Structured professional judgment (SPJ) instruments are used at various points in justice decision making, including release to parole and sanctioning of adjudicated youth. In other disciplines, these tools are called “expert” or “consensus” systems. SPJ models identify factors to consider in assessing risk (usually based on prior research and/or theory), but items are not scored. At the end of every assessment, workers simply assign a level of risk, presumably based on the overall profile derived from factors assessed. It is one of the most troubling developments in risk assessment in recent years.

Rating risk factors but not summing them to create an overall risk score has long proved problematic. As far back as the 1950s, assessment studies demonstrated that while highly trained clinicians could reliably rate individual risk factors, they had little success in predicting outcomes. In one seminal study of this issue, Bleckner (1954) found that simply summing the ratings given individual factors by clinicians produced far more accurate predictions of outcomes than individual clinicians could provide. A study of SPJs in child welfare found they were neither valid nor reliable and were significantly outperformed by additive actuarial models (Baird & Wagner, 2009). Given this history, coupled with the relative success of actuarial tools in the justice field, it seems strange that anyone would suggest SPJs have advantages over risk assessment systems where factors are weighted and scored to produce a risk level.

Several assessments serve as examples of problems with the SPJ approach. Some used in the adult criminal justice system (e.g., the PCL-R) were not developed specifically for corrections. These systems were originally intended for use in the mental health system to assist in diagnosing psychopathy and gradually made their way into corrections. Recent studies have raised issues concerning reliability and validity (see, for example, Singh, Frazal, Ralitza, & Buchanan, 2014; Yang, Wong, & Coid, 2010). Their use in release decisions is especially controversial and currently is being challenged in court in California.

One SPJ assessment, the SAVRY, was developed specifically for the juvenile justice system to assess a youth’s potential for violence. The SAVRY is particularly important because it has gained considerable support in recent years, due in part to the MacArthur Foundation's Models for Change initiative. The SAVRY is used in the United States and several European countries.
Vincent and colleagues claimed that “the SAVRY has reported the best predictive accuracy of any instrument based on available research” (Vincent, Terry, & Maney, 2009). The lead author cited, as evidence for this claim, a “meta-analysis” of 11 studies of the SAVRY (Yang et al., 2010). However, a detailed review of those studies reveals the following.

1. Though the study is presented as a meta-analysis, the review really represents a basic compendium of very small studies. The average sample size of studies cited in this article is 113 cases; no study included more than 176 cases, and four studies included fewer than 100 cases.

2. These studies were from a variety of countries including Canada, Spain, the Netherlands, England, and Germany. Policies and practices undoubtedly varied greatly across these countries, seriously diminishing the value of any attempt to combine results. Only one study was conducted in the United States, which presented results indicating poor predictive validity.

3. Validity measures presented in the meta-analysis were limited to correlation coefficients and ROC values. No data were provided regarding the ability of SAVRY to discriminate by risk level.

4. Follow-up periods varied substantially. In some studies, a standard follow-up period was not used.

These issues are important. First, small study samples offer little in terms of knowledge advancement. When small samples are divided into three or four different risk categories, not to mention gender or racial/ethnic groupings, the number of cases in each category is too small to produce stable and meaningful statistics. The total combined sample from all 11 SAVRY studies was 1,239, a figure frequently exceeded by single studies of other risk assessment models (see, for example, Gottfredson & Snyder, 2005; and Baird et al., 2013). The Gottfredson and Snyder study sample alone comprised more than 9,500 cases. Secondly, combining the results of small studies from agencies with widely disparate policies, procedures, and offender populations is problematic at best. Third, the US study, the second largest cited, found the SAVRY produced a correlation of .15 with recidivism. It is not uncommon to find a single factor (such as prior delinquencies) that correlates at a higher level. Finally, results from Canada far exceeded those produced in other countries, raising questions regarding transferability.

SAVRY proponents who served as consultants on the Models for Change initiative generally ignored the work of many other researchers, citing problems with short actuarial risk models that simply do not exist. For example, in Risk Assessment in Juvenile Justice: A Guidebook for Implementation, they state that without evidence or citation, “unfortunately, there is no brief risk tool currently available that can adequately identify low-risk youth” (Vincent, Guy, & Grisso, 2013, p. 36). This is not true. The guidebook fails to consider findings from a major study conducted in 2005 for the Office of Juvenile Justice and Delinquency Prevention by highly respected researchers (Gottfredson & Snyder, 2005). These researchers developed a nine-
item scale that effectively divided cases into five risk categories with substantial differences in outcomes, accurately identifying low-risk cases. In addition, the National Council on Crime and Delinquency (NCCD) developed an 11-item scale that outperformed longer instruments in identifying low-risk youth (Baird et al., 2013).

Further, some studies conducted to measure the validity of the SAVRY focused on an analysis of “scores” derived by summing items from SAVRY subscales (Gretton & Abramowitz, 2002; Catchpole & Gretton, 2003). While studies of total or subscale “scores” may indicate some relationship between SAVRY scores and outcomes, they say little about “summary risk level” assigned by workers. Some studies have, in fact, found surprisingly low correlation between scores and summary ratings, an indication that subjectivity may play a significant role in assigning overall risk ratings (Gretton & Abramowitz, 2002).

Inter-rater reliability citations used to support the SAVRY are also frequently based on very limited studies. In one study, SAVRY ratings were completed by two teams, each composed of two staff members with advanced degrees (Lodewijks, Dorleijers, & Ruiter, 2008). Thus, assignments to risk levels were agreed upon by two individuals. This does not represent what occurs in most US probation agencies, where assessments are completed independently by dozens, if not hundreds, of staff with very different levels of experience and education. The small size of the study (N=25), combined with the fact that it was completed on cases from a European country with different policies, procedures, and laws, renders the reliability results rather meaningless to any US jurisdiction. Further review of the analysis of individual scale factors conducted by these researchers demonstrates just how spurious relationships derived from small samples can be: Several items shown in prior research to be related to recidivism and violence were inversely correlated with outcomes in this study. This is counterintuitive and at odds with violence theory at the foundation of the SAVRY. This is very likely an artifact of the small sample size.

Models for Change consultants have also been selective regarding results produced by the SAVRY in Louisiana. A 2011 publication linked reductions in the use of residential placement to the SAVRY (Models for Change Knowledge Brief, 2011). It may be likely that these reductions are related more to increases in community-based programs introduced as part of the program. No data on the relationship between SAVRY ratings and outcomes are presented.

The Models for Change consultants did, to their credit, conduct an inter-rater reliability study of the SAVRY, concluding that ratings among staff members were highly reliable. But this finding stands in sharp contrast to other findings regarding SPJs (Baird & Wagner, 2000; D’Andrade, Benton, & Austin, 2005). The difference in findings may well be attributable to study methodology.
There is no perfect way to conduct reliability studies. Results obtained from such studies always will be mere estimates of the level of reliability attained in the field. Replicating the actual field assessment process as closely as possible probably produces the best estimates. Whereas NCCD used case files from four jurisdictions in its study of SPJs in child welfare (Baird, Wagner, Healy, & Johnson, 1999), Models for Change consultants used vignettes created specifically for the study. Using case files to measure reliability may well produce a more accurate estimate of what will be achieved in actual practice. Case files represent information typically collected and available to assessors.

A plethora of research indicates that SPJs do not provide the degree of structure needed to ensure reliability among raters in large, diverse agencies (D’Andrade et al., 2005). Agreement among independent raters is often well below acceptable standards, indicating the SPJ process is simply too subjective to provide a high level of consistency among staff members.

In sum, evidence supporting the SAVRY is not well established and is based almost entirely on very small studies that often fail to report on the actual levels of discrimination achieved. There also is research showing that SPJ models lack the level of validity needed to improve decision making. Most actuarial models allow workers to override the risk level derived through scoring if they know something about the case that indicates a higher or lower designation. Overrides often require supervisory review and can be tracked to help ensure fidelity to the system. This builds in flexibility to include judgment of the staff member conducting the assessment, while placing needed control on the use of subjective judgment. SPJs lack the structure needed in justice decision making and represent a step backward in practice. Agencies should avoid using these tools.
References


