VanBenschoten, Branch Chief, Probation and Pretrial Services Office, Administrative Office of the United States Courts, Washington, DC. Correspondence concerning this article should be addressed to Thomas H. Cohen, Administrative Office of the U.S. Courts, One Columbus Circle, NE, Washington, DC 20544. Email: Thomas_cohen@ao.uscourts.gov. A longer version of this paper has been published in the academic journal *Criminology and Public Policy* (See Cohen, Lowenkamp, & VanBenschoten, 2016). Readers interested in a longer version of this paper should contact the authors for more information.

commencement of his or her supervision term. For details of studies describing the construction and validation of the PCRA, see Johnson et al. (2011), Lowenkamp et al. (2013), and Lowenkamp, Holsinger, and Cohen (2015).

Although the predictive utility of the PCRA has been demonstrated, we have only recently begun exploring how this instrument can measure changes in offender risk over multiple assessments and observe how changes in risk are correlated with subsequent recidivism activity. A follow-up PCRA validation study conducted by Lowenkamp et al. (2013) found that offenders whose risk classification increased were more likely to recidivate compared to their counterparts with stable or decreasing risk classifications. In a more recent publication, Cohen and VanBenschoten (2014) found that many offenders initially classified at the highest risk levels moved to a lower risk category in their second assessment and that offenders experiencing improvements in their risk levels were less likely to have their supervision terms revoked compared to offenders with stable or increased risk classifications.

Method

Study Population

We began our inquiry by obtaining data on all offenders within the federal probation system who received an initial PCRA assessment between August 1, 2010, and October 15, 2012.[1] This data extract resulted in us obtaining information on 107,754 offenders with at least one PCRA assessment. From this population of 107,754 offenders, we excluded 43,038 offenders who were not reassessed during the study time frame. Offenders may not receive reassessments for numerous reasons. For instance, prior to the next assessment, they may be revoked, receive an early or successful termination, or be placed on administrative supervision involving minimal officer contact. Ultimately, an offender's initial risk classification influences the type of disposition that might occur before the next assessment. For example, nearly three-fifths of low-risk offenders without second assessments were successfully terminated from supervision before their next assessment, while similar percentages of high-risk offenders were revoked from supervision before receiving their next assessment (see Appendix Table 1).[2] The fact that sizable numbers of offenders with one assessment were never reassessed is intrinsic to most studies examining the relationship between changes in risk characteristics and recidivism (Howard & Dixon, 2013), and illustrates the point that these findings are applicable only to those offenders who received at least two PCRA assessments in our study time frame.

From the initial extract of 107,754 offenders, 64,716 received at least two PCRA assessments between August 2010 and October 2013, which represents the time we stopped tracking these offenders. We used the PCRA assessment rather than the actual supervision start date to anchor this study because when the PCRA was rolled out, PCRAs were done on offenders who may have been well into their supervision term. We decided not to restrict our study population to offenders with short time periods between their supervision start and PCRA assessment dates because we were focused on examining the relationship between changing PCRA risk scores and recidivism irrespective of how long the offender had been on supervision.

The PCRA assessments and re-assessments were conducted as part of the operational supervision duties of federal probation officers. An average of nine months separated the first from the second PCRA assessment. Descriptive information about the study population is provided in Table 1. This table shows that 85 percent of offenders in the study were sentenced to a term of supervised release, meaning that they had finished an incarceration term with the Federal Bureau of Prisons; the remainder had been sentenced directly to a term of probation. According to the PCRA, 78 percent of offenders with at least two PCRA assessments were initially classified as either low (34 percent) or low/moderate (44 percent) risk, while 18 percent were moderate and 5 percent were high risk. A combined 76 percent of offenders examined were either non-Hispanic whites (38 percent) or blacks (38 percent), while another 19 percent were Hispanic. Over four-fifths were male and the average age was 40 years.

TABLE 1.
Characteristics of federally supervised offenders in study sample

Offender characteristics	Descriptive information
Initial PCRA risk classification	
Low	**34%**
Low/Moderate	**44%**
Moderate	**18%**
High	**5%**
Supervised release	**85%**
Offender race and ethnicity	
American Indian or Alaska Native	**2%**
Asian or Pacific Islander	**3%**
Black or African American	**38%**
Hispanic, any race	**19%**
White, not Hispanic	**38%**
Male offender	**82%**
Mean age	**40.1**
Number of offenders	**64,716**

Assessing Change in Offender Risk

The PCRA Scoring Mechanism

Understanding the PCRA scoring mechanism is essential to comprehending how change in risk is measured. Federal probation officers assess an offender's risk of recidivating by scoring offenders on 15 static and dynamic risk predictors. The 15 scored risk predictors can be aggregated into five domains.[3] The first of these involves an offender's criminal history. The criminal history domain is static and includes six risk predictors measuring an offender's prior criminal behavior (AOUSC, 2011). The remaining four PCRA domains assess an offender's dynamic criminogenic characteristics in the areas of education/employment (3 predictors), substance abuse (2 predictors), social networks (3 predictors), and supervision attitudes (1 predictor) (AOUSC, 2011; Johnson et al., 2011; Lowenkamp et al., 2013).

[1] We excluded initial PCRA assessments that occurred after October 2012 because at the time these data files were generated our recidivism measures tracked offenders until October 2013. Obtaining initial PCRA assessments that occurred after October 2012 would not have allowed for sufficient follow-up time between the second PCRA assessment and arrest outcomes.

[2] See Appendix Table 1 comparing the risk characteristics and outcomes of offenders with one versus multiple PCRAs.

[3] This paper only covers changes in the scored PCRA items. For further information about the non-scored PCRA items, see the AOUSC's report that summarizes the PCRA risk tool (AOUSC, 2011).

Of the 15 scored PCRA risk predictors, 13 are assigned values of one, if present, or otherwise zero. The two exceptions are the criminal history factors of prior arrest (3 potential points) and age at intake (2 potential points). In theory, offenders can receive a combined PCRA score ranging from 0 to 18. Of the 18 possible points on the PCRA, nine points appear in the dynamic sections and can be changed. These continuous scores translate into the following four risk categories: low (0-5 points), low/moderate (6-9 points), moderate (10-12 points), or high (13 or more points).[4] These risk categories inform officers about an offender's probability of recidivating and provide guidance on the intensity of supervision that should be directed to a particular offender (AOUSC, 2011; Johnson et al., 2011; Lowenkamp et al., 2013).

How We Measured Change in PCRA Risk Between Two Time Points

In this study, we operationalize changes in an offender's PCRA risk classification through three approaches. First, we explore changes in risk classification by examining the proportion of offenders in each risk category who, at their second assessment, either remained in the same risk category or were reclassified into a higher or lower risk category. Next, we calculate actual point changes in PCRA scores between assessments. Specifically, we subtracted the overall second score from the overall first score to measure how many offenders experienced a one, two, or three or more point increase or decrease in their total score by the next assessment.[5] Last, we explored the percentage of offenders witnessing either a higher or lower score in any of the dynamic domains of education/employment, substance abuse, social networks, or supervision attitudes. Through these approaches, we explore the extent to which change in risk is associated with higher or lower recidivism outcomes.

Measuring Recidivism Outcomes

Recidivism is our primary outcome measure. Recidivism is defined as an arrest for either a felony or misdemeanor offense within one year after the second assessment date. We standardized the follow-up times by tracking only those offenders whose arrest behavior could be observed for 12 months or more after the second assessment. The arrest event was counted only if they were arrested within 12 months after their second PCRA. This standardization resulted in the study sample being reduced from 64,716 to 32,647 offenders. Tracking the study sample within the same uniform time frame allows us to overcome a problem inherent in many recidivism studies where some offenders are followed for longer time periods than others.[6]

Analytical Objectives

By measuring change in offender risk and analyzing the relationship between changes in risk and arrest outcomes, we can address the following research issues.

- What percent of offenders are reclassified from a higher to lower PCRA risk category between assessments or vice versa, and what is the relationship between changes in risk categories and rearrest?
- How many offenders experience a 1, 2, or 3 or more point increase or decrease in their total PCRA scores between assessments, and to what extent are changes in the total PCRA risk scores associated with rearrest?
- Which of the dynamic PCRA domains are most amenable to change, and how is rearrest related to increased or decreased domain scores? For example, does getting a job reduce the probability of arrest to the same extent as obtaining support from a network of prosocial friends or mentors?

Results

Changes in PCRA Risk Classifications, Overall Risk Scores, and Domains

Figure 1 depicts the percent of offenders moving from one risk classification to another between their first and second PCRA assessments by initial risk classification. This figure indicates that many high-risk offenders improve by moving to a lower-risk level by their next assessment. Among offenders initially classified as high risk, 38 percent moved to a lower-risk category in their second assessment; moreover, 27 percent of moderate-risk offenders were reclassified into a lower-risk group at their second assessment. Although not shown, most offenders reclassified to a lower risk level move down only one level (e.g., high to moderate risk). Ninety-two percent of the low-risk offenders and 84 percent of the low/moderate risk offenders demonstrated stability in risk (no change). Further, only seven percent of the low/moderate risk offenders demonstrated a reduction in risk.

Figure 2 focuses on changes in the overall PCRA risk scores and analyzes these changes by an offender's initial risk classification. Unlike Figure 1, this figure shows the percentage of offenders with a 1, 2, or 3 or more point increase or decrease in their total risk scores. In general, the total scores improved the most for high- and moderate-risk offenders. For example, 50 percent of high- and 41 percent of moderate-risk offenders saw reductions by 1 or more points between assessments. Smaller percentages of low-moderate and low-risk offenders have reductions of a point or more in their scores at 25 percent and 13 percent, respectively.[7] The percentage of offenders with increasing scores did not differ as much among the risk categories. For example, the percent of offenders with increasing scores ranged from 17 percent for high-risk to 22 percent for low-moderate and moderate-risk offenders.

Figure 3 presents information on the percentage of offenders with an increase, decrease, or stable score for each of the PCRA domains. Information on the fluctuations in domain scores is analyzed by the offender's initial risk classification. This figure shows the domain of education/employment being the most amenable to change. This was especially the case for offenders in the high-risk category. For instance, 35 percent of high-risk offenders witnessed improvements in their education/employment scores, while 24 percent and 21 percent saw improvements in their substance abuse and social network scores. Similar to the high-risk population,

---

[4] We note that the PCRA is currently undergoing a revision which will involve the integration of a violence assessment into the instrument and result in offenders being placed into 12 different risk groups. At the time of this study, the revised PCRA had not yet been implemented; hence, we continue anchoring our offender population into the four risk groups discussed above.

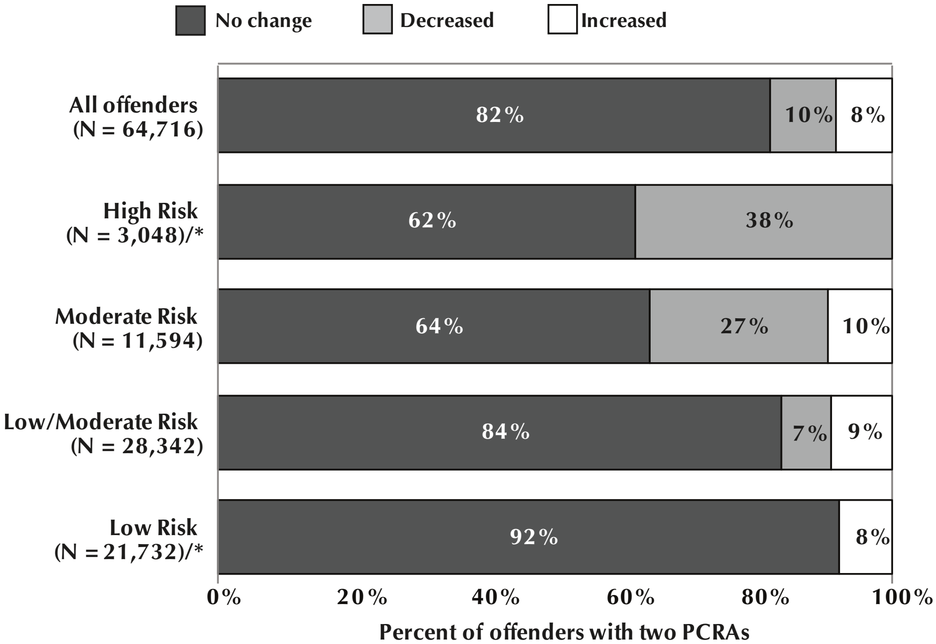
[5] Changes in PCRA scores above or below +/– 4 points were recoded into +/-3 points, as relatively few offenders saw their PCRA scores increase or decrease by 4 or more points.

[6] Although we were unable to track the reoffending behavior for about half of the 64,716 offenders with at least two PCRA assessments, we compared the PCRA risk factors for both groups of offenders using cross tabulations and chi-square tests and for the most part, found negligible differences in their risk characteristics.

[7] The percentage of offenders demonstrating a one point or greater reduction in risk is calculated by adding up the percentages that demonstrated one, two, or three or more point decreases. For example, 19 percent of high-risk offenders demonstrated a one-point decrease in risk, 14 percent a two-point decrease, and 17 percent a three or more point decrease. Adding these three values together equals 50 percent.

FIGURE 1.

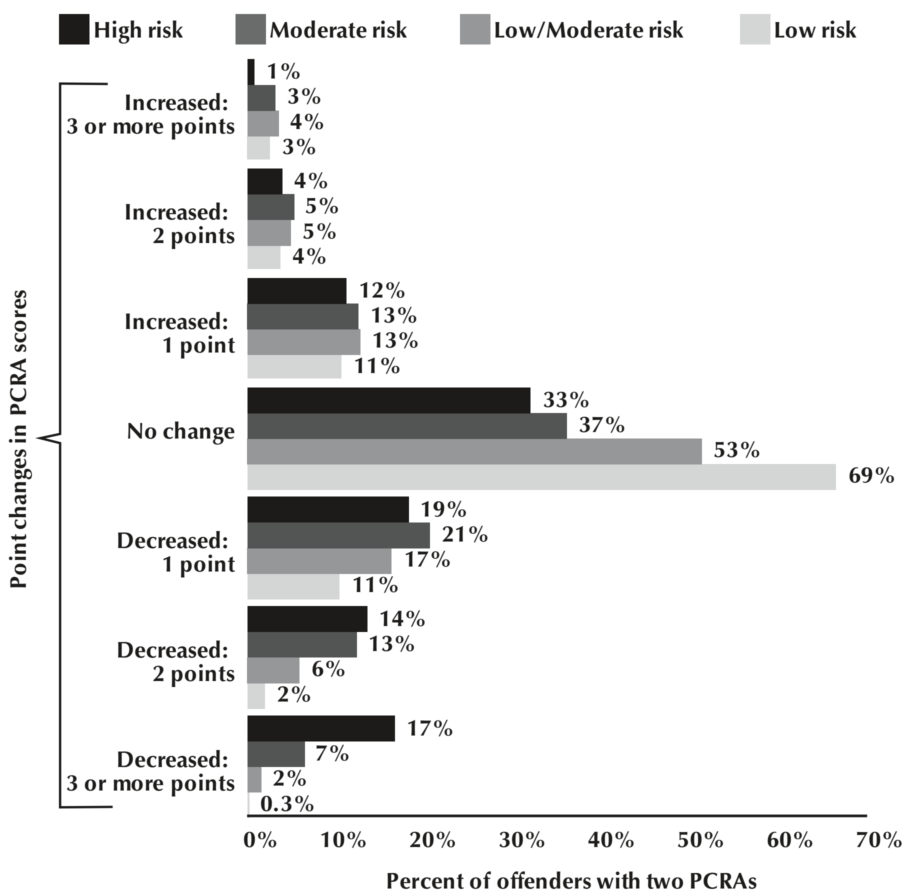
Changes in risk classification levels for offenders with at least two PCRAs, by initial risk classification



Note: *Offenders with the lowest PCRA risk classification cannot receive a decrease in their PCRA risk level and offenders in the highest risk classification cannot receive an increase in their risk level.

FIGURE 2.

Point changes in PCRA scores for offenders, by initial risk classification



moderate-risk offenders witnessed the most change in their education/employment scores. Thirty percent of moderate-risk offenders recorded improvement (decreases) in the education/employment score; in comparison, the percentage of moderate-risk offenders with improvements in any of the other domains did not exceed 15 percent.

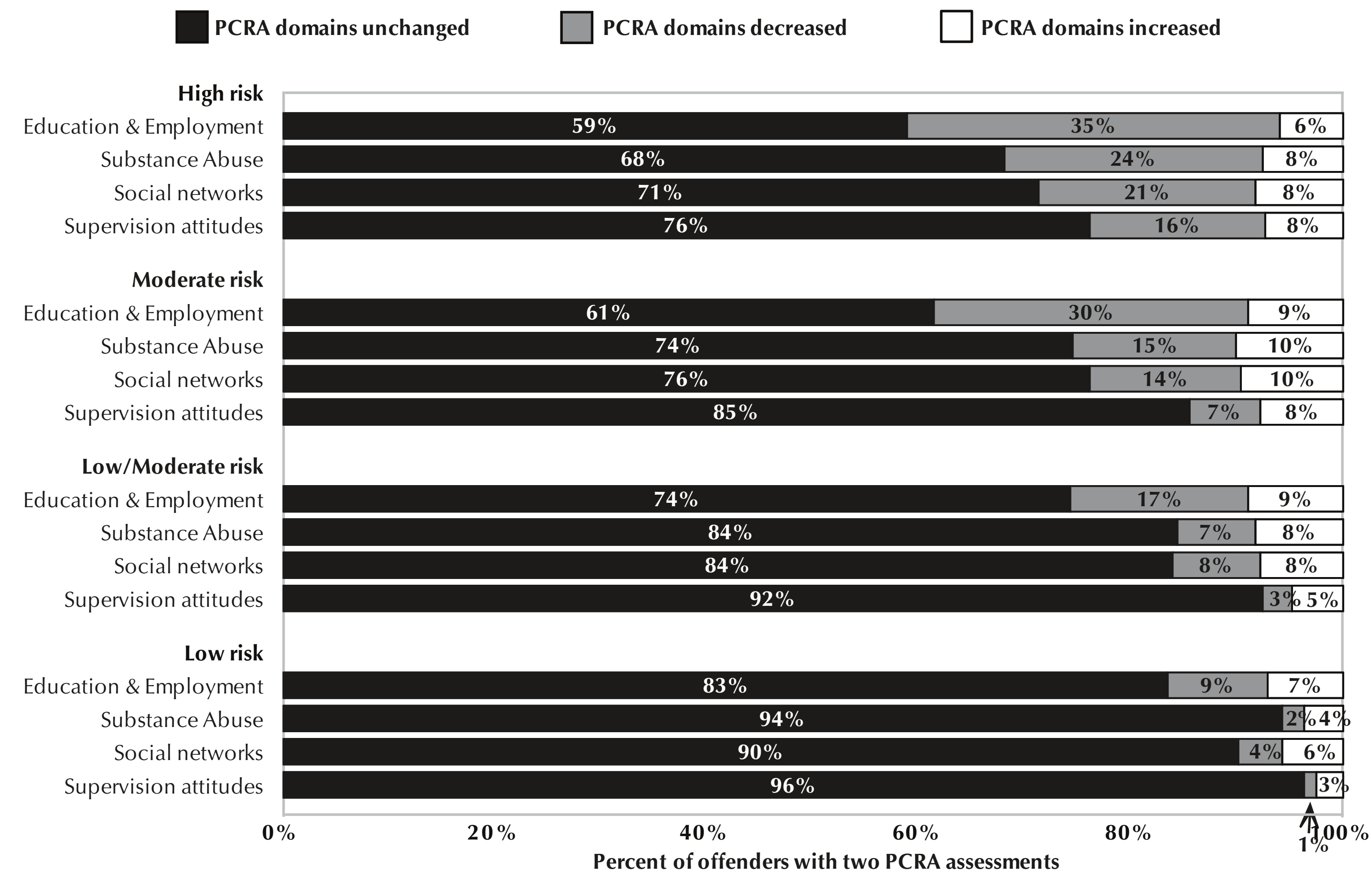
Relationship between changes in risk classification and recidivism

Figure 4 examines the relationship between changes in risk classification and arrest outcomes. Offenders with reduced risk levels were less likely to be arrested compared to offenders whose risk classifications remained unchanged or increased.[8] High-risk offenders who remained in the same risk category, for example, were one and a half times more likely to be arrested for felony or misdemeanor offenses (49 percent) compared to high-risk offenders with lowered risk classifications (33 percent). Among moderate-risk offenders, 49 percent were arrested if their risk classification increased and 30 percent had an arrest if their risk classification remained unchanged; however, for those moderate-risk offenders with a decrease in their risk levels, 18 percent were arrested for a new offense. The same pattern of reduced risk levels being associated with lower arrest rates and increasing risk classifications being associated with higher arrest rates also held for low-moderate and low-risk offenders.

The relationship between changes in the total scores—intra-risk category—and arrest outcomes is investigated in Figure 5. One major finding for high- and moderate-risk offenders is that larger decreases in risk scores were associated with more substantial declines in the likelihood of arrest compared to smaller decreases. For example, high-risk offenders with a reduction in risk of 3 or more points had a lower arrest rate (28 percent arrested) than high-risk offenders with a 1 point reduction in their total risk score (44 percent arrested). In fact, moderate- and high-risk offenders with 1 point reductions in their total scores were arrested at rates that were relatively similar to their counterparts whose scores were unchanged between the assessment periods. Another finding involves the interplay between reduced scores and arrest rates for low-moderate and low-risk offenders. Reductions in the risk score

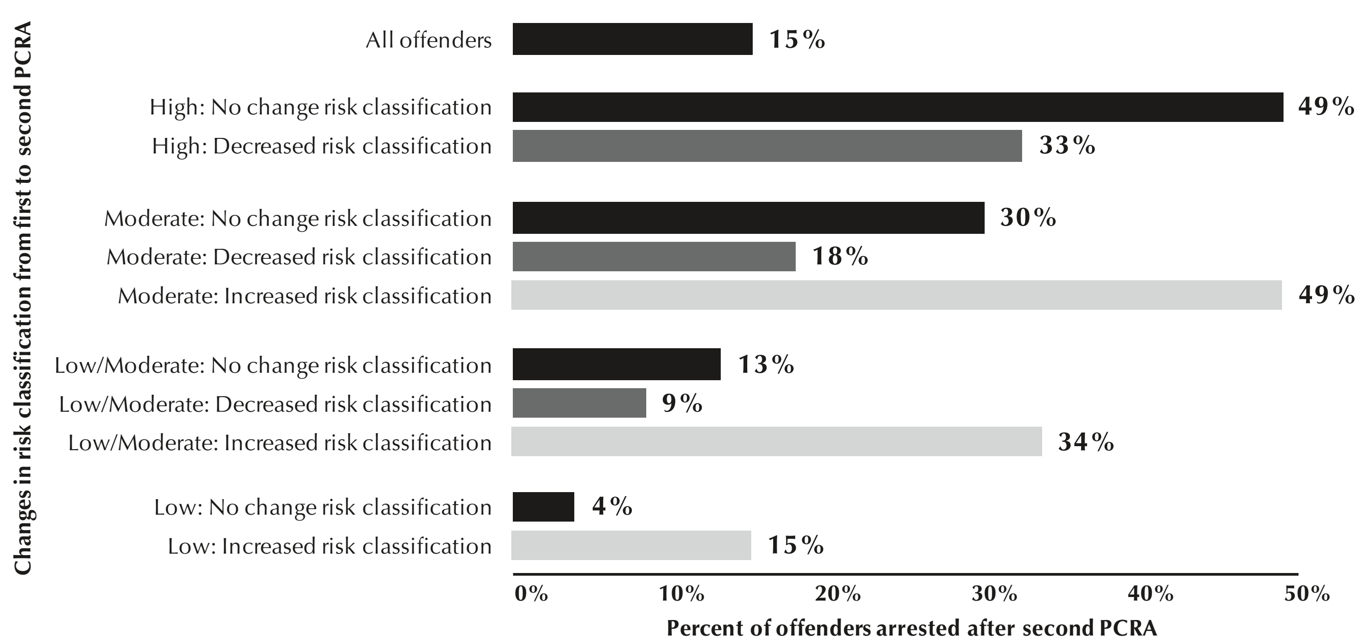
[8] For the recidivism section of this paper (Figures 4, 5, and 6), offenders were counted as arrested if they received new arrests for felony or misdemeanor offenses within 12 months of the second PCRA assessment.

FIGURE 3.
Changes in individual PCRA domains for offenders between first and second assessments, by initial risk classification



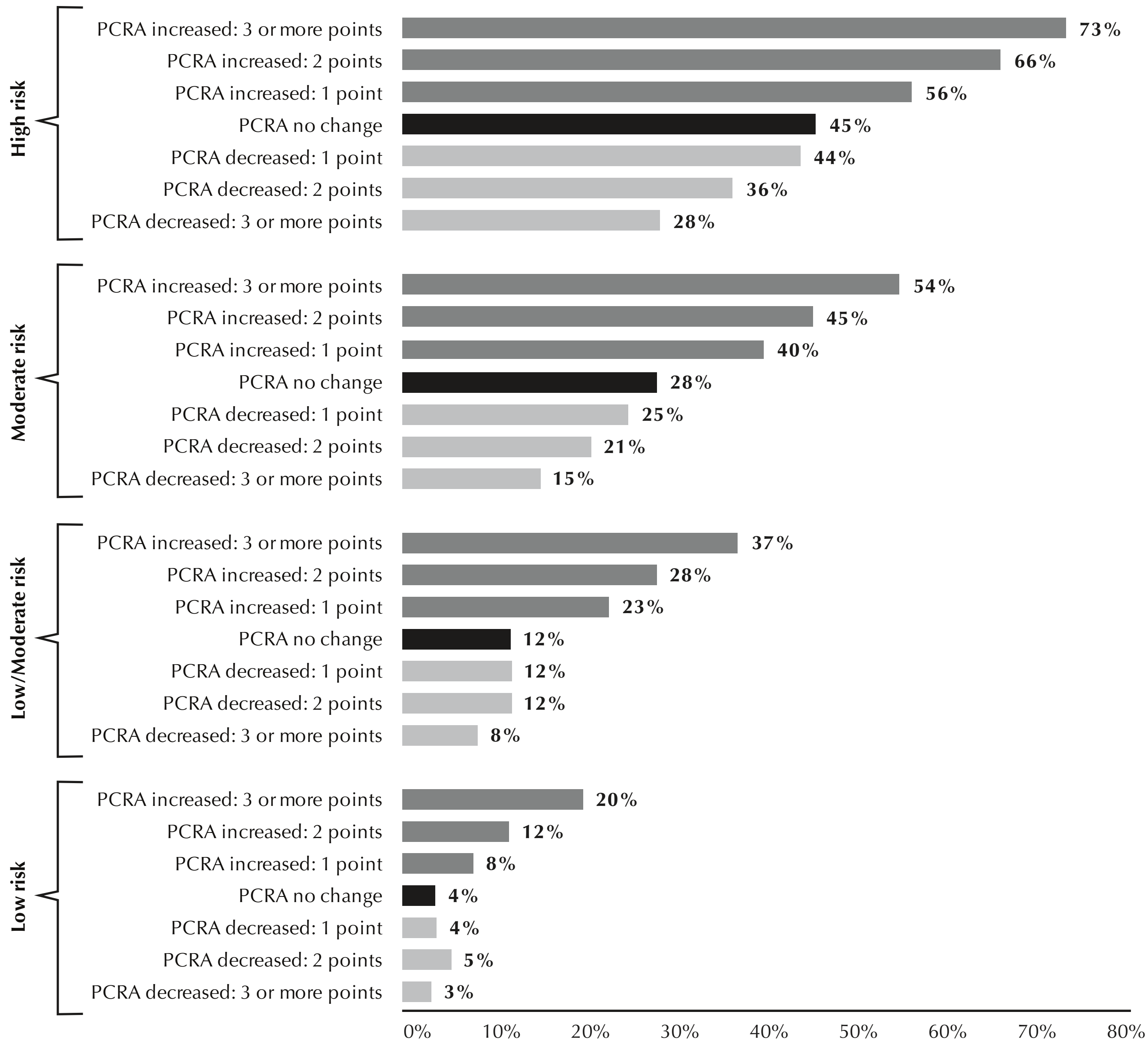
Note: Changes in criminal history scores not shown.

FIGURE 4.
Relationship between changes in PCRA categories and offender arrest outcomes, by initial risk classification



Note: Figure tracks a subset of offenders followed for at least one year after their second PCRA.
Changes represent re-classification of offenders into different risk categories.

FIGURE 5.
Relationship between changes in PCRA scores and arrest outcomes, by initial risk classification



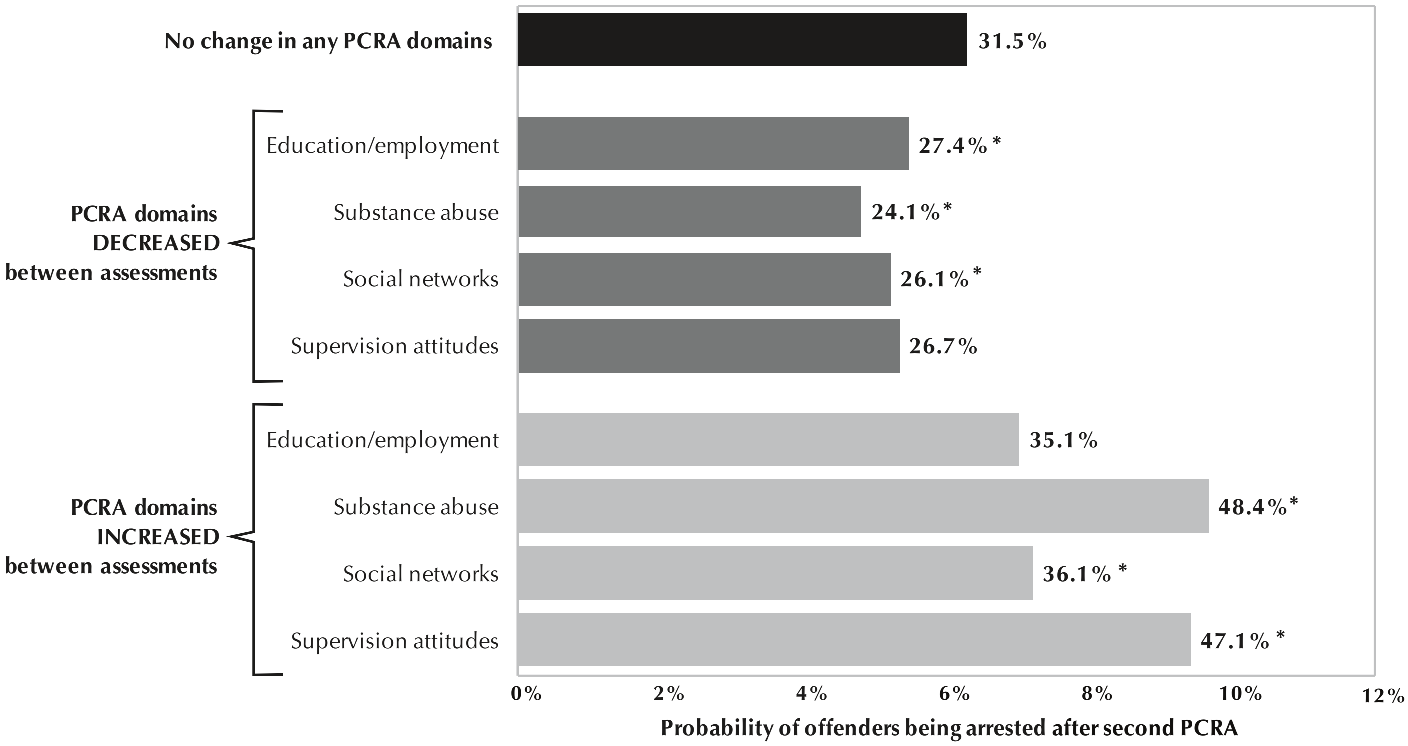
for low and low-moderate risk offenders were not consistently associated with appreciable reductions in arrest rates. This was especially the case for low-risk offenders, whose arrest rates were essentially the same regardless of whether the overall PCRA score improved by 1, 2, or 3 or more points.[9]

Increasing risk scores were associated with higher arrest rates across risk categories. For example, low-risk offenders with a 3 or more point increase in their score had an arrest rate that was almost double that of low-risk offenders with a two-point increase in risk. Finally, it is important to acknowledge that even a one-point increase of the PCRA score was associated with substantial increases in the likelihood of arrest throughout the risk continuum.

A final component of this analysis examines the relationship between offenders with increasing or decreasing PCRA domain scores and rearrests. We examine this by calculating the predicted probabilities of arrest within 12 months after the second assessment for male offenders in the combined high/moderate-risk categories (see Figure 6) and in the low-risk category (see Figure 7). These predicted probabilities were generated through a statistical technique (logistic regression) that allows us to examine the relationship between changes in the individual PCRA domains and recidivism while holding constant other factors that might be correlated with arrest outcomes. For example, we can use this approach to explore the individual contribution of decreased substance abuse scores to recidivism reduction while keeping the other domains unchanged and controlling for other factors such as initial PCRA baseline scores and race/ethnicity. In the predicted probability analysis, we compare arrest probabilities for offenders with increased or decreased scores to

[9] Subsequent regression analyses showed no statistically significant differences between the odds of arrest for high- and moderate-risk offenders with unchanging vs. one-point reductions in their PCRA scores. Offenders with improving PCRA scores of two or more points, however, were significantly less likely to be arrested compared to offenders with no changes in their PCRA scores.

FIGURE 6.
Predicted probability of arrest for all male high and moderate risk offenders with increased or decreased PCRA domain scores



Significant differences are noted by an asterisk.
Note: Figure only shows variation in predicted probability of arrest by changes in the PCRA domain scores.
Other variables in model not shown.
* p < .05

offenders with no changes in their scores.

Figure 6 shows that 32 percent of high/moderate-risk male offenders with no changes in their PCRA domains were predicted to have an arrest within 12 months of their PCRA re-assessment. In comparison, high/moderate-risk offenders with decreased domain scores, for the most part, were significantly less likely have a new arrest. For example, high/moderate-risk male offenders with decreased education/employment, substance abuse, and social network scores had an arrest likelihood ranging from 24 percent to 27 percent.[10] Since the predicted arrest probabilities associated with improvement in education/employment, substance abuse, and social networks were relatively similar, one cannot discern that decreases in one domain resulted in greater reductions in the likelihood of arrest than decreases in another domain.

Increased substance abuse and supervision attitude scores were more closely related to an offender's arrest probability than increased education/employment and social network scores. For instance, nearly half of high/moderate-risk offenders with worsening substance abuse (48 percent arrest probability) or supervision attitude (47 percent arrest probability) scores were predicted to be arrested within 12 months after the second PCRA assessment. Among high/moderate-risk offenders with job losses or weakening social networks, arrest probabilities were 35 percent and 36 percent, respectively. [11]

Figure 7 shows the predicted probabilities of arrest for low-risk male offenders. Unlike higher-risk offenders, low-risk offenders with improving PCRA scores did not witness significant reductions in their arrest probabilities. For example, the predicted probability of low-risk male offenders with decreased domain scores being rearrested was about 3 percent. Low-risk male offenders with no changes in their PCRA domains, in comparison, had a predicted arrest probability of 4 percent. For low-risk offenders with worsening PCRA domain scores, deteriorations in substance abuse or supervision attitudes resulted in higher arrest probabilities than those of offenders with increasing education/employment and social network scores.

Policy Implications

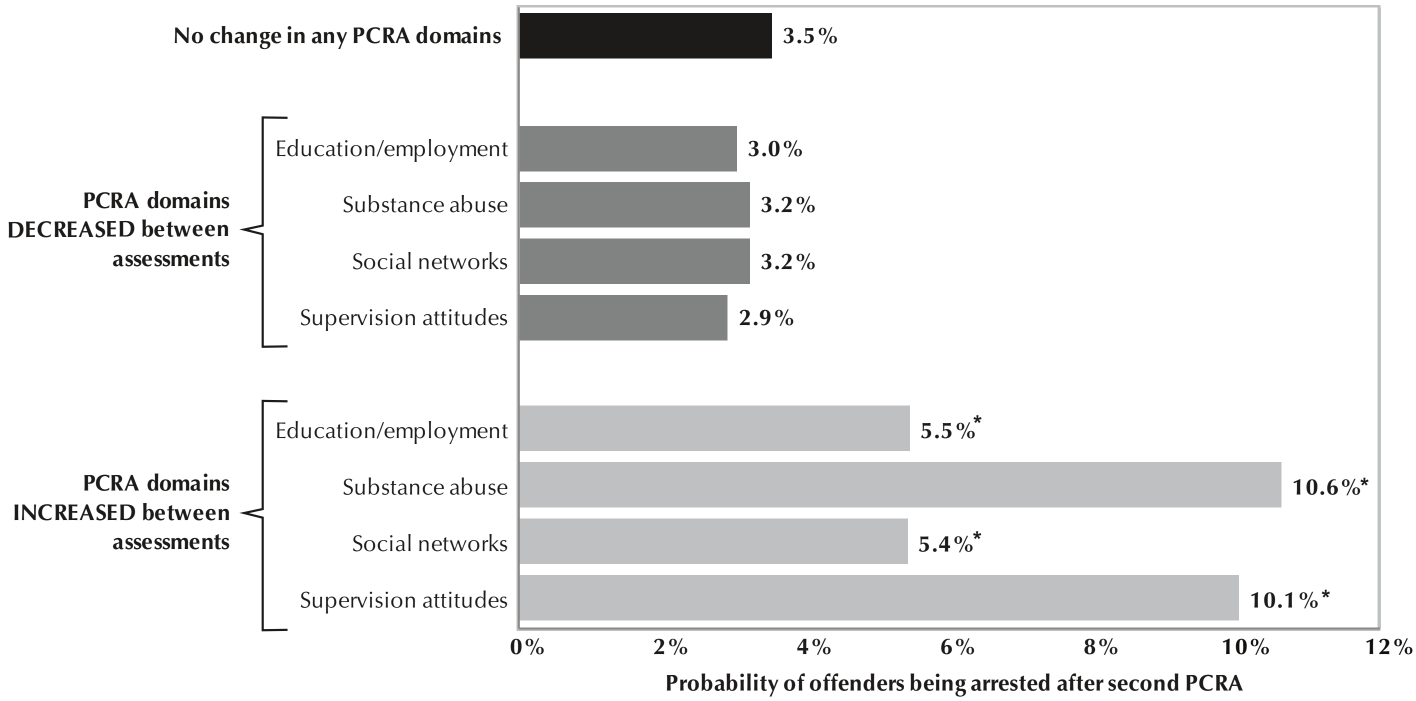
This analysis provides officers with information about how changes in offender risk levels can influence the likelihood of arrest. It clearly shows that low-moderate, moderate-, and high-risk offenders on federal supervision with decreased risk classifications were less likely to recidivate compared to their counterparts whose risk level either remained unchanged or increased. Conversely, higher recidivism rates were associated with increases in offender risk across all risk categories. These findings are consistent with the risk principle of the risk, needs, and responsivity (RNR) model that suggests officers reduce the intensity of supervision services to offenders with decreasing risk levels once those decreases have stabilized (Andrews & Bonta, 2010). Alternatively, probation officers should pay closer attention and intensify supervision

[10] While improving supervision attitude scores were also associated with reduced arrest probabilities, the effect was not significantly different compared to offenders with no changes in their PCRA domain scores.

[11] Although not shown, we found somewhat similar patterns between improving and worsening PCRA domain scores and recidivism outcomes for low-moderate risk offenders. The only notable differences were that improving education/employment scores had no significant relationship to arrest, while improving supervision attitude scores were significantly related to arrest outcomes for these offenders.

FIGURE 7.

Predicted probability of arrest for all male low risk offenders under federal supervision with increased or decreased PCRA domain scores



Significant differences are noted by an asterisk.
Note: Figure only shows variation in predicted probability of arrest by changes in the PCRA domain scores. Other variables in model not shown.
* $p < .05$

services for those offenders reclassified into higher risk levels.

We also show that offenders in the high- and moderate-risk categories were less likely to be rearrested if they demonstrated improvements in their substance abuse, social networks, or education/employment domains, while offenders in the low-moderate risk category were arrested less frequently when their substance abuse, social networks, or supervision attitude scores improved. Based on these findings, we cannot make any recommendations on which PCRA domain to target first for intervention. Our research suggests that ameliorating any existing domain should reduce recidivism and that decisions about what should be targeted first should be individualized to the offender. Our findings also suggest that the lowering of an offender's overall PCRA score by several points reduces the likelihood of recidivism to a greater extent than a one-point reduction.

For offenders with increasing PCRA scores, we show that increasing risk scores of any magnitude were related to higher arrest likelihoods. Moreover, the most significant increases in recidivism occurred for offenders with higher substance abuse and supervision attitude scores. These findings suggest that probation officers consider paying close attention to offenders with any increases in their overall PCRA scores, with particular emphasis on those whose substance abuse or supervision attitudes showed signs of worsening.

Finally, the lowest-risk offenders did not benefit from reductions in their domain scores. To reiterate, decreasing PCRA domain scores were not associated with reduced arrest probabilities for offenders in the lowest risk category. This finding is highly consistent with the risk principle, which advocates expending time and resources on the highest-risk offenders (Andrews, Bonta, & Hoge, 1990). Specifically, probation officers should carefully consider whether to provide resources and services to low-risk offenders who do not seem to benefit from efforts aimed at reducing their criminal risk factors (Vose, Smith, & Cullen, 2013). At the same time, these findings indicate that officers should monitor low-risk offenders and respond accordingly if increases in risk are seen.

References

Administrative Office of the U.S. Courts. (AOUSC) (2011). *An overview of the federal Post Conviction Risk Assessment*. Washington, DC: Administrative Office of the U.S. Courts.

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

Andrews, D., & Bonta, J. (2010). *The psychology of criminal conduct (5th Ed.)*. Cincinnati, OH: Anderson Publishing.

Cohen, T., & VanBenschoten, S. (2014). Does the risk of recidivism for supervised offenders improve over time?: Examining changes in the dynamic risk characteristics for offenders under federal supervision. *Federal Probation*, *78* (2), 15-21.

Cohen, T., Lowenkamp, C., & VanBenschoten, S. (2016). Does change in risk matter?: Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology and Public Policy*, *15*(2) 263-296.

Howard, P., & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism: Casual dynamic risk factors in the OASys violence predictor. *Law and Human Behavior*, *37*(3), 163-174.

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

Lowenkamp, C., Johnson, J., VanBenschoten, S., Robinson, C., & Holsinger, A. (2013). The Federal Post Conviction Risk Assessment

(PCRA): A construction and validation study. *Psychological Services*, *10*(1), 87-96.

Lowenkamp, C., Holsinger, A., & Cohen, T. (2015). PCRA revisited: Testing the validity of the federal Post Conviction Risk Assessment (PCRA). *Psychological Services*, *12*(2), 149-157.

Vose, B., Smith, P., & Cullen, F. (2013). Predictive validity and the impact of change in total LSI-R score on recidivism. *Criminal Justice and Behavior*, *40*(12), 1383-1396.

APPENDIX TABLE 1.

Comparing scored PCRA characteristics and case outcomes for offenders placed on federal supervision with one vs. multiple PCRAs, by initial risk classification

	High risk		Moderate risk		Low/Moderate risk		Low risk	
Descriptive statistics	**Two PCRAs**	**One PCRA**	**Two PCRAs**	**One PCRA**	**Two PCRAs**	**One PCRA**	**Two PCRAs**	**One PCRA**
Disposition after first PCRA								
Case still open	50%	22%	62%	29%	66%	31%	65%	37%
Successful termination	11%	18%	16%	34%	24%	56%	31%	61%
Revocation	39%	60%	22%	38%	10%	13%	3%	2%
Mean initial PCRA scores								
Criminal History	7.33	7.32	6.55	6.50*	5.02	5.01	1.81	1.74*
Education & Employment	2.52	2.52	1.87	1.84*	1.06	1.03*	0.55	0.50*
Substance Abuse	1.11	1.05*	0.56	0.54*	0.22	0.19*	0.07	0.04*
Social networks	2.24	2.34*	1.58	1.66*	1.10	1.10	0.72	0.66*
Supervision attitudes	0.52	0.63*	0.21	0.28*	0.08	0.09*	0.04	0.03*
Number of offenders	3,048	2,066	11,594	5,955	28,342	14,887	21,732	20,130

Note: *T-test of mean differences denotes significant difference at the .05 level.

CONTRIBUTORS *To This Issue*

Kristin Bechtel

Managing Associate, Crime and Justice Institute. Ph.D., University of Cincinnati. Co-author of "Trademarks, Press Releases, and Policy: Will Rigorous Research Get in the Way?" *Federal Probation* (June 2016).

John Bentley

Chief Probation and Pretrial Services Officer, U.S. District Court, District of South Dakota. Master's degree, Loyola University, Chicago.

Thomas H. Cohen

Social Science Analyst, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. Previsously, Statistician, Bureau of Justice Statistics. Ph.D., Rutgers University School of Criminal Justice. J.D., University of Maryland School of Law. Co-author of "The Supervision of Low-risk Offenders: How the Low-risk Policy Has Changed Federal Supervision Practices without Compromising Community Safety," *Federal Probation* (June 2016).

James Doyle

Of counsel to the Boston firm of Bassil, Klovee & Budreau. Former Visiting Fellow at the National Institute of Justice. B.A., Trinity College; J.D., Northwestern University School of Law; LL.M, Georgetown University Law Center. Author of "Learning From Error in American Criminal Justice," *Journal of Criminal Law and Criminology* (2010); "Learning from Error in the Criminal Justice System: Sentinel Event Reviews" in *Special Report, Mending Justice: Sentinel Event Reviews* (NIJ 2014).

Anthony W. Flores

Assistant Professor, California State University, Bakersfield. Ph.D., University of Cincinnati. Author of "The Importance of Measuring Offender Change," *Criminology and Public Policy,15*(2), 1-3 (April 2016).

Nancy Beatty Gregoire

Deputy Chief, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. M.A., American University. Author of "Federal Probation Joins the World of DNA Collection," *Federal Probation* (June 2002).

James L. Johnson

Program Administrator, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. Ph.D., American University. Co-author of "State of the System: Federal Probation and Pretrial Services," *Federal Probation* (September 2015).

Cynthia A. Kempinen

Retired Deputy Director, Pennsylvania Commission on Sentencing. Ph.D., Penn State University. Co-author of "How Long After? A Natural Experiment Assessing the Impact of the Length of Aftercare Service Delivery on Recidivism" (2011).

Lauren K. Knoth

Graduate Student, Penn State University. M.S., Penn State University.

Christopher T. Lowenkamp

Social Science Analyst, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. Ph.D., University of Cincinnati. Co-author of "Trademarks, Press Releases, and Policy: Will Rigorous Research Get in the Way?" *Federal Probation* (June 2016).

R. Barry Ruback

Professor of Criminology and Sociology, Penn State University. J.D., University of Texas. Ph.D., University of Pittsburgh.

Ralph Serin

Professor, Carleton University, Ottawa. Ph.D., Queen's University.

Janette Sheil

Chief, Program Oversight Branch, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. Co-author of "Accountability and Collaboration: Strengthening Our System Through Office Reviews," *Federal Probation* (December 2015).

Michelle Spidell

Probation and Pretrial Services Administrator, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. Master's degree from Creighton University.

Leigh A. Tinik

Research Project Manager, Pennsylvania Commission on Sentencing. M.A., Penn State University. Co-author of "Disproportional Imprisonment of Black and Hispanic Males: Sentencing Discretion, Processing Outcomes, and Policy Structures," *Justice Quarterly, 33*(4) (2014).

Patricia Trevino

Probation Administrator, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. M.A. in Theoretical and Behavioral Counseling, Wayne State University.

Scott W. VanBenschoten

Chief, Program Development Branch, Probation and Pretrial Services Office, Administrative Office of the U.S. Courts. M.S.W., Virginia Commonwealth University. M.P.A., The American University. Co-author of "Does the Risk of Recidivism for Supervised Offenders Improve Over Time? Examining Changes in the Dynamic Risk Characteristics for Offenders under Federal Supervision," *Federal Probation* (Sept. 2014).

United States Government

INFORMATION

Order Processing Code

***5876**

Federal PROBATION

a journal of correctional philosophy and practice

Please send me _____ subscription(s) to *Federal Probation* at $16.50 each ($22.40 foreign) per year. The total cost of my order is $______________.

Price includes shipping and handling and is subject to change.

____

(PLEASE TYPE OR PRINT) NAME OR TITLE

____

COMPANY NAME ROOM, FLOOR, OR SUITE

____

STREET ADDRESS

____

CITY STATE ZIP CODE + 4

____

DAYTIME PHONE INCLUDING AREA CODE

____

PURCHASE ORDER NUMBER (OPTIONAL)

Method of Payment

❑ Check payable to: Superintendent of Documents

❑ GPO Deposit Account ______________________________

❑ Credit Card ❑ Visa ❑ MasterCard ❑ Discover

____

CARD NUMBER EXPIRATION DATE

____

AUTHORIZED SIGNATURE

ORDER ONLINE FROM

http://bookstore.gpo.gov

FAX YOUR ORDER TO

202-512-2104

PHONE YOUR ORDER TO

202-512-1800 or
866-512-1800

MAIL YOUR ORDER TO

Superintendent of Documents
P.O. Box 979050
St. Louis, MO 63197-9000

IMPORTANT!

Please include your completed order form with your remittance

thank you for your order!

FEDERAL PROBATION

Administrative Office
of the United States Courts

Washington, DC 20544