© 2011 American Psychological Association
0147-7307/11/$12.00 DOI: 10.1037/h0093974

# Field Reliability of the SAVRY With Juvenile Probation Officers: Implications for Training

Gina M. Vincent, Laura S. Guy, Samantha L. Fusco, and Bernice G. Gershenson
University of Massachusetts Medical School

Two complimentary studies were conducted to investigate the inter-rater reliability and performance of juvenile justice personnel when conducting the Structured Assessment of Violence Risk for Youth (SAVRY). Study 1 reports the performance on four standardized vignettes of 408 juvenile probation officers (JPOs) and social workers rating the SAVRY as part of their training. JPOs had high agreement with the expert consensus on the SAVRY rating of overall risk and total scores, but those trained by a peer master trainer outperformed those trained by an expert. Study 2 examined the field reliability of the SAVRY on 80 young offender cases rated by a JPO and a trained research assistant. In the field, intra-class correlation coefficients were 'excellent' for SAVRY total and most domain scores, and were 'good' for overall risk ratings. Results suggest that the SAVRY and structured professional judgment can be used reliably in the field by juvenile justice personnel and is comparable to reliability indices reported in more lab-like research studies; however, replication is essential.

*Keywords:* SAVRY, field reliability, juvenile justice, juvenile probation, risk assessment

Assessments of the likelihood of future violence or serious re-offending are relevant for a variety of legal decisions in the juvenile justice system. Juvenile justice personnel and judges are routinely expected to make decisions involving the handling of youth who come into contact with the law across various points in the system. These decisions generally involve determinations of which youths pose a threat to public safety, which youths are likely to benefit from interventions, and which interventions are most likely to result in a positive change (Fagan & Zimring, 2000; Mulvey, 2005). According to the *risk principle*, both public safety concerns and intervention planning should be driven, in part, by an offender's likelihood of re-offending. Generally, those at highest risk to reoffend should receive the most intensive services, whereas low risk cases are unlikely to re-offend even if no or minimal services are provided (Andrews & Bonta, 2002; Andrews, Bonta, & Hoge, 1990; Andrews & Dowden, 2006).

As such, case processing decisions must start with reliable and valid identification of offenders' level of risk for violence or serious offending and their specific treatment needs for reducing the risk (Andrews, 1989; Austin, 2006; Grisso, 2005). Several scholars and professionals have written about the importance of using some form of structured decision-making tool or assessment of risk for reoffending in criminal and juvenile justice (Gottfredson & Tonry, 1988; Grisso, Vincent, & Seagrave, 2005; Hoge, 2002). In the past decade, many risk assessment instruments have been developed and validated for use with young offenders (see Hoge, 2002; Otto & Douglas 2009; Vincent, Terry, & Maney, 2009 for reviews). The use of risk assessment has grown from 33% of state juvenile justice systems in 1990 to 86% by 2003 (Griffin & Bozynski, 2003).

Several tools have established good predictive accuracy as demonstrated in peer-reviewed research by independent parties (meaning a party who does not have a financial interest in the tool; Schwalbe, 2007; Olver, Stockdale, & Wormith, 2009). Some tools, such as the *Washington State Juvenile Court Assessment* (WSJCA; Barnoski, 2004; Barnoski & Markussen, 2005), were designed specifically for use by probation officers, whereas others, such as the *Structured Assessment of Violence Risk for Youth* (SAVRY; Borum, Bartel, & Forth, 2003/2006), were designed for use by a larger group of professionals including psychologists, social workers (SWs), and non-clinical staff such as probation officers.

Although the notion of implementing risk for recidivism assessment tools in juvenile probation, on the surface, seems like it would increase the consistency and validity of decision-making, these tools could have deleterious effects if the tools were unreliable or invalid. Among the instruments that have demonstrated sound inter-rater reliability, which unfortunately are few (see review by Vincent et al., 2009), reliability typically has been estab-

This article was published Online First April 27, 2011.

Gina M. Vincent, Laura S. Guy, Samantha L. Fusco, and Bernice G. Gershenson, Center for Mental Health Services Research, Department of Psychiatry, University of Massachusetts Medical School.

This research was funded by the John D. and Catherine T. MacArthur Foundation as part of the Models for Change Research Network. The authors would like to acknowledge Dr. Debra DePrato, MD, Associate Clinical Professor of Public Health, for making this study possible; the Louisiana Office of Juvenile Justice, particularly Mary Livers, PhD, Deputy Secretary and Kelly Clement, Regional Manager, for organizing all of the probation officer trainings and assisting us with gathering data; Patrick Bartel, PhD, for his SAVRY training; and both Dr. Bartel and Randy Borum, PsyD, for their guidance around implementing the SAVRY. Finally, the authors wish to thank our site research associates, Joshua Everett, MA, Brady Holtzclaw, MA, and Brittany Foreman.

Correspondence concerning this article should be addressed to Gina M. Vincent, Center for Mental Health Services Research, Department of Psychiatry, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655, USA. E-mail: Gina.Vincent@umassmed.edu

lished only in "laboratory-like settings" using trained research assistants who completed the assessments based on file review. In most cases, the reliability of these psychological measures in the field, where raters have added pressures of time constraints and political demands for incarceration of high profile cases, remains uncertain. As such, this study was designed to address the question of whether a valid risk assessment tool, specifically the SAVRY, could be completed reliably by juvenile probation officers (JPOs) in the field.

## Research on the Field Reliability of Psychological Measures

The importance of testing the inter-rater agreement of risk assessment or forensic assessment tools was highlighted recently in a set of field reliability studies of the Psychopathy Checklist-Revised (PCL-R; Hare, 2003), the most widely used forensic assessment instrument in clinical practice. The studies examined PCL-R scores generated for offenders on trial for sexually violent predator commitments by clinicians from opposing sides of the case (Boccaccini, Turner, & Murrie, 2008; Murrie, Boccaccini, Johnson, & Janke, 2008). The agreement between prosecution- and defense-retained evaluators using an intra-class correlation coefficient (ICC—for absolute agreement for a single evaluator) was only .39. This agreement is markedly different than that reported in the PCL-R manual from more "laboratory-like studies," which tend to base reliability estimates on agreement between trained graduate students or clinicians conducting the PCL-R for research purposes ($\text{ICC}_1$s range from .86 to .94). The authors were able to explain 34% of the variance in scores as a result of evaluator-level differences (some evaluators had a tendency towards assigning higher scores) using multi-level modeling. Unfortunately, 53% of the variance was still unexplained. They posited several explanations for the larger discrepancies between ratings found in the field versus the "lab," including the possibility of adversarial allegiance. In other words, the side paying for the evaluator, be it prosecution or defense, may have had some bearing on the evaluator's final PCL-R ratings.

The studies from Boccaccini et al. (2008) and Murrie et al. (2008) raise serious questions about the reliability of forensic assessments when in the hands of practitioners. Although the design had strong ecological validity, there were several factors that could have affected inter-rater agreement that could not be controlled. Most notably, the uncontrolled factors included differences in the 21 examiners' training on the assessment tool (which was unknown) and consistency between examiners in the amount of information or quality of the interview with the offender that was used to rate the PCL-R.

## The Structured Assessment of Violence Risk in Youth (SAVRY)

The SAVRY (Borum et al., 2006) was designed to assess violence risk in adolescents aged 12–18 years. However, several studies of young offender populations have demonstrated that the SAVRY has high predictive accuracy for both violent and non-violent re-offending (e.g., Lodewijks, Doreleijers, & de Ruiter, 2008; Welsh, Schmidt, McKinnon, Chattha, & Meyers, 2008). Its predictive validity has been demonstrated in forensic and young offender populations in at least ten studies published by independent researchers (see Borum, Lodewijks, Bartel, & Forth, 2009 for a review). A recent meta-analysis reported moderate effect sizes (weighted *r*) for SAVRY total scores of .38 for nonviolent re-offending and .30 for violent re-offending (Olver et al., 2009). These effect sizes are comparable to or better than other risk assessment tools for youth (see Vincent et al., 2009 for a review).

The SAVRY protocol comprises six items defining Protective Factors (which may lower the likelihood of risk) and 24 items defining Risk Factors (which may increase the likelihood of risk). Evaluators also are able to designate additional risk and protective factors, recognizing that some cases may present circumstances that are not included among the SAVRY factors. Because the SAVRY uses a structured professional judgment approach (as opposed to a formulaic actuarial approach), the final determination of an examinee's overall level of risk for violence is the evaluator's Summary Risk Rating (SRR; Low, Moderate, High risk) based on the examiner's professional judgment as informed by a systematic appraisal of relevant factors. This assures that examiners assess risk factors that are empirically associated with violence, consider the applicability of these factors to the specific examinee's risk for reoffending (otherwise known as *criminogenic needs*), and classify the severity to make their final SRR. Thus, a prime advantage of the SAVRY is that it not only structures the process by which a valid decision is made about the likelihood that one will recidivate, but it also includes dynamic risk factors that identify need areas that may be effective targets for treatment.

Inter-rater reliability (IRR) for the SAVRY has been acceptable to excellent as reported in several studies by independent researchers. In a recent review, Borum et al. (2009) summarized results of IRR analyses across six studies to date. $\text{ICC}_1$s ranged from .81 to .97 for numeric total scores and .72 to .95 for Summary Risk Ratings. Those IRR analyses were based on SAVRYs double-rated by trained master's level clinicians or research assistants, typically based on file information without an interview with examinees. Examiner agreement for probation officers has not been reported.

## Methods of Assessing Inter-Rater Agreement

There are a few statistical methods commonly used to quantify inter-rater agreement. First, it is helpful to report percentage agreement across raters because of its understandability, but it does not account for variance between raters. Percentage agreement also may exaggerate the apparent level of inter-rater agreement because it does not control for agreements that occur simply by chance. Kappa is a measure of agreement that controls for chance (Cohen, 1960) and can be calculated across raters even when the rater-pairs across subjects differ (Fleiss, 1981). There are versions of the kappa that can be used with multiple raters and multi-category, as opposed to dichotomous, ratings (Chen, Zaebst, & Seel, 2005; Green, 1997). Pearson's r also is used, but this index does not account for additive and multiplicative biases between raters, as does the ICC. A main advantage of the ICC is that it produces an index of agreement while accounting for variance across raters (Shrout & Fleiss, 1979).

As Shrout and Fleiss (1979) noted, there are many versions of the ICC that can produce very different results or similar results that vary in how they should be interpreted and their assumptions. Thus, it is important for researchers to select the appropriate method for their design and to note which version they used. The two types of models differ in terms of whether (a) each subject is rated by a different and random selection of a pool of raters or (b) each subject is rated by the same raters. For the latter model, researchers may use an ICC in which *consistency* (examines whether raters are giving higher and lower scores to the same subjects regardless of the numerical value of the score) or *absolute agreement* (examines differences in the actual scores given by raters) is investigated.

There is some variability in the interpretation of the magnitude of ICCs. Referring to weighted kappa (which is mathematically equivalent to ICC), Cicchetti and Sparrow (1981) defined reliability indices below .40 as “poor,” .40 to .59 as “fair,” .60 to .74 as “good,” and .75 or above as “excellent.” Landis and Koch (1977) defined reliability indices between .61 and .80 as “substantial” and between .81 and 1.00 as “almost perfect.” Fleiss (1986) suggested the following classifications for single measure $\text{ICC}_s$: $\text{ICC} \geq 0.75$, excellent; $0.60 < \text{ICC} < 0.75$, good; $0.40 < \text{ICC} < 0.60$, moderate; and $\text{ICC} < 0.40$, poor. Basically, a tool should strive for ICCs of at least .61 and preferably above .75.

## The Present Study

Few researchers have reported on the inter-rater reliability of risk assessment tools in the field, especially in juvenile justice settings. The present studies investigated the SAVRY’s field inter-rater reliability among trained JPOs and juvenile justice SWs. Study 1 examined the performance of a large sample of JPOs and SWs across four standardized case vignettes which they completed as part of their SAVRY training. It is important to note that because these case vignettes were part of the JPOs’ and SWs’ training, one should expect a certain amount of errors on the ratings of these cases, which does not necessarily reflect how well JPOs or SWs performed in the field after training. Consequently, Study 2 was designed to examine the actual reliability of SAVRY ratings in the field among a subsample of these JPOs. JPOs conducted interviews with randomly selected youth probationers and parents that were observed by a second rater. Then, both the JPOs and second raters rated the SAVRY after being exposed to the same interview and having access to the same file information. Thus, our investigation of inter-rater reliability controlled for consistency in the amount and type of information used to rate the SAVRY.

## Study 1: Accuracy of Probation Officer Ratings on Standardized Case Vignettes

This study examined ratings on four standardized practice-case vignettes completed by JPOs and SWs shortly after they completed separate 2-day SAVRY training workshops. These practice-case vignettes were the required follow-up component of the juvenile justice personnel’s SAVRY training, which had to be completed before they used the SAVRY in their day-to-day operations with young offenders. The researchers obtained the data afterward to evaluate JPOs’ and SWs’ *performance* (or accuracy) by comparing their ratings on these cases to expert consensus ratings. This study was concerned with rater accuracy not rater agreement or reliability, per se. The analyses for Study 1 addressed two questions. First, were certain SAVRY items consistently more difficult for juvenile justice personnel to rate accurately? Second, did performance on the practice-case vignettes differ as a function of profession (JPO vs. SW) or the type of training received?

### Method

**Sample.** Raters were 349 (85.5%) JPOs and 59 (14.5%) SWs from one of 16 juvenile probation offices and three secure custody facilities in Louisiana ($N = 408$). Four of the juvenile probation offices were local (governed by the parish) and the remaining offices and facilities were operated by the state (i.e., Louisiana Department of Public Safety and Corrections, Youth Services, Office of Juvenile Justice; OJJ). Overall, 58% of participants were female (54% of JPOs and 79% of SWs). A small majority of participants were White (56%) and 43% were Black/African American. A higher percentage of SWs than JPOs were female (79 vs. 54%, respectively) and Black (86 vs. 34%, respectively).[1] The JPOs and SWs conducted ratings on case vignettes as part of their mandated SAVRY training.

**Measures.**

***The structured assessment of violence risk in youth.*** As described earlier, the SAVRY (Borum et al., 2006) is a risk assessment tool for use with adolescents based on the structured professional judgment model. The SAVRY comprises 24 risk factor items that are coded based on statements in the manual describing the conditions under which a case receives a Low, Moderate, or High rating on each item. The six protective factors are rated as ‘Present’ or ‘Absent.’ As is typical in practice with rating scales like the SAVRY, raters were trained to use “sliders” when rating items where there was not enough information to make a firm decision. For example, if it was unclear whether the case should receive a Low versus a Moderate on a particular item, the rater could give a Low ‘slider up’ (Low+) or a Moderate ‘slider down’ (Moderate−).

The SAVRY risk items are divided conceptually into three domains: *Historical* (10 items), *Social/Contextual* (six items), and *Individual/Clinical* (eight items). The Historical items primarily are static in nature whereas the Individual/Clinical and Social/Contextual items primarily are dynamic. The SRR is the examiner’s final determination of the examinee’s relative likelihood to commit violence in the future (Low, Moderate, or High). For the purposes of this study, a SAVRY was considered invalid if five or more items (i.e., 17% of items) were not rated. The rationale for this low threshold was that the vignettes contained sufficient information to rate most items and, therefore, few if any items should have been missing.

***Case vignettes.*** The researchers created four case vignettes to be used for post-workshop training purposes. The vignettes were based on real cases of adjudicated young offenders ob-

[1] The race figures reported here were based on data from the Louisiana Office of Juvenile Justice. Unfortunately, the researchers did not have race data for each specific JPO or SW in Study 1 so it was not possible to conduct any later comparisons by race.

tained (de-identified) from reports written by JPOs and forensic psychologists. Reports were selected to be adapted for use as training vignettes based on the richness of the data available in the report. We aimed to develop vignettes that differed in terms of the type of index offense, delinquency history, and psychosocial characteristics of the youth/family. The primary changes made to the de-identified reports included simplifying language and removing clinical or technical terms that we anticipated JPOs might not understand, and adding information to ensure that each SAVRY item could be rated. Vignettes ranged from five to nine pages and were labeled Brad, Emily, Krista, or Wendy.

The researchers developed expert consensus ratings and scoring rationale for each SAVRY item and the SRR for all four cases. The process involved two to four trained forensic psychology doctoral graduates providing individual ratings on all cases, in addition to one of the authors of the SAVRY (Dr. Patrick Bartel). For the majority of items, all raters assigned the same rating. The number of items for which ratings were not unanimous across raters was as follows: 15 for Brad, 16 for Emily, 16 for Krista, and four for Wendy. For most of these items, however, three of the four raters gave the same rating. For the SRR, there were no major disagreements aside from two of the raters on the Krista case. One of the researchers assigned consensus ratings after a review of the raters' rationales and discussion with Patrick Bartel to resolve scoring differences. The final consensus SRRs across cases were: Brad (High), Emily (High, slider down), Krista (Moderate, slider down), and Wendy (Moderate, slider up).

**Procedures.** Each JPO and SW completed a training workshop on the SAVRY that covered information about the trajectories of youth offending, developmental issues, research on risk factors, and the SAVRY scoring criteria. The workshops included rating the SAVRY for two case vignettes, which were reviewed and discussed as a group. The workshops were conducted at each individual probation office or secure facility. JPOs ($n = 70$) and SWs ($n = 59$) from five of these sites completed a 2-day workshop with an expert (one of the SAVRY authors or one of the researchers). JPOs ($n = 274$) from the other 13 probation offices completed 1-day workshops from a local peer master trainer (these were exceptional probation officers at each site who completed training in a 2-day workshop by Dr. Bartel and were charged with training the rest of their staff on the SAVRY). Thus, all SWs were trained by an expert but less than one-quarter of the JPOs were trained by an expert.

All JPOs and SWs were expected to complete a minimum of two post-training case vignettes, but most completed three. After completing each case vignette, raters received feedback as a group in a 2-hour training session from an expert or peer master trainer regarding how well their ratings corresponded to the consensus ratings and they were provided with the rationale for each item rating. Feedback was received prior to completion of their next case and so on. The researchers obtained the case ratings for each JPO and SW from the master trainers at each site.

**Data analyses.** Rater *performance or accuracy* on the SAVRY for the case vignettes was operationalized in two ways. First, the percent of disagreement (% of JPOs/SWs with ratings that deviated from the expert consensus) was calculated for each SAVRY item and SRR for each of the four vignettes. Percent of disagreement in this case should be seen as a measure of the % of incorrect items, not as a measure of across rater disagreement. To account for sliders, ratings falling on either side of the slider were considered to be in agreement with the consensus rating. For example, if the consensus rating was Moderate+ then both Moderate and High ratings were considered to be in agreement with the consensus.[2] Second, for some analyses performance was considered the overall number of "correct" item ratings (meaning items corresponding with the expert consensus while considering sliders) on the SAVRY total and four domain scores.

The order in which vignettes were distributed for post-workshop training varied across some offices where it was not controlled by the researchers. Therefore, for the analyses of difference in performance depending on profession or training type, it was necessary to account for both vignette difficulty and potential practice effects. *Practice* was operationalized as the timing of the case completion, be it the rater's first, second, or third case. Differences in performance were examined using two-way analyses of variance (ANOVAs) with number of items correct on the SAVRY total, four domain scores, and the SRR as dependent variables.

## Results

**Difficulty of SAVRY items.** After removing the 21 invalid SAVRYs, there were 647 SAVRY case ratings completed by 349 JPOs and SWs across the four case vignettes. First, the accuracy of field ratings as compared to the expert consensus ratings was examined for each vignette. Table 1 displays the percentage of JPOs and SWs who gave item and SRR ratings that were not in agreement with the expert consensus, and the average number of *incorrect* items within each SAVRY domain and the SAVRY total score. The last column in Table 1 displays the percentage of disagreement for raters across all vignettes. Items with 40% or more of raters disagreeing with the expert consensus were considered to be especially difficult.

**Differences in performance by training.** To review, the sample comprised SAVRY ratings on vignettes by individuals who were trained in one of three ways: SWs trained by an expert in a 2-day workshop, JPOs trained by an expert in a two-day workshop, and JPOs trained by a peer master trainer in 1 day. Our goal was to examine differences in performance by profession and type of training while accounting for extraneous factors such as vignette difficulty (Table 1 indicates that the Emily and Krista cases received the most incorrect item ratings on average), practice effects, and rater-gender. Based on some exploratory analyses, it was determined that (a) practice effects were not common and appeared to be dependent on the vignette, and (b) there was a

[2] The ideal approach would have been to examine absolute agreement, meaning an item rating of Low+ or Moderate- would both be acceptable answers if the consensus rating was Low+ or Moderate−. Unfortunately, the researchers did not receive all of the training rating data in a format that provided the sliders so it was not possible to examine absolute agreement.

rater-gender effect on the Historical and Protective Factor domains.[3]

Since there were no consistent practice effects, we accounted for both practice and vignette difficulty by calculating each JPO's and SW's average performance (average number of correct items) across the two or three vignettes they completed (since only 24 JPOs completed four vignettes, ratings for the fourth vignettes were not included in the averages) for the SAVRY total score, four SAVRY domain scores, and the SRR. Gender of rater was included as a factor in a series of two-way ANOVAs to examine the average performance by gender and training type or profession. These analyses excluded the 32 raters for whom gender was missing.

We began by comparing SWs ($n = 57$) to JPOs who completed a 2-day training with an expert ($n = 58$) on their performance on SAVRY total and domain scores using two-way ANOVAs (Profession × Gender). The omnibus $F$ tests were significant at the .008 level ($p$ was set at .008 following a Bonferroni correction) for two of the six analyses. There were significant main effects for Profession on the Social/Contextual (SWs—$M = 4.97$, $SD = 0.90$ versus JPOs—$M = 4.33$, $SD = 1.00$; $F[1, 111] = 10.64$, $p = .001$) and Protective Factor (SWs—$M = 4.31$, $SD = 0.68$ versus JPOs—$M = 3.67$, $SD = 0.97$; $F[1, 111] = 16.11$, $p < .001$) domains such that SWs were more adept at making these ratings than JPOs given equivalent training. According to standardized effect sizes (calculated based on the marginal means), the magnitude of the differences between groups was large; for the Social/Contextual domain, Cohen's $d = .67$ and for the Protective Factor domain, Cohen's $d = .76$.[4] However, these differences did not have a notable impact at the test level. Specifically, there were no significant differences in performance on the SAVRY total score or in the average number of times raters gave a SRR rating consistent with the expert consensus rating. There were no significant gender effects.

Next we compared the performance in SAVRY ratings between SWs and JPOs trained by an expert ($n = 115$) and JPOs trained by a peer master trainer ($n = 223$) in two-way ANOVAs (Training Type × Gender). The omnibus $F$-tests were significant at the .008 level ($p$ was set at .008 following a Bonferroni correction) for two of the six analyses, such that JPOs trained by a peer outperformed those trained by an expert (regardless of educational background) and female raters outperformed male raters. For SAVRY total scores, there was a significant main effect for Training Type (expert-trained—$M = 15.97$, $SD = 2.44$ versus peer-trained—$M = 16.94$, $SD = 2.31$; $F[1, 334] = 13.13$, $p < .001$; $d = .41$) and Gender (males—$M = 16.26$, $SD = 2.42$ versus females—$M = 16.85$, $SD = 2.35$; $F[1, 334] = 5.38$, $p = .02$; $d = .25$). On the Historical domain, there was also a significant main effect for Training Type (expert-trained—$M = 6.09$, $SD = 1.47$ versus peer-trained—$M = 6.61$, $SD = 1.40$; $F[1, 334] = 9.43$, $p = .002$; $d = .36$) and Gender (males—$M = 6.20$, $SD = 1.42$ versus females—$M = 6.60$, $SD = 1.44$; $F[1, 334] = 5.14$, $p = .02$; $d = .28$). There was a similar trend on both the Individual and Social/Contextual domains, but these did not achieve statistical significance. Where differences were significant, the magnitude of the differences between those trained by an expert and those trained by a peer was moderate and gender effects were small. There was no appreciable difference in performance due to gender or training type on the SRR or Protective Factor domain.

## Discussion

Overall, results from the case vignettes demonstrated that JPOs and SWs generally were better at making the SAVRY summary risk rating than item ratings, having disagreements on the SRR in only 19% of cases. This bodes well for use of the SAVRY with juvenile justice personnel because the SRR is used to communicate a youth's level of risk in practice. It should be noted that in most of these cases, with the exception of the Brad vignette, the expert consensus rating for the SRR involved a slider (e.g., Moderate to High or Moderate to Low) so we considered responses on either side of the slider to be agreement. The Brad case had only one correct SRR answer, and so the threshold for agreement was higher; this may explain why there was more disagreement on the SRR for Brad despite the relatively high agreement on his SAVRY items.

Seven of the 30 SAVRY items were consistently difficult for raters. Past Supervision/Intervention Failures, Parental/Caregiver Criminality, and Early Caregiver Disruption were rated incorrectly for over half of the cases. Childhood History of Maltreatment, Risk Taking/Impulsivity, Strong Attachment and Bonds, and Positive Attitude toward Intervention and Authority were rated incorrectly by 40–49% of raters. It is possible these items were particularly difficult to rate based on a vignette as opposed to a live case. Alternatively, the manual descriptions for these particular items may be in need of revision.

There were two very noteworthy findings pertaining to training that have large practical implications. First, JPOs performed as well on the SAVRY as SWs, given equivalent SAVRY training, with respect to the most practical aspects of the SAVRY; namely, the overall rating of risk and the SAVRY total scores. However, JPOS were less adept than SWs with equivalent training at rating items in the Social/Contextual and Protective Factor domains. Second, JPOs who received a 1-day

[3] We examined whether practice and gender of rater were potential covariates by conducting five two-way (Vignette × Timing of Vignette [a measure of practice]) ANOVAs and five two-way (Vignette × Gender) ANOVAs, one for performance on the SAVRY total score and then for each of the four domain scores. This approach was preferred to one omnibus MANOVA because it allowed a liberal examination of potential covariates. We conducted these analyses across all cases rather than across raters, again to take a more liberal approach. There was not a significant main effect for Timing of Vignette on performance on any domain except the Protective Factor domain ($F[2, 611] = 3.39$, $p = .03$). For three domains, there was a significant interaction between Vignette and Timing of Vignette on performance (Historical—$F[6, 611] = 7.47$, $p = .008$; Individual— $F[6, 611] = 7.47$, $p = .006$; Protective—$F[6, 611] = 7.69$, $p < .001$). Most importantly, there was a significant main effect of Vignette on performance for every domain and the SAVRY total score. For the gender analyses, there was a main effect of Gender on performance on the Historical domain, such that females ($M = 6.75$; $SE = .11$) performed better than males ($M = 6.39$; $SE = .13$; $F[1, 564] = 4.47$, $p = .04$), and on the Protective Factor domain, such that males ($M = 3.86$; $SE = .09$) performed better than females ($M = 3.60$; $SE = .08$; $F[1, 564] = 4.79$, $p = .03$). Results of all analyses are available from the senior author.

[4] According to Cohen (1988, 1992), a small-sized correlation is $r = \pm.10$, a moderate-sized correlation is $r = \pm.30$, and a large correlation is $r = \pm.50$. The corresponding thresholds for standardized mean differences (i.e., Cohen's $d$) are 0.2, 0.5, and 0.8.

Table 1
*Standardized Vignette Ratings: JPO/SW Ratings % Incorrect (% Deviating From Consensus) by Vignette*

| | Emily (*n* = 252) | Brad (*n* = 130) | Wendy (*n* = 69) | Krista (*n* = 196) | Total (*N* = 647) |
|---|---|---|---|---|---|
| *Summary risk rating* | 6% | 51% | 3% | 20% | 19% |
| Risk factor items | | | | | |
| 1. History of violence | 23% | 9% | 0% | 28% | 20% |
| 2. History of non-violent offending | 37% | 18% | 4% | 23% | 26% |
| 3. Early initiation of violence | 41% | 28% | 44% | 33% | 36% |
| **4. Past supervision/intervention failures** | **56%** | **37%** | **59%** | **69%** | **56%** |
| 5. History of self harm or suicide attempts | 13% | 1% | 42% | 21% | 16% |
| 6. Exposure to violence in the home | 6% | 0% | 25% | 5% | 6% |
| **7. Childhood history of maltreatment** | **17%** | **35%** | **57%** | **62%** | **40%** |
| **8. Parental/caregiver criminality** | **61%** | **34%** | **74%** | **45%** | **52%** |
| **9. Early caregiver disruption** | **63%** | **48%** | **67%** | **77%** | **65%** |
| 10. Poor school achievement | 49% | 39% | 7% | 9% | 30% |
| 11. Peer delinquency | 12% | 18% | 8% | 4% | 10% |
| 12. Peer rejection | 2% | 3% | 23% | 27% | 12% |
| 13. Stress and poor coping | 35% | 6% | 58% | 33% | 31% |
| 14. Poor parental management | 7% | 15% | 3% | 22% | 13% |
| 15. Lack of personal/social support | 18% | 51% | 36% | 29% | 30% |
| 16. Community disorganization | 30% | 42% | — | 3% | 23% |
| 17. Negative attitudes | 16% | 35% | 23% | 43% | 29% |
| **18. Risk taking/Impulsivity** | **62%** | **46%** | **71%** | **27%** | **49%** |
| 19. Substance use difficulties | 20% | 31% | 55% | 2% | 21% |
| 20. Anger management problems | 38% | 6% | 19% | 15% | 23% |
| 21. Low empathy and remorse | 39% | 29% | 3% | 46% | 35% |
| 22. Attention-deficit/hyperactivity difficulties | 62% | 10% | 7% | 13% | 31% |
| 23. Poor compliance | 12% | 36% | 7% | 40% | 25% |
| 24. Low interest/commitment to school | 44% | 39% | 10% | 8% | 28% |
| Protective factor items | | | | | |
| 1. Prosocial involvement | 25% | 15% | 60% | 2% | 20% |
| 2. Strong social support | 28% | 53% | 12% | 32% | 33% |
| **3. Strong attachment and bonds** | **41%** | **55%** | **19%** | **43%** | **42%** |
| **4. Positive attitude intervention/authority** | **33%** | **59%** | — | **40%** | **41%** |
| 5. Strong commitment to school | 49% | 48% | — | 8% | 35% |
| 6. Resilient personality traits | 0% | 22% | 54% | 0% | 10% |
| *Domains: median # of items disagreement* | | | | | |
| SAVRY risk total (out of 24) | 8 | 6 | 7 | 8 | 7 (29%) |
| Historical risk domain (out of 10) | 4 | 2 | 4 | 4 | 3 (30%) |
| Social/contextual risk domain (out of 6) | 1 | 1 | 2 | 1 | 1 (17%) |
| Individual risk domain (out of 8) | 3 | 2 | 2 | 2 | 2 (25%) |
| Protective factors domain (out of 6) | 2 | 3 | 3 | 1 | 2 |

*Note*. Bolded items represent those with 40% or more of JPOs/SWs disagreeing with the expert consensus ratings. Blank cells represent circumstances where there was not sufficient information in the vignette for rating the item.

workshop on the SAVRY by a peer master trainer achieved a level of accuracy that actually outperformed individuals trained in 2-day workshops by an expert with respect to some domain ratings. In fact, JPOs trained by a peer appeared to perform better in the areas that were more difficult for them (i.e., ratings of Social/ Contextual and Protective Factor items) than JPOs trained by an expert.

## Study 2: Inter-Rater Reliability on Youth Probationer Cases

The goal of Study 2 was to examine the inter-rater reliability on the SAVRY as used in practice with randomly selected, live youth probationers. Raters in Study 2 were a subset of Study 1 JPOs. Study 2 data were SAVRY ratings made after JPOs completed the training workshops and post-training vignettes described in Study 1. JPOs conducted their pre- or post-disposition assessments with youth and parents while being observed by a trained research assistant (second rater). Both independently scored the SAVRY using the same file and interview information.

### Method

**Participants.** The raters in this study were 36 JPOs from three of the probation offices included in Study 1 (two local probation offices and one state probation office). The JPOs averaged 35 years of age (*SD* = 9.05 years), 58% were male and most were Black/African-American (61%; White, 39%). JPOs had worked in juvenile justice settings for 6.34 years (*SD* = 5.92 years) on average. The second raters were trained psychology graduate student research associates (RAs) at two offices and a probation officer RA at the third office. The SAVRY ratings completed by the 36 JPOs and the 3 RAs came from a random sample of 80 adjudicated young offenders seen in the probation offices for a pre-disposition report or because they were sentenced to probation. The mean age of young offenders was 15 years ($SD$ = 1.41). Most youths were male (81%) and Black/African American (63%; White, 33%; Biracial/Other, 4%). Research agreements were obtained from each office included in this study and approved by the institutional review board at the University of Massachusetts Medical School.

**Measures.**

***SAVRY.*** When used in practice, each SAVRY item "is rated on a three-level scale, but is not assigned a numerical value" (Borum et al., 2006, p. 12). However, for research purposes, items typically are 'scored' as 0, 1, or 2 (see Borum et al., 2009) to allow for statistical analysis. To be consistent with existing research on the SAVRY, in this study total and risk factor domain scores were calculated by assigning a 0 (Low), 1 (Moderate), or 2 (High) to the 24 risk items and summing the items. Protective factor items were scored as 0 (Absent) or 1 (Present) and summed to yield a total protective factor score. Similarly, SRR ratings were assigned values of 0 (Low), 1 (Moderate), or 2 (High).

**Procedures.**

***Training.*** All of the JPOs completed the 2-day SAVRY workshop by an expert and at least three subsequent practice vignettes described under Study 1. Two of the three RAs completed the exact same training regimen. The third RA was trained individually by one of the investigators. This RA also attended a booster SAVRY training, completed the same practice vignettes, and received individualized feedback. RAs were required to demonstrate an agreement rate of at least 90% on the last of the three practice vignettes.

Each SAVRY was rated based on a review of available files (e.g., court files, treatment records) and interviews conducted by JPOs with the youth, with a parent, and also with the youth and parent together. The interviews were conducted using a uniform semi-structured risk interview created by the researchers. The RA reviewed the same file information as the JPO and observed the JPO's interview conducted in vivo. The JPO and RA rated the SAVRY independently. The majority of JPOs in this sample completed SAVRY ratings for two youth ($n$ = 24, 67%). Ten JPOs completed three SAVRY cases and two JPOs completed only one case.

***Data analysis.*** This design required ICCs using a twoway random effects model for absolute agreement. This version is appropriate for research designs using a random sample of raters selected from a larger population of judges where each rater rates the same subjects or targets (Shrout & Fleiss, 1979). $ICC_1$ denotes the single rater reliability and $ICC_2$ denotes the *averaged rater reliability* (reliability of the averaged rating across raters). Some rater-pairs overlap, but no pair of raters rated the same youth (i.e., ratings from the same JPO/RA pair exist for multiple cases, but each youth was rated only once by a single JPO/RA pair; reliability estimates involved comparisons of JPO and RA scores of the same youth). We calculated $ICC_1$ values as the primary index of reliability (because in practice SAVRY ratings typically are based on ratings of a single JPO). $ICC_2$ values also were calculated to provide an estimate of the potential reliability of averaged ratings (i.e., ratings from more than one JPO for the same youth).

## Results

**Descriptives.** The average total SAVRY score of the 80 youths for ratings made by JPOs was 14.4 ($SD$ = 7.89; range, 0–38). JPOs' ratings on each of the four scales were: Historical ($M$ = 6.04; $SD$ = 3.63; range, 0–19), Social/ Contextual ($M$ = 3.21; $SD$ = 2.19; range, 0–9), Individual ($M$ = 5.15; $SD$ = 3.55; range, 0–15), and Protective ($M$ = 3.51; $SD$ = 1.95; range, 0–6). The average total SAVRY score for ratings made by the RA was 14.93 ($SD$ = 8.11, range, 1–37). RAs' ratings for the scales were: Historical ($M$ = 6.48; $SD$ = 3.70; range, 0–15), Social/Contextual ($M$ = 3.38; $SD$ = 2.19; range, 0–10), Individual ($M$ = 5.08; $SD$ = 3.75; range, 0–16), and Protective ($M$ = 3.35; $SD$ = 1.90; range, 0–6). With respect to the summary risk ratings (SRR), JPOs rated 47.5% of cases Low, 37.5% Moderate, and 15% High risk. RAs rated 41% of cases Low, 41% of cases Moderate, and 17.5% of cases High risk.

**Inter-rater reliability estimates.** The ICCs for the SAVRY indices are listed in Table 2. $ICC_1$ values for the total (.86), Historical (.81), Individual/Clinical (.86), and Protective (.83) Domains were all excellent. The $ICC_1$ for the Social/Contextual Domain (.67) was substantially lower but still fell in the "good" range. All $ICC_2$ values were excellent (all equal to or greater than .80), and followed the same pattern as the $ICC_1$ values. For the SRR, the $ICC_1$ value was good (.71) and the $ICC_2$ value (.83) was excellent.

With respect to the item level, there was a wide range in $ICC_1$ values. Of the 30 risk and protective items, 19 had $ICC_1$ values equal to or greater than .60. As expected, $ICC_2$ values were higher, with 23 of the 30 values being equal to or greater than .70. For the Historical Domain items, $ICC_1$ ranged from .34 (Past Supervision/ Intervention Failures) to .82 (History of Self Harm or Suicide Attempts). $ICC_1$ values were greater than .60 for seven of the 10 items. With respect to the Social/Contextual Domain, $ICC_1$ values ranged from .35 (Stress and Poor Coping) to .76 (Community Disorganization). Only two of the six $ICC_1$ values were greater than .60. For the Individual/Clinical Domain, $ICC_1$ values ranged from .50 (Risk Taking/Impulsivity) to .84 (Substance Use Difficulties). Seven of the eight $ICC_1$ values were equal to or greater than .60. Finally, on the Protective Factor Domain, $ICC_1$ values ranged from .49 (Resilient Personality Traits) to .75 (Strong Attachment and Bonds). Half of the items had $ICC_1$ values greater than .60.

Agreement for the SRR for violence is summarized in Table 3. Of the 34 JPOs who rated two or three cases, 12 (35.3%) had perfect agreement with the RA's SRR, 19 (55.9%) had one disagreement, and 3 (3.8%) had two disagreements. Of the 80 cases, each JPO and RA in the rater-pair gave the same SRR for 55 cases (69% agreement), and there were no major disagreements (low vs. high risk errors).

**Patterns of SRR disagreement by JPO.** It is possible that some JPOs have a particular pattern of assigning risk ratings. Put simply, some JPOs may have a tendency to give all High, all Low, or even all Moderate ratings. In order to investigate whether patterns existed within JPOs, we examined in greater detail the SRR for violence ratings given by the 10 JPOs who rated three SAVRY cases. Most JPOs had some variability in the SRRs they assigned across the three youth assessed. Only three of the 10 JPOs had no variability in their three SAVRY ratings; one JPO rated all

Table 2
*Intra-Class Correlation Coefficients for Juvenile Probation Cases (N = 80)*

| | $ICC_1$ | 95% C.I. | $ICC_2$ | 95% C.I. |
|---|---|---|---|---|
| Summary risk rating (violence) | .71 | .58–.80 | .83 | .73–.89 |
| SAVRY total score | .86 | .79–.91 | .93 | .89–.95 |
| *Historical risk scale* | .81 | .71–.87 | .89 | .83–.93 |
| 1. History of violence | .65 | .46–.77 | .78 | .63–.87 |
| 2. History of non-violent offending | .71 | .58–.80 | .83 | .73–.89 |
| 3. Early initiation of violence | .62 | .47–.74 | .77 | .64–.85 |
| **4. Past supervision/intervention failures** | **.34** | **.14–.52** | **.51** | **.24–.69** |
| 5. History of self harm or suicide attempts | .82 | .74–.88 | .90 | .85–.94 |
| 6. Exposure to violence in the home | .75 | .64–.83 | .86 | .78–.91 |
| 7. Childhood history of maltreatment | .63 | .48–.75 | .77 | .65–.86 |
| **8. Parental/caregiver criminality** | **.56** | **.38–.69** | **.71** | **.55–.82** |
| **9. Early caregiver disruption** | **.52** | **.34–.67** | **.69** | **.51–.80** |
| 10. Poor school achievement | .64 | .49–.75 | .78 | .65–.86 |
| *Social/contextual Risk Scale* | .67 | .53–.77 | .80 | .69–.87 |
| 11. Peer delinquency | .68 | .54–.78 | .81 | .70–.88 |
| **12. Peer rejection** | **.44** | **.24–.60** | **.61** | **.39–.75** |
| **13. Stress and poor coping** | **.35** | **.15–.53** | **.52** | **.26–.69** |
| **14. Poor parental management** | **.48** | **.29–.63** | **.65** | **.45–.77** |
| **15. Lack of personal/social support** | **.54** | **.37–.68** | **.70** | **.54–.81** |
| 16. Community disorganization | .76 | .65–.85 | .87 | .78–.92 |
| *Individual risk scale* | .86 | .79–.91 | .92 | .88–.95 |
| 17. Negative attitudes | .75 | .63–.83 | .86 | .77-.91 |
| **18. Risk taking/impulsivity** | **.50** | **.31–.65** | **.67** | **.48–.79** |
| 19. Substance use diffiulties | .84 | .76–.89 | .91 | .86–.94 |
| 20. Anger management problems | .74 | .62–.83 | .85 | .77–.91 |
| 21. Low empathy and remorse | .70 | .57–.80 | .83 | .73–.89 |
| 22. Attention-deficit/hyperactivity difficulties | .75 | .64–.83 | .86 | .78–.91 |
| 23. Poor compliance | .67 | .53–.78 | .80 | .69–.88 |
| 24. Low interest/commitment to school | .67 | .53–.78 | .80 | .69–.88 |
| *Protective factors* | .83 | .74–.89 | .91 | .85–.94 |
| 1. Prosocial involvement | .74 | .63–.83 | .85 | .77–.91 |
| **2. Strong social support** | **.54** | **.36–.68** | **.70** | **.53–.81** |
| 3. Strong attachment and bonds | .75 | .63–.83 | .85 | .77–.91 |
| 4. Positive attitude toward intervention and authority | .62 | .46–.74 | .77 | .63–.85 |
| **5. Strong commitment to school** | **.54** | **.36–.68** | **.70** | **.53–.81** |
| **6. Resilient personality traits** | **.49** | **.30–.64** | **.66** | **.46–.78** |

*Note*. $ICC_1$ = single-rater ratings; $ICC_2$ = averaged ratings. All values are significant at $p < .01$. Items with $ICC_1$'s less than .61 (a "good" ICC) are in bold.

three cases as Low risk, and the other two JPOs rated all three cases as Moderate risk.

Among the 10 JPOs, only one JPO had perfect agreement with the RA's ratings across all three cases (this was the JPO who assigned ratings of Low for each case). Of the remaining nine JPOs, seven JPOs had one disagreement and two JPOs had two disagreements. Therefore, across the 30 cases rated by these 10 JPOs, there were 11 instances in which the JPO and RA assigned a different SRR for the same youth. Table 3 provides the concordance between JPOs and RAs on the SRR. Disagreements tended to occur slightly more often in the direction of the JPO assigning a higher rating than the RA. Both $ICC_1$ and $ICC_2$ values calculated using a *consistency of ratings* approach were the same as when calculated using *absolute agreement*.

Table 3
*Agreement Between 10 JPOs and Three RAs on Summary Risk Ratings for 80 Young Offenders*

| | RA rating | | | |
|---|---|---|---|---|
| JPO rating | Low | Moderate | High | $Total_{JPO}$ |
| Low | 27 | 11 | 0 | 38 |
| Moderate | 6 | 19 | 5 | 30 |
| High | 0 | 3 | 9 | 12 |
| $Total_{RA}$ | 33 | 33 | 14 | 80 |

*Note*. $ICC_1$ (single-rater ratings) = .71; $ICC_2$ (averaged rat-ings) = .83. $Total_{JPO}$ refers to the row totals (*n* cases) of Low, Moderate, and High ratings made by the JPO. $Total_{RA}$ refers to the column totals (*n* cases) of Low, Moderate, and High ratings made by the RA.

## Discussion

This study demonstrated that good agreement—with no major category errors—was achieved for the final structured professional judgment of low, moderate, or high between JPOs and other examiners who had similar training and access to identical information. According to thresholds cited earlier, results indicated that the inter-rater reliability of JPOs for making SAVRY SRRs was substantial or good. Indices of agreement for the total score were somewhat higher than for the SRR. However, because the SRR is a single item, its reliability would be expected to be lower than that of the total score, which is based on multiple items. The reliability for most SAVRY domains was excellent with the exception of the Social/Contextual domain, which fell in the 'good' range. At the item level, the majority (19) had ICCs in the good range; nine items were in the fair range, and two items had poor reliability. This study provides evidence that juvenile justice personnel, given the proper training and a semi-structured interview script, can reliably rate the SAVRY items and use their structured professional judgment to assign categorical levels of risk for violence.

This study is limited in the information it can provide regarding tendencies of specific juvenile justice personnel to rate all examinees in one direction when using their structured professional judgment because ratings for more than two cases were available for only 10 JPOs. However, among those 10, only three appeared to have a specific tendency in their ratings and one of these matched the ratings of the second examiner. As such, exploration of the prevalence and impact of rater-specific scoring biases with a larger sample of JPOs is warranted.

## General Discussion

This work reports the first comprehensive studies of the field reliability with juvenile justice personnel for the SAVRY, an instrument for which research has reported excellent reliability in laboratory-like research designs. Indeed, to our knowledge this is the first reliability study of a risk for general violence and re-offending instrument to examine both ratings on standardized vignettes and ratings in the field with live cases using juvenile justice personnel. It appears the only such tool for use with young offenders for which field reliability data are available in the peer-reviewed literature is the *Youth Level of Service/Case Management Inventory* (YLS/CMI; Hoge & Andrews, 2006). Comparing professionals and probation officers, Schmidt, Hoge, and Robertson (2005) reported ICCs on the YLS/CMI subscales that ranged from .71 to .85, with the exception of Peer Relations (.61). Our findings for the SAVRY (Study 2) were comparable to that of the YLS/CMI and previous laboratory research on the SAVRY with excellent reliability on the total score ($\text{ICC}_1 = .86$) and most subscales ($\text{ICC}_1$'s ranging .81–.86), with the exception of the Social/Contextual domain ($\text{ICC}_1 = .67$). The reliability of the SRR was good ($\text{ICC}_1 = .71$) and, importantly, there were no major category errors, which is evidence that JPOs were consistent with other examiners in the application of their structured professional judgment. The results of the studies presented here have significant implications for use of the SAVRY in the field with juvenile justice personnel and for the tool developers.

### Implications for use of the SAVRY by Juvenile Justice Personnel

Conceivably, risk assessment tools will have the largest impact in the hands of well-trained probation officers and case managers as opposed to limiting use to mental health professionals. The use of a mental health professional or forensic evaluator for every young offender processed through the juvenile justice system is neither realistic nor cost-effective. Fortunately, the performance of JPOs in Study 1 demonstrated that they were as qualified as SWs to use their structured professional judgment and to rate most areas of the SAVRY given equivalent training. Study 1 demonstrated that performance for the summary risk ratings was excellent across professions and training types, with only 19% of personnel missing the 'correct' SRR on the case vignettes, and Study 2 demonstrated that there was good agreement between trained raters in the field. On the case vignettes, JPOs had greater difficulty than SWs rating items on the Social/Contextual and Protective Factor domains. Nevertheless, overall performance on these items was high across all raters.

Reliability in the field (Study 2) paralleled the case vignette findings where again JPOs appeared to have considerable difficulty in rating Protective Factor and Social/Contextual items, with several showing fair to poor reliability. These particular SAVRY items refer to risk factors such as personal support, poor parenting, and deviant peer involvement. These are areas one would expect JPOs to be skilled at assessing given their educational background and training. Also, they used semi-structured parent and youth interviews in the field that contained questions that should have yielded the requisite data to rate these items. Perhaps JPOs had difficulty obtaining reliable information pertaining to parenting practices and youths' peer groups from parents and youth probationers in the field. Alternatively, the finding may have been the result of a methods effect. Specifically, if the JPOs had previous exposure to any of the youth they assessed on the SAVRY, they would have been privy to details about family and peers that the research associates would not have known. Thus, familiarity with the youth could have adversely affected the agreement between the JPOs and researcher. We did not document prior exposure to cases in Study 2. We suggest field reliability studies consider such information in the future.

Another possible explanation for the relatively poorer agreement for the Protective Factor and Social/Contextual items is that JPOs may have had a difficult time learning how to rate the items from an expert who was not in their field and did not speak their 'language.' All JPOs in Study 2 were trained by an expert rather than a peer. The JPOs who had been trained by a peer master trainer performed as well as SWs at identifying these types of problems on the case vignettes. Indeed, they performed significantly better than SWs and more rigorously trained JPOs on most aspects of rating the SAVRY. This is consistent with findings about peer-training in other types of settings where individuals trained by peers performed equally as well as, and sometimes better than, those trained by professional trainers. Kurtz, Robins and Schork (1997), for example, demonstrated that participants trained in health and safety by a peer trainer identified more closely with the trainer, and reported changing their behavior after training more often than workers trained by professionals. Fremouw and Feindler (1978) found that using a peer model for teaching study skills was just as effective as using a professional model. Peers are often seen as 'insiders' to others in the work place, and individuals that regard someone as similar to them may strive to perform as effectively. The findings of the current study are meaningful because they provide support for the effectiveness of the peer master trainer model, which is more cost-effective and feasible for juvenile justice agencies than bringing in an expert every time staff members require training.

## Implications for the SAVRY Test Manual and Procedures

The combined results from the studies here indicated that certain items on the SAVRY were associated with consistently poorer performance and reliability over multiple cases, both in standardized vignettes and in the field. This is persuasive evidence for the need for clarification of certain item descriptions in the SAVRY test manual. Specifically, Risk Factor items that may benefit from revision include Past Supervision/Intervention Failures, Parental/ Caregiver Criminality, Early Caregiver Disruption, Childhood History of Maltreatment, Risk Taking/ Impulsivity, Stress and Poor Coping, and several items on the Protective Factors domain.

The key to improving field reliability on the Social/ Contextual items also may lie in providing juvenile justice personnel with more guidance as to the proper questions to ask examinees. Implementation of the SAVRY in the probation offices in this study did involve use of a semi-structured interview, which JPOs were required to complete as per office policy. However, the interview may have been lacking in the Social/Contextual domain areas. Any previous exposure to the youth, the youth's peers, or the youth's family could bias the way JPOs rate these items, so more structure might be beneficial. A tested, semi-structured interview specifically for use by juvenile justice personnel might be a step in the right direction. Of course, such an interview script might not be as important in certain settings or for certain professions. For example, in mental health settings, SPJ tools are intended to be integrated into existing clinical practice, and most if not all items should be able to be rated based on a comprehensive clinical interview that typically would be conducted even when not using the tool.

Because the SAVRY is an SPJ tool, the relatively poor inter-rater reliability for certain items is perhaps more important than for actuarially developed schemes, where items are summed into a total score, thereby masking measurement error due to any individual items (generally the more items included in a score, the higher the reliability). In the SPJ model, users are guided to arrive at a final estimate of risk following a consideration of which factors are present and the salience of each individual item for the examinee's risk for violence, making the SRR susceptible to item unreliability. Despite some of the poor item reliability and accuracy shown in both studies, the agreement and accuracy for the SRR was good. As such it does not appear that these items impacted the risk judgments of raters, although a test of the SAVRY's predictive validity will be essential to determine this.

Another potential problem arising from poor item level reliability is the impact on evaluators' recommendations regarding risk management and intervention strategies. Because item level considerations play such an important role when making recommendations regarding risk-reducing interventions, efforts to improve on item level inter-rater agreement (e.g., through better training, manual revision, etc.) are critical. However, this issue is not specific to the SPJ model given some actuarial tools also would use specific items to guide treatment planning.

## Limitations and Future Directions

In Study 1, it is possible that the skills of individuals trained by an expert and those trained by a peer would achieve equivalence with sufficient practice. Unfortunately, we could not conduct adequate analyses for improvement across the three waves of practice cases because the data in Study 1 lacked variability in JPOs' third practice case (i.e., we did not have any third practice cases for JPOs trained by an expert). Based on the limited data we did have, there were no appreciable practice effects. Instead improvement was dependent on the vignette.

Another limitation with the data in Study 1 was that the researchers did not consistently obtain the sliders along with JPOs'/ SWs' ratings, making it impossible to examine absolute accuracy (e.g., a rater would have to give a Moderate+ rating to be considered correct if the consensus rating was also Moderate+, rather than considering both ratings of Moderate and High to be correct). We also were unable to examine differences in performance by race of the rater, which may have accounted for some of the differences between SWs and JPOs given the proportion of Black/ African-American was much higher for SWs than JPOs. In general, examinations of the optimal number of post-training practice cases needed to achieve adequate performance and any potential rater biases due to race of the examinee or rater are crucial areas for future research. Any future investigations should examine rater accuracy using sliders.

Another limitation with this study was that the mere presence of an RA and involvement of researchers might have led to higher reliability on the SAVRY than what would typically be found in the field. The probation offices in these studies were involved in a rigorous implementation process that included a semi-structured interview for staff and completion of multiple case vignettes prior to use of the SAVRY in the field. Although these procedures are consistent with best practice, it is unknown how often agencies adhere to these practices. Additionally, the field inter-rater examination in Study 2 began shortly after the JPOs completed their SAVRY training. It would be interesting to see if there was any deterioration in their ratings over time (*rater drift*) by conducting the same study a year after training. Part of the policy in all of these probation offices was to conduct booster trainings on the SAVRY with all staff every 6 months to prevent rater drift. Examination of the effectiveness of these booster trainings is another area for future research. The best evidence, of course, for the generalizability of findings reported here would be replication in other sites shortly after training, and again a year later.

Finally, it is important to note that although reliability places an upper limit on validity, the concepts are distinct. This study cannot speak to whether the SAVRY field ratings reported here will predict reoffending accurately. However, recent research indicated that SAVRYs completed by juvenile detention personnel were associated with future violent and non-violent re-offending (Vincent, Chapman, & Cook, 2011). An important step in the validation process of the SPJ model of violence risk assessment is the demonstration that evaluators' structured professional ratings regarding level of risk or 'case seriousness' (communicated vis-à-vis low, moderate, or high) are associated with future violence or offending (Douglas & Kropp, 2002). Although commentators (e.g., Quinsey, Harris, Rice, & Cormier, 2006) have expressed concerns regarding the utility of the SRR in terms of its putative subjectivity and relation to future offending, a growing body of research has demonstrated nearly consistently that SRRs are as or more strongly related to violence than the total scores of SPJ tools, or than risk estimates generated by actuarial tools (see Douglas & Reeves, 2010). An important subsequent inquiry will be the association between re-offending and the SRRs made by JPOs in Study 2, which will be explored in future research. However, we must acknowledge that these studies are difficult to conduct after a tool is in place and has been implemented properly. If agencies follow a sound protocol for intervening to reduce risk among youth they identified with the SAVRY as being at moderate or high risk, we would expect to observe a decreased re-offense rate and, consequently, a reduction in the predictive validity of the SAVRY.

## References

Andrews, D. A. (1989). Recidivism is predictable and can be influenced: Using risk assessments to reduce recidivism. *Forum on Corrections Research, 1*(2), 11–18.

Andrews, D. A., & Bonta, J. (2002). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19–52. doi:10.1177/00938548900 17001004

Andrews, D. A., & Dowden, C. (2006). Risk principle of case classification in correctional treatment: A meta-analytic investigation. *International Journal of Offender Therapy and Comparative Criminology, 50,* 88–100. doi:10.1177/030662 4X05282556

Austin, J. (2006). How much risk can we take? The misuse of risk assessment in corrections. *Federal Probation, 70*(2), 58–63.

Barnoski, R. (2004). *Assessing risk for re-offense: Validating the Washington State Juvenile Court Assessment* (Report No. 04-03-1201). Olympia: Washington State Institute for Public Policy.

Barnoski, R., & Markussen, S. (2005). Washington state juvenile court assessment. In T. Grisso, G. Vincent, & D. Seagrave (Eds.), *Mental health screening and assessment in juvenile justice* (pp. 271–282). New York: Guilford Press.

Boccaccini, M. T., Turner, D., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower psychopathy scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, & Law, 14,* 262–283. doi:10.1037/a0014523

Borum, R., Bartel, P., & Forth, A. (2003/2006). *Structured Assessment of Violence Risk in Youth (SAVRY). Odessa, FL: Psychological Assessment Resources, Inc.*

Borum, R., Lodewijks, H., Bartel, P., & Forth, A. (2009). Structured Assessment of Violence Risk in Youth (SAVRY). In K. Douglas & R. Otto (Eds.), *Handbook of Violence Risk Assessment* (pp. 63–80). New York: Routledge.

Chen, B., Zaebst, D., & Seel, L. (2005). A macro to calculate kappa statistics for categorizations by multiple raters [cited 2005 Nov 29]. In SUGI 30 Proceedings, Philadelphia, PA, April 10–13, 2005, from http://www2.sas.com/proceedings/sugi30/155-30.pdf

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,* 127–137.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. doi:10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Douglas, K. S., & Kropp, P. R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior, 29,* 617–658.

Douglas, K. S., & Reeves, K. (2010). The HCR-20 violence risk assessment scheme: Overview and re-view of the research. In R. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 147–185). Oxford: Routledge/Taylor & Francis.

Fagan, J., & Zimring, F. E. (Eds.). (2000). *The changing borders of juvenile justice: Transfer of adolescents to the criminal court.* Chicago: The University of Chicago Press.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5,* 105–112. doi:10.1177/014662168100500115

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fremouw, W. J., & Feindler, E. L. (1978). Peer versus professional models for study skills training. *Journal of Counseling Psychology, 25*(6), 576–580. doi:10.1037/0022-0167.25.6.576

Gottfredson, D., & Tonry, M. (1988). *Prediction and classification: Criminal justice decision-making.* Chicago: Chicago University Press.

Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference*, pp. 1110–1115.

Griffin, P., & Bozynski, M. (2003). *National overviews: State juvenile justice profiles.* Retrieved November 5, 2003, from http://www.ncjj.org/stateprofiles/

Grisso, T. (2005). Why we need mental health screening and assessment in juvenile justice programs. In T. Grisso, G. Vincent, & D. Seagrave (Eds.), *Mental health screening and assessment in juvenile justice* (pp. 3–21). New York: Guilford Press.

Grisso, T., Vincent, G. M., & Seagrave, D. (2005). *Mental health screening and assessment in juvenile justice.* New York: Guilford Press.

Hare, R. D. (2003). *Manual for the Hare Psychopathy Checklist— revised* (2nd ed.). Toronto: Multi-Health Systems.

Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior, 29,* 380–396. doi:10.1177/0093854802029004003

Hoge, R. D., & Andrews, D. A. (2006). *Youth Level of Service/Case Management Inventory: User's manual.* North Tonawanda, NY: Multi-Health Systems.

Kurtz, J. R., Robins, T. G., & Schork, M. A. (1997). An evaluation of peer and professional trainers in a union-based occupational health and safety training program. *Journal of Occupational and Environmental Medicine, 39*(7), 661–671.

Landis, J., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data. Biometrics, 33,* 159–174.

Lodewijks, H. P. B., Doreleijers, T. A. H., & de Ruiter, C. (2008). SAVRY risk assessment in violent Dutch adolescents: Relation to sentencing and recidivism. *Criminal Justice and Behavior, 35,* 696–709. doi:10.1177/0093854808316146

Mulvey, E. P. (2005). Risk Assessment in Juvenile Justice Policy and Practice. In K. Heilbrun, N. E. Sevin Goldstein, & R. E. Redding (Eds.), *Juvenile delinquency: Prevention, assessment, and intervention* (pp. 209–231). New York: Oxford University Press.

Murrie, D. C., Boccaccini, M., Johnson, J., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in Sexually Violent Predator trials suggest partisan allegiance in forensic evaluation? *Law and Human Behavior, 32,* 352–362. doi:10.1007/s10979-007-9097-5

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior, 36,* 329–353. doi:10.1177/0093854809331457

Otto, R. K., & Douglas, K. S. (Eds.). (2009). *Handbook of violence risk assessment.* New York: Routledge/Taylor & Francis Group.

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

Schmidt, F., Hoge, R., & Robertson, L. (2005). Reliability and validity analyses of the Youth Level of Services/Case Management Inventory. *Criminal Justice and Behavior, 32*(3), 329–344. doi:10.1177/0093854804274373

Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31,* 449–462. doi:10.1007/ s10979-006-9071-7

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86,* 420–428. doi: 10.1037/0033-2909.86.2.420

Vincent, G. M., Chapman, J., & Cook, N. E. (2011). Risk/Needs assessment in juvenile justice: Predictive validity of the SAVRY, racial differences, and contribution of needs factors. *Criminal Justice and Behavior, 38*(1), 42–62. doi:10.1177/009385481 0386000

Vincent, G. M., Terry, A., & Maney, S. (2009). Risk/Needs tools for antisocial behavior and violence among youthful populations. In J. Andrade (Ed.), *Handbook of Violence Risk Assessment and Treatment for Forensic Mental Health Practitioners* (pp. 337–424). New York: Springer.

Welsh, J., Schmidt, F., McKinnon, L., Chattha, H., & Meyers, J. (2008). A comparative study of adolescent risk assessment instruments: predictive and incremental validity. *Assessment, 15,* 104–115.