Risk Evaluation: Maximizing Risk Accuracy

The first presentation to the Sex Offender Recidivism Commission (SORC) was intended to give a brief overview of the history and mechanics of risk assessment as it has been applied to managing sex offenders. The presentation attempted to place the practices currently used in Massachusetts in an historic, social/political, and methodological context in the hope of guiding discussion about strategies that might be pursued for improving the psychometric reliability and empirical validity of assessment in the state, so that dispositional decisions about the treatment and management of sex offenders might be improved, and public safety might be enhanced.

Brief History of Risk Assessment

Bonta (1996) identified the use of unstructured professional opinion as the *first generation* of risk assessment procedures. This strategy involved assessments that neither specified relevant items nor prescribed a method for combing items to determine risk level. Such unrestricted, unguided clinical prediction has long been recognized as an unreliable and undependable metric for predicting future violence (Monahan, 2007).

The introduction of empirical evidence to guide assessment demarcates Bonta's *second generation* of risk assessment. Hanson and Morton-Bourgon (2009) identified a number of strategies in this second generation. Structured clinical guidelines (SCG) address the issue of which items should be considered. The more sophisticated provide clear anchors and numeric values for recommended items, but none give guidance on how to combine these items. Consequently, SCGs provide no tables linking summary scores to recidivism rates. Empirical actuarials comprise empirically derived items with well-defined, quantitative anchors for rating. They specify the method for combining these items into an overall score, and they provide tables linking the summary scores to recidivism rates. Mechanical actuarials are like their empirical

counterparts in quantifying items and prescribing algorithms for combining items, but they do not provide tables linking the resultant summary scores to predicted recidivism rates. In a practical context empirical and mechanical actuarials can be applied directly, or evaluators can be allowed to adjust their scores using evidence purportedly external to the actuarial.

We are currently in the *third generation*, which is less well researched. The second generation focused on static risk factors, which are fixed or historical factors that cannot be changed. The third generation has introduced the assessment of dynamic risk factors or "criminogenic needs." Dynamic risk factors are characteristics that are both capable of change and their change is associated with modifications (up or down) in recidivism risk.

Historical and Socio-Political Context for Evaluating the MA SORB Classification Factors

The MA Classification Factors for sex offenders were developed in the mid 1990s. The instrument is a SCG because it suggests the domains that evaluators should consider in their judgments about assigning offenders to tiers or risk categories, but it does not have rules on how to combine or weigh items in reaching a decision. Moreover, its items do not have specific anchors, do not provide clear cutoffs for presence or absence of domains, do not result in the assignment of numerical values to item judgments, and at times conflate multiple domains within a single item. Thus, it is not possible to evaluate the reliability or predictive validity of these items or to use empirical research to improve the items or how they are combined in the instrument. One could only generally assess the reliability and predictive validity of the ultimate level recommendations of evaluators, if such independent judgment data were systematically recorded. It is less sophisticated than the more quantitative SCGs, and thus although historically it would be classified as a second-generation instrument, it falls short of other SCGs and is significantly inferior to empirical and mechanical actuarials.

Massachusetts is not alone in its use of suboptimal instruments to classify sex offenders. De facto "tiering" (i.e., categorizing sex offenders in some manner for differential dispositional decisions) occurs in 98% of the states. Only 6% of states use standard mechanical actuarials to make their decisions about offender classification, and an additional 6% have generated their own mechanical actuarials. Two other states with MA (6%) use SCGs. The remaining 80% either do not specify criteria for decisions (17%) or simply use crime categories for classification (63%).

Comparing the Efficacy of Risk Assessment Strategies

The two essential determiners of whether a particular risk assessment strategy is viable are measures of reliability and validity. The former assesses the accuracy or freedom from measurement error of a strategy, which in this area is typically assessed by the agreement between independent raters and the covariations among items in a scale. Validity addresses the question about whether a construct measures what it is purported to measure. In risk assessment the ability of a strategy to predict recidivism is the critical test of validity that determines whether the strategy does what it purports to do.

The reliability of the MA Classification Factors has never been established. The lack of specification of judgment criteria suggests that in its current format it would not achieve adequate levels of interrater reliability. Covariation among its items cannot be calculated in its present format.

A recent meta-analysis by Hanson and Morton-Bourgon (2009) found that empirical and mechanical actuarials were significantly more accurate than SCGs and unstructured judgments in predicting sexual recidivism among sex offenders. This study also found that when clinicians adjusted scores, the resultant scores showed lower predictive accuracy than unadjusted scores.

Zgoba et al. (2015) found in their four-state follow-up study that the crime-based Adam Walsh Act (AWA) criteria either did not predict sexual recidivism at all or in the case of Florida significantly predicted in the opposite direction. This study clearly indicates that simple crime-based sorting of sex offenders, the most common classification process across states, is not a viable tiering strategy. The state-generated tiering systems examined in Zgoba et al. performed better than AWA criteria, but did not reach statistically significant levels of prediction accuracy. The Minnesota actuarial, the Minnesota Sex Offender Screening Tool Revised (MnSOST-R; Epperson, Kaul, & Hesselton, 1999) has been successful in other contexts (e.g., Knight & Thornton, 2007), suggesting that the poor performance of the state instruments in Zgoba et al. might be due to the practice of allowing clinical adjustment of their actuarials in determining tier assignment. A substantial literature has consistently found that mechanical actuarials are superior in predictive accuracy to both clinical judgments and judgments that allow clinical adjustments (Grove, Zald, Lebow, Snitz, & Nelson, 2000), and the reasons for this superiority have been documented (Grove & Meehl, 1996).

These studies, which are representative of the general empirical literature, provide a context both for evaluating the efficacy of the MA Classification Factors and for recommending strategies to improve it. They indicate that the current tiering classification strategy is suboptimal, and they provide two models for improving the accuracy of our decision making—(a) adopting an already well-validated Empirical Actuarial like the Static-99R (e.g., Oregon); or (b) attempting to transform the current criteria into an empirical actuarial (like New Jersey's Registrant Risk Assessment Scale).

Advantages and Disadvantages of Different Improvement Strategies

Adopting, as Oregon did, an already validated empirical actuarial has the advantages that one can choose a classification strategy that (a) uses items empirically supported by the current research literature based on extensive follow-up data, (b) provides specified, anchored criteria for items with quantitative item assignments, (c) has a specific algorithm for combining items into a total score, and (d) proposes recidivism rates based on specific scores. Moreover, the adoption of this strategy can be supplemented by the addition of standard dynamic risk assessment tools that, if applied mechanically, can both increase predictive accuracy and permit the assessment of risk change (e.g., Hanson, Helmus, & Harris, 2015; Thornton & Knight, 2015). The disadvantages of this strategy are that (a) the actuarial would not be fashioned specifically for the local state environment, and (b) because one would be tied to a standard instrument, one may be less likely to assess the instrument for continuous improvement. It is essential for accurate decisions to calibrate risk instruments to local samples and to continuously monitor such calibration (Helmas, Hanson, Thornton, Babchishin, & Harris, 2012).

Alternatively, if we begin with the current classification system as a point of departure and follow the example of those states that have attempted to generate their own actuarials, we would have the advantage of being able to create a classification tool that is (a) uniquely tethered to the local sex offender sample and matched to the state's individual decision processes, and (b) amenable to continuous improvement and responsive to ongoing feedback and evaluation. A model for how such a strategy could be implemented was discussed. The proposed implementation, however, illustrated the considerable disadvantages of this strategy. These included (a) the significant amount of resources that would have to be allocated to the process of revising the current criteria so that they are quantifiable, can be reliably applied, and have

predictive validity, and (b) the long wait that would be necessary to allow a prospective study of the new instrument's predictive accuracy (at least 5 years). Thus, the transformation of the current classification criteria into a reliable and valid instrument would be costly. Moreover, years would pass before it would be possible to gather sufficient evidence to support its validity and to allow calibration of its scores with recidivism frequencies. In contrast, if a standard empirical actuarial were adopted, there would be a considerably faster transition to functionality, and the implementation would be less costly.

Regardless of the strategy chosen, remaining with the status quo is not scientifically defensible. Whatever strategy is ultimately chosen, it must include the establishment of adequate reliability, clear mechanical rules for combining items to generate risk scores, clear mechanical rules for using dynamic risk assessments that would be useful in treatment and monitoring change, and built-in procedures for assessing efficacy and continuous improvement. Moreover, the New Jersey experience with implementing its risk assessment procedures has taught us that continuous monitoring of evaluator training and reliability is essential (Lanterman, Boyle, & Ragusa-Salerno, 2014). Subsequent presentations addressed the additional needs of taking into account special populations (e.g., juveniles, women, adults with either major mental illness or intellectual disabilities) when fashioning risk tools.

References

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.

Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Alexander, W., & Goldman, R. (1999). *Minnesota sex offender screening tool - Revised (MnSost-R): Development performance, and recommended risk level cut scores*. Retrieved from www.psychology.iastate.edu/faculty/epperson

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public, Policy, and Law, 2(2),* 293-323.

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.

Hanson, R. K., Helmus, L., & Harris, A. J. R. (2015). Assessing the risk and needs of supervised sexual offenders: A prospective study using Static-99R and STABLE-2007. *Criminal Justice and Behavior.* doi: 10.1177/0093854815602094

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

Helmas, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior, 39(9),* 1148-1171. doi: 10.1177/0093854812443648

Lanterman, J. L., Boyle, D. J., & Ragusa-Salerno, L. M. (2014). Sex offender risk assessment, sources of variation, and the implications of misuse. *Criminal Justice and Behavior, 41(7),* 822-843.

Knight, R. A., & Thornton, D. (2007). *Evaluating and Improving Risk Assessment Schemes for Sexual Recidivism: A Long-Term Follow-Up of Convicted Sexual Offenders.* Final Report, NCJ 217618, http://nij.ncjrs.gov/publications

Monahan, J. (2007). Clinical and actuarial predictions of violence. In D. Faigman, D. Kaye, M. Saks, J. Sanders, & E. Cheng (Eds.), *Modern scientific evidence: The law and science of expert testimony* (pp. 122-147). St. Paul, MN: West Publishing.

Thornton, D., & Knight, R. A. (2015). Construction and validation of SRA-Need Assessment. *Sexual Abuse: Journal of Research and Treatment, 27(4),* 360-375. doi: 10.1177/1079063213511120.

Zgoba, K. M., Miner, M., Levenson, J., Knight, R., Letourneau, E., & Thornton, D. (2015). The Adam Walsh Act: An examination of sex offender risk and classification systems using data from four states. *Sexual Abuse: A Journal of Research and Treatment*. doi: 10.1177/1079063215569543