Are Pretrial Services Officers Reliable in Rating Pretrial Risk Assessment Tools?[1]

Patrick J. Kennealy
Travis County Community Justice Services

THERE WERE OVER 10 million arrests in the United States during 2016 (Federal Bureau of Investigation, 2017). Once arrested, the decision to release or detain the accused pending trial is made by the court. This bail decision is typically made by weighing the risk of failure to appear at future court dates, the likelihood of new arrests prior to the disposition of the case, and other considerations (VanNostrand, 2007). In jurisdictions where available, pretrial services agencies assist the court throughout this process. Foremost among the responsibilities of pretrial services agencies is the collection of information to inform the bail decision and the decision on whether release conditions such as curfew, electronic monitoring, or alcohol and drug testing are necessary (VanNostrand, Rose, & Weibrecht, 2011).

The importance of the collection of information to inform the bail decision is highlighted by the consequences of pretrial detention on defendants (Farnworth & Horan, 1980; Ulmer, 2012; Bechtel, Holsinger, Lowenkamp, & Warren, 2016). For example, Demuth (2003) reported that pretrial detention harms the defendants' capacity to maintain employment, meet family obligations, and participate in the development and execution of their legal defense. Further, in a study of over 90,000 federal defendants, Oleson and colleagues (2017) investigated the association between pretrial detention and sentencing outcomes when controlling for the seriousness of the offense and criminal history. They reported that release pending trial was associated with less serious sentences, whereas pretrial detention was linked to more serious sentences. Similar findings were reported by Oleson, Lowenkamp, Cadigan, VanNostrand, and Wooldredge (2016) in a sample of 1,723 United States federal court cases.

Even when defendants are released rather than detained, unnecessary conditions of release can be harmful (Cadigan & Lowenkamp, 2011). This is particularly true of defendants who present with a low level of risk for pretrial supervision failure. In a study of federal defendants, VanNostrand and Keebler (2009) found that requiring location monitoring as a condition of release for low-risk defendants resulted in a 112 percent increase in the likelihood of pretrial supervision failure relative to low-risk defendants without this condition. In light of these consequences, there have been growing efforts to ensure that decisions on pretrial release and detention are fair and consistent.

One of the common ways to improve the pretrial recommendation-making process has been the development and use of pretrial risk assessment instruments. These instruments are designed to assess the risk that defendants will a) fail to appear in court or b) be arrested for new criminal activity if released from custody pending trial (VanNostrand & Keebler, 2007). Pretrial risk assessment instruments typically measure a combination of static (i.e., unchanging) and dynamic (i.e., changeable) risk factors. Static items often include the current charge, criminal history (e.g., previous arrests and incarcerations), and previous failures to appear, while the dynamic items may focus on employment, residential stability, ties to the community, and drug use (VanNostrand & Lowenkamp, 2013; Bechtel, Lowenkamp, & Holsinger, 2011). Depending on the given instrument, the tool is completed based on an interview with the defendant, a review of file information, or a combination of both.

Although the use of pretrial risk assessment instruments has been linked to more release recommendations and lower jail populations (Cooprider, 2009), the basic use of pretrial risk assessment instruments alone does not guarantee these benefits (Mamalian, 2011). The utility of the tool is dependent on implementation, as with any risk assessment instrument in criminal justice contexts (Mamalian, 2011; Latessa & Lovins, 2010). One essential aspect of the implementation of a risk assessment tool is the demonstration of inter-rater reliability (Bechtel, Holsinger, Lowenkamp, & Warren, 2016). This is the degree to which two raters agree on the rating of a given case when provided with the same information.

A lack of reliability can have devastating consequences for defendants and the broader community. When unreliably rated, the recommendation to release a defendant on bond could vary widely as a function of the pretrial services officer who performed the assessment. For example, if mistakenly rated as high risk, a low-risk defendant may not be recommended for release on bond. Alternatively, a high-risk defendant misclassified as low risk could be released on bond. This would

[1] Acknowledgments: This article expresses the views of the author and not necessarily the views of the organization with which he is affiliated. The author would like to thank Kasey Wada, Fernando Romero, Gerald Rodriguez, Stacy Brown, Daniel McCoy-Bae, and Rodolpho Pérez, Jr. for their support and assistance in conducting this research.

be a misallocation of resources in the best-case scenario, but potentially harmful for the defendant and his or her family or the community in the worst-case scenario. Therefore, inter-rater reliability is a prerequisite of the use of risk assessment instruments as an evidence-based practice (Latessa & Lovins, 2010).

Despite the demonstrated consequences of detainment on defendants (Oleson et al., 2017; Oleson et al., 2016; Ulmer, 2012), there is a dearth of published research on the inter-rater reliability of pretrial risk assessment tools. In fact, reviews of the pretrial risk assessment literature failed to find a single study that reported the inter-rater reliability of pretrial services officers in scoring such instruments (Bechtel, Holsinger, Lowenkamp, & Warren, 2016). To help address this need, the present study investigates the inter-rater reliability of pretrial services officers in rating a pretrial risk assessment tool. Specifically, we assess the inter-rater reliability of the Ohio Risk Assessment System-Pretrial Assessment Tool (ORAS-PAT; Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009) at the item-, total-, and summary risk classification-level. Findings of this study offer implications for the use of risk assessment tools in pretrial services to inform jail release decisions.

Method

Study Design

To investigate the inter-rater reliability of the ORAS-PAT, we identified all pretrial services officers who regularly rate defendants on this measure in a single county agency. Next, we ascertained a list of all cases (n = 3,445) rated by these 21 pretrial services officers during a two-month period (i.e., September and October of 2017). With this list, we randomly selected five cases rated by each of these 21 pretrial services officers, resulting in a total of 105 cases. However, one case file could not be located and another file was missing the ORAS-PAT scoring form. This left a total of 103 cases with complete information for use in this study.

In turn, two pretrial services supervisors were tasked with performing secondary ratings on the ORAS-PAT for these 103 cases. One of these supervisors was the agency's lead trainer for the ORAS-PAT and the other frequently performed audits on ORAS-PAT ratings. Additionally, both of these supervisors have regularly performed ORAS-PAT ratings in their time with the agency. These secondary ratings were performed with notes from the original semi-structured interview along with a review of relevant information from the defendant's official file.

Participants

Primary raters were 21 pretrial services officers from a county pretrial services agency in a large southwestern state. The pretrial services officers all had at least a bachelor-level degree. This group was mostly female (66.7 percent) and Hispanic (61.9 percent; White, 23.8 percent; African-American, 9.5 percent; Other, 4.8 percent). On average, the pretrial services officers were approximately 33 years old (M = 32.8, SD = 11.3) and had worked for the agency 5.6 years (SD = 9.2). Each of these pretrial services officers rated about 37 defendants (M = 36.6, SD = 46.5) on the ORAS-PAT a month.

The ORAS-PAT ratings were completed by the pretrial services officers on 103 defendants. The defendants were mostly male (79.6 percent) and approximately 33 years old (M = 33.2, SD = 12.9). Further, this group was 35.9 percent White, 32.0 percent Hispanic, 30.1 percent African-American, and 1.9 percent Asian-Pacific Islander.

Measure

The Ohio Risk Assessment System-Pretrial Assessment Tool (ORAS-PAT; Latessa et al., 2009) is a pretrial risk assessment instrument that was developed to inform pretrial release decisions. The instrument features 7 items and is scored based on an interview with the defendant and a review of official file information. Each item is either dichotomous or rated on a 3-point Likert scale. Items assess age at first arrest, history of failure-to-appear warrants and incarcerations, employment status, residential stability, and drug use. Scores on these items are summed to render a total score, which is then converted into a summary risk classification (i.e., low, medium, or high risk). Although there is a lack of examinations of the reliability of the ORAS-PAT, research indicates that the instrument demonstrates predictive utility for criminal justice outcomes (Latessa et al., 2009). Descriptive information, including means and standard deviations for primary and secondary ORAS-PAT ratings, is presented in Table 1.

Analyses

The inter-rater reliability between pretrial services officers and supervisors is estimated with weighted Kappa. This statistic is suitable for use on categorical items and ratings. Kappa identifies the variance in a set of ratings that is due to the cases that were rated instead of measurement error (Cohen, 2001). In other words, Kappa estimates the degree that pretrial services officers and pretrial services supervisors can reliably rate a defendant when controlling for chance agreement. All analyses in this study were performed in STATA 13.1.

To maintain consistency with other research on the reliability of risk assessment instruments (Vincent, Guy, Fusco, & Gershenson, 2012; Kennealy, Skeem, & Hernandez, 2016), we adopt the reliability standards of Cicchetti and Sparrow (1981). "Excellent" values are .75 and greater, "good" values are between .60 and .74, "fair" values are between .40 and .59, and "poor" values are less than .40 (Cicchetti & Sparrow, 1981). Rather than depending solely on these reliability standard labels, we also encourage consideration of the individual Kappa values for full information.

TABLE 1.
Descriptive Information on Primary and Secondary ORAS-PAT Ratings

ORAS-PAT	Primary Rater Mean (SD)	Secondary Rater Mean (SD)
1. Age At First Arrest	.89 (.31)	.90 (.30)
2. Number of Failure-To-Appear Warrants in Past 24 Months	.08 (.33)	.11 (.42)
3. Three or more Prior Jail Incarcerations	.37 (.49)	.38 (.49)
4. Employed at the Time of Arrest	.47 (.75)	.56 (.80)
5. Residential Stability	.37 (.49)	.37 (.49)
6. Illegal Drug Use during Past Six Months	.35 (.48)	.43 (.50)
7. Severe Drug Use Problem	.05 (.22)	.06 (.24)
Total Score	**2.54 (1.58)**	**2.78 (1.78)**
Summary Risk Level	**.44 (.57)**	**.58 (.65)**

Notes: n = 103. SD = standard deviation.

Results

To test the inter-rater reliability of the ORAS-PAT, we treated the secondary ratings of pretrial services supervisors as the criterion and compared them to the item-, total-, and summary risk classification-level scores of pretrial services officers (see Table 2). First, we assessed the inter-rater reliability of the ORAS-PAT at the item-level. Kappa values ranged from .72 to .94 on ORAS-PAT items. These values fall in the "good" to "excellent" ranges of Kappa values (Cicchetti & Sparrow, 1981). In fact, Kappa values fell in the "excellent" range for 6 of 7 ORAS-PAT items. To help contextualize these findings, we also calculated the percent agreement between raters on each of these items. The same rating was obtained by pretrial services officers and pretrial services supervisors over 90 percent of the time for each ORAS-PAT item.

Second, we assessed the inter-rater reliability of the ORAS-PAT total risk score. The sum of all 7 ORAS-PAT items, this score ranges from 0 to 9 and is used to make summary risk classifications based on the instrument's established cut-off scores. Here, a weighted Kappa of .82 (SE = .06) was found between raters, which is considered "excellent" (Cicchetti & Sparrow, 1981). Further, the same total risk score was obtained by pretrial services officers and pretrial services supervisors in 74.5 percent of cases.

Finally, we assessed inter-rater reliability at the summary risk classification. This classification as low, moderate, or high risk has the strongest impact on release recommendations. A "good" (Cicchetti & Sparrow, 1981) weighted Kappa value of .72 (SE = .08) was observed between pretrial services officers and pretrial services supervisors. Of the 21 pretrial services officers, 9 (42.9 percent) had zero disagreements, 6 (28.6 percent) had one disagreement, and 6 (28.6 percent) had two disagreements with the supervisors on the summary risk classification. The exact same summary risk classification was obtained by pretrial services officers and pretrial services supervisors in 83.3 percent of 103 cases.

Discussion

This study is one of the first investigations of the inter-rater reliability of a pretrial risk assessment tool as completed by pretrial services professionals. The key finding of this study is that pretrial services officers can reliably rate defendants on a pretrial risk assessment tool. That is, pretrial services officers generally rated defendants in the same manner as pretrial services supervisors in this study. In fact, "good" to "excellent" inter-rater reliability was observed for all 7 items, the total score, and the summary risk classification of the ORAS-PAT. These findings have important implications for the fidelity of pretrial risk assessment tools in the field of pretrial services.

Foremost, these findings limit concerns that pretrial services officers are not able to rate defendants on these instruments in a reliable manner during the course of their everyday job responsibilities. When a defendant is inaccurately rated, the risk of failure to appear in court or arrest for a new crime may be misrepresented. For example, a defendant who is misclassified as "high" risk may be detained despite presenting minimal risk of not appearing in court or of new arrests. However, the present study demonstrates that it is possible for pretrial services officers to use pretrial risk assessment tools in a reliable manner, which helps alleviate concerns about misclassification and the resulting consequences.

Although the findings are promising, the limitations of this study must be carefully considered. Of most concern, the secondary ORAS-PAT ratings provided by pretrial services supervisors were dependent on information that was collected by the pretrial services officers who performed the primary ORAS-PAT ratings. Specifically, the scoring of items on employment, residential stability, and drug use required information gathered during the semi-structured interview of the defendant. As such, any information that was not recorded in the notes of the semi-structured interview with the defendant would not be available to the pretrial services supervisor. This could potentially result in pretrial services supervisors either a) having less information available to inform their rating or b) duplicating the mistakes the pretrial services officers made during the primary rating. These concerns are somewhat mitigated by the fact that there was a general consistency in Kappa values between the items scored based on the semi-structured interview and the three criminal history items that are coded from official records.

Nonetheless, the field would benefit from additional research using different research designs that have the potential to help address the primary weakness of the present study. For example, one promising option is the development of training reliability cases that consist of either a video recording of a semi-structured interview with a defendant or a vignette, both of which would be presented with excerpts of relevant file information (for an example, see Kennealy et al., 2017). These training cases can then be systematically administered to staff members in a given agency. Although limited in ecological validity, this approach reduces the potential for bias by ensuring that all raters have the exact same information available to them when rating a case. Alternatively, another promising design would be having a secondary rater observe the primary rater's semi-structured interview with a defendant (for an example, see Vincent et al., 2012). Both the primary and secondary raters would have access to the same file information on the defendant. This approach helps avoid the ecological validity issues of training reliability cases that feature video recordings or vignettes, but introduces the possibility of an experimenter effect, because the primary rater is being observed and that may alter his or her

TABLE 2.
Inter-rater Reliability of Pretrial Services Officers on ORAS-PAT

ORAS-PAT	Percent Agreement	Weighted Kappa	Standard Error
1. Age At First Arrest	97.1%	.84	.10
2. Number of Failure-To-Appear Warrants in Past 24 Months	96.1%	.72	.08
3. Three or more Prior Jail Incarcerations	97.1%	.94	.10
4. Employed at the Time of Arrest	92.2%	.84	.08
5. Residential Stability	96.1%	.92	.10
6. Illegal Drug Use during Past Six Months	92.2%	.84	.10
7. Severe Drug Use Problem	99.0%	.90	.10
Total Score	**74.5%**	**.82**	**.06**
Summary Risk Classification	**83.3%**	**.72**	**.08**

Notes: n = 103. All weighted Kappa values are significant at p < .001.

typical actions. Together, a combination of separate studies employing different methodological designs offers the best opportunity to better understand the inter-rater reliability of pretrial risk assessment instruments.

In conclusion, the present study on reliability and previous research on validity (Latessa et al., 2009; Bechtel et al., 2016) show that pretrial risk assessment instruments can help inform bail decision making. However, the benefits of using a pretrial risk assessment instrument (e.g., increased recommendations for release on bail; Cooprider, 2009) can only be achieved when the tool is used with fidelity (Mamalian, 2011; Bechtel et al., 2016; Vincent et al., 2012). This requires a thorough implementation plan that ensures the ongoing training and support of staff members along with quality assurance (Bechtel et al., 2017). Evaluations of quality assurance should include assessments of the instrument's a) reliability and b) predictive utility for important pretrial-related outcomes (e.g., failure to appear in court and new arrests while released on bond). Any pretrial services agency either implementing or currently using a pretrial risk assessment tool should carefully consider these components of quality assurance to maximize outcomes for defendants and public safety.

References

Bechtel, K., Holsinger, A. M., Lowenkamp, C. T., & Warren, M. J. (2016). A meta-analytic review of pretrial research: Risk assessment, bond type, and interventions. *American Journal of Criminal Justice, 42*, 443-467. https://doi.org/10.1007/s12103-016-9367-1

Bechtel, K., Lowenkamp, C. T., & Holsinger, A. M. (2011). Identifying the predictors of pretrial failure: A meta-analysis. *Federal Probation, 75*, 78-87.

Cadigan, T. P., & Lowenkamp, C. T. (2011). Implementing risk assessment in the federal pretrial services system. *Federal Probation, 75*, 30-34.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*, 127-37.

Cohen, B. H. (2001). *Explaining psychological statistics* (2nd edition). New York, NY: John Wiley & Sons, Inc.

Cooprider, K. (2009). Pretrial risk assessment and case classification: A case study. *Federal Probation, 73*, 12-15.

Demuth, S. (2003). Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of Hispanic, Black, and White felony arrestees. *Criminology, 41*, 873-907. https://doi.org/10.1111/j.1745-9125.2003.tb01007.x

Farnworth, M., & Horan, P. (1980). Separate justice: An analysis of race differences in court processes. *Social Science, 9,* 381-399. http://dx.doi.org/10.1016/S0049-089X(80)80004-4Federal Bureau of Investigation. (2017). Crime in the United States, 2016. Retrieved https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/resource-pages/about-cius.pdf.

Kennealy, P. J., Skeem, J. L., & Hernandez, I. R. (2017). Does staff see what experts see? Accuracy of front line staff in scoring juveniles' risk factors. *Psychological Assessment, 29*, 26-34. http://dx.doi.org/10.1037/pas0000316

Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims & Offenders, 5*, 203-219. http://dx.doi.org/10.1080/15564886.2010.485900

Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and validation of the Ohio risk assessment system: Final report*. University of Cincinnati.

Mamalian, C.A. (2011). *State of the science of pretrial risk assessment*. Pretrial Justice Institute. Retrieved from https://www.bja.gov/publications/pji_pretrialriskassessment.pdf.

Oleson, J. C., Lowenkamp, C. T., Wooldredge, J., VanNostrand, M., & Cadigan, T. (2017). The sentencing consequences of federal pretrial detention. *Crime and Delinquency, 63*, 313-333. http://dx.doi.org/10.1177/0011128714551406

Oleson, J. C., Lowenkamp, C. T., VanNostrand, M., Cadigan, T., & Wooldredge, J. (2016). The effect of pretrial detention on sentencing in two federal districts. *Justice Quarterly, 33*, 1103-1122. http://dx.doi.org/10.1080/07418825.2014.959035

Ulmer, J. T. (2012). Recent developments and new directions in sentencing research. *Justice Quarterly, 29*, 1-40. http://dx.doi.org/10.1080/07418825.2011.624115

VanNostrand, M. (2007). *Legal and evidence based practices: Application of legal principles, laws, and research to the field of pretrial services*. Crime and Justice Institute. Retrieved from https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=248724.

VanNostrand, M., & Keebler, G. (2007). Our journey toward pretrial justice. *Federal Probation, 71*, 20-25.

VanNostrand, M., & Keebler, G. (2009). Pretrial risk assessment in the federal court. *Federal Probation, 72*, 2-29.

VanNostrand, M., & Lowenkamp, C. (2013). *Assessing pretrial risk without a defendant interview*. Laura and John Arnold Foundation. Retrieved from http://www.arnoldfoundation.org/wp-content/uploads/2014/02/LJAF_Report_no-interview_FNL.pdf.

VanNostrand, M., Rose, K., & Weibrecht, K. (2011). *State of the science of pretrial release recommendations and supervision*. Pretrial Justice Institute. Retrieved from https://www.pretrial.org/download/research/PJI%20State%20of%20the%20Science%20Pretrial%20Recommendations%20and%20Supervision%20(2011).pdf.

Vincent, G. M., Guy, L. S., Fusco, S. L., & Gershenson, B. G. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior, 36*, 225–236. http://dx.doi.org/10.1037/h0093974