Assessing Risk of Reoffending in Adolescents Who Have Committed a Sexual Offense: The Accuracy of Clinical Judgments After Completion of Risk Assessment Instruments

Natasha Elkovitch
University of Nebraska–Lincoln, nelkovitch@huskers.unl.edu

Jodi L. Viljoen
Simon Fraser University, jviljoen@sfu.ca

Mario J. Scalora
University of Nebraska-Lincoln, mscalora1@unl.edu

Daniel Ullman
Lincoln Regional Center, Lincoln, NE

Assessing Risk of Reoffending in Adolescents Who Have Committed a Sexual Offense: The Accuracy of Clinical Judgments After Completion of Risk Assessment Instruments

Natasha Elkovitch, M.A., University of Nebraska—Lincoln

Jodi L. Viljoen, Ph.D., Simon Fraser University

Mario J. Scalora, Ph.D., University of Nebraska—Lincoln

Daniel Ullman, Ph.D., Lincoln Regional Center

Corresponding author — Natasha Elkovitch, Department of Psychology, University of Nebraska—Lincoln, Lincoln, NE 68588, U.S.A; email nelkovit@bigred.unl.edu

Abstract

As courts often rely on clinicians when differentiating between sexually abusive youth at a low versus high risk of reoffense, understanding factors that contribute to accuracy in assessment of risk is imperative. The present study built on existing research by examining (1) the accuracy of clinical judgments of risk made after completing risk assessment instruments, (2) whether instrument-informed clinical judgments made with a high degree of confidence are associated with greater accuracy, and (3) the risk assessment instruments and subscales most predictive of clinical judgments. Raters assessed each youth's (n = 166) risk of reoffending after completing the SAVRY and J-SOAP-II. Raters were not able to predict detected cases of either sexual recidivism or nonsexual violent recidivism above chance, and a high degree of rater confidence was not associated with higher levels of accuracy. Total scores on the J-SOAP-II were predictive of instrument-informed clinical judgments of sexual risk, and total scores on the SAVRY of nonsexual risk.

Introduction

Although adult men commit the majority of sexual assaults, a significant minority of sexual assaults are committed by youth. Some estimates indicate that 16% of forcible rapes are committed by youth under the age of 18 (Snyder, 2005). The Federal Bureau of Investigation's National Incident-Based Reporting System has indicated that offenders under the age of 18 are responsible for approximately one of every five sexual assaults (Snyder, 2005), and one of every three sexual assaults involving victims under the age of 18 (Snyder, 2001). Additionally, youth are responsible for 40% of the sexual assaults involving children under the age of 6 (Snyder, 2001).

However, the literature varies considerably regarding the rate at which youth sexually reoffend. While some studies have reported low sexual reoffense rates (see Hecker, Scoular, Righthand, & Nangle, 2002; Martinez, Flores, & Rosenfeld, 2007; Mazur & Michael, 1992; Parks & Bard, 2006; Smith & Monastersky, 1986), others have reported relatively high sexual recidivism rates (see Gretton, McBride, Hare, O'Shaughnessy & Kumka, 2001; Långström, 2002; Rubenstein, Yeager, Goodstein, & Lewis, 1993). Worling and Långström (2003) point out that these discrepant findings are likely a function of sampling and methodological differences, including the characteristics of adolescents under investigation, the type and impact of interventions, the measurement of recidivism (e.g. whether the outcome is dichotomized), and the length of the follow-up period. However, despite the variation in reoffense rates across samples, the literature indicates there is indeed a population that is of high risk for future violence (Barbaree, Hudson, & Seto, 1993; Righthand et al., 2005).

The identification of youth who may be at increased risk of reoffending has been referred to as one of the ''highest-profile tasks'' of mental health professionals (Lewis & Webster, 2004, p. 401). Courts often rely on clinicians when differentiating between youth at a low risk versus high risk of reoffense, and decisions based on evaluations can have a great impact (Prentky & Righthand, 2003). When examining the potential of future offending, risk estimates often inform the courts regarding prosecution, detention placement decisions, level of security necessary, and when, if ever, a youth can be released back into the community (Hoge, 2002; Prentky & Righthand, 2003). Youth's risk for future violence is routinely considered in decisions regarding transfer of youth to adult court (*Kent v. United States,* 1966), and in a number of states adolescent sex offenders believed to be at high risk for future violence are placed on sex offender registries (Caldwell, 2002; Trivits & Reppucci, 2002). For example, North Carolina currently requires registration of youth age 11 and older, but only for those who were found by the court to be a danger to the community (N.C. Gen. Stat. §§14-208.26(a)–(a1), 2005).

Despite the frequency with which clinicians are asked to make judgments about adolescents' risk of violence, serious concerns have been expressed regarding clinician's ability to accurately assess risk for violence (see *Barefoot v. Estelle,* 1983; Menzies, Webster, McMain, Staley, & Scaglione, 1994; Monahan, 1992). Early reviews of the research on prediction of dangerousness suggest that mental health professionals have limited ability to predict future violent behavior. In fact, some literature contends that, at best, clinical judgments made with unstructured prediction methods have a ''modest, better-than-chance level of accuracy'' (Mossman, 1994, p. 790). More recent reviews, however, indicate that accurate

predictions of violence are indeed attainable (Borum, 1996; Douglas & Webster, 1999; Otto, 2000). For example, a number of studies have supported the predictive validity of adult risk assessment instruments that follow the actuarial and structured professional judgment model (e.g. Static-99, Hanson & Thornton, 1999; HCR-20, Webster, Douglas, Eaves, & Hart, 1997).

While several instruments designed for use with adolescents appear promising (e.g. Juvenile Sex Offender Assessment Protocol, J-SOAP-II, Prentky & Righthand, 2003; Structured Assessment of Violence Risk in Youth, SAVRY, Borum, Bartel, & Forth, 2003), research on the predictive validity of these instruments is limited and studies have yielded mixed findings. For example, Viljoen and colleagues (2008) found that, while the total score on the J-SOAP-II predicted nonsexual violence, it did not significantly predict sexual recidivism. In other samples of sexually abusive adolescents, however, the J-SOAP-II has been able to significantly predict sexually reoffending (see, e.g., Martinez et al., 2007; Prentky, 2006). Nonetheless, in clinical practice, even when clinicians use risk assessment instruments, in the end, it may still be seen as necessary or desirable to come to a clinical judgment about the youth. Many instruments designed for youth, such as the J-SOAP-II, do not have cut-off scores for classifying a youth as low or high risk, and in fact, at the present time, the authors of the J-SOAP-II recommend that judgments regarding youths' risk of reoffending not be made exclusively on the basis of their JSOAP-II scores (Righthand et al., 2005). Thus, it is unclear whether, when given the option, clinicians will make a judgment that differs from the degree of risk indicated by available instruments. In addition, guidelines often encourage mental health professionals to gather information from various sources before deriving overall conclusions (e.g. Committee on Ethical Guidelines for Forensic Psychologists, 1991), and there may be a belief held by mental health professionals that by using multiple instruments one may be able to compensate for the limitations of a single instrument (see Righthand et al., 2005; Seto, 2005). While Seto (2005) did not find that combining actuarial risk scales (using three analytical approaches) in the prediction of adult sexual recidivism was advantageous, considering multiple instruments when making judgments may be useful in the prediction of adolescent recidivism. The primary aim of the present research is to examine the predictive validity of clinical judgments of risk after completion of both the J-SOAP-II and the SAVRY. We refer to these judgments as instrument-informed clinical judgments.

While recent research has investigated conditions under which predictions of violence are more accurate or less accurate (Douglas & Ogloff, 2003), a relatively understudied, but potentially important, construct in this regard is the confidence with which risk judgments are made (Douglas & Ogloff, 2003; McNeil, Sandberg, & Binder, 1998; Rabinowitz & Garelik-Wyler, 1999). While Rabinowitz and Garelik-Wyler (1999) did not find a significant confidence–accuracy relationship in a sample of adult psychiatric patients, studies with larger cell sizes have, perhaps surprisingly, found evidence that confidence may indeed be associated with the accuracy of risk assessments. In a sample of adult psychiatric patients, McNeil et al. (1998) found a significant relationship between confidence and accuracy of physicians' short-term clinical estimates of risk. Douglas and Ogloff (2003) also found a significant relationship between confidence and the accuracy of risk judgments made with actuarial and structured professional assessment models in a sample of adult forensic patients.

These studies have important implications for clinical and legal practice and suggest that risk assessment decisions paired with low clinician confidence should be given particular consideration and re-evaluation (Lewis & Webster, 2004). Clinicians who express a high degree of confidence in their judgments about whether an adolescent will sexually reoffend may be very convincing to judges, juries, probation officers, and treatment teams who must make decisions regarding youths' treatment and placements. However, to date, no research has examined the confidence–accuracy relationship in an adolescent forensic sample. The second aim of the present research is to fill this void and examine the relationship between confidence and accuracy in rater's instrument-informed clinical judgments for youth's risk for post-release sexual and nonsexual violence.

While examining factors that contribute to mental health professionals' accuracy, it is also important to examine how clinicians use available risk assessment instruments to formulate final risk judgments. A large literature has focused on how clinicians should use risk assessment instruments (e.g. Borum, 1996; Viljoen, Elkovitch, & Ullman, 2007); however, little research has examined how clinicians actually utilize instruments when formulating risk judgments. Elbogen, Huss, Tompkins, and Scalora (2005) emphasize that it "will not matter whether a particular risk factor predicts violence if no effort is made to determine whether or how those risk factors actually are used in practice" (p. 134). To date, no research has examined factors that are predictive of clinicians' judgments of risk in an adolescent forensic population. Thus, the final aim of the present research is to examine to what extent risk assessment instruments, including instrument subscales, are used by raters when formulating final risk judgments.

The Current Study

Data for the current study are part of a larger study on the predictive validity of the JSOAP-II, an instrument designed to assess risk of sexual and nonsexual violence in adolescents who have sexually offended, and the SAVRY, an instrument designed to assess risk of violence more broadly (Viljoen et al., 2008). The current study builds on existing research by examining three questions. First, how accurate are clinical judgments of risk made after completing two risk assessment instruments, the JSOAP-II and the SAVRY? Based on the findings of previous work (Viljoen et al., 2008), we hypothesize that clinical judgments of nonsexual violence made after completing the instruments would be accurate, as these instruments predicted serious nonsexual violence. However, as the instruments did not significantly predict sexual violence (Viljoen et al., 2008), it is uncertain whether clinical judgments of sexual violence made after using these instruments would be accurate. It is possible that clinical judgments would be able to compensate for limitations in total scores on these instruments.

Second, are instrument-informed clinical judgments of risk made with a high degree of confidence more accurate? While recent research has suggested that higher confidence may be associated with increased accuracy in adult violence risk assessments, to date no research has examined this question in the context of risk assessments with an adolescent forensic population. Thus, it is unclear whether, in this sample, this positive relationship will emerge.

Third, what factors are most predictive of instrument-informed clinical judgments of risk? The J-SOAP-II was specifically designed to assess risk for violence in adolescents with a history of sexual offending or sexually coercive behavior, while the SAVRY was designed to assess risk for violence in adolescents more broadly. Thus, raters may place more weight on the J-SOAP-II total scores and its subscales when formulating judgments regarding sexual risk, and more weight on the SAVRY and its subscales for judgments regarding nonsexual risk. On the other hand, because raters completed both instruments, it is possible that raters may consider both instruments when formulating risk judgments.

Method

Participants

The present study included 166 male adolescents who had been referred to a residential sex offender program in the Midwestern United States. To be admitted into the treatment program, all youth had to undergo an interview and meet admission criteria, including appropriate age (between 13 and 17 years), intellectual and adaptive functioning at least at the borderline level, adjudicated delinquent of a sexual offense and mandated to receive treatment, and demonstrated self-control that would allow functioning in an open, unlocked treatment program. Further, to be included in our analyses, a period of at least 250 days must have elapsed since program completion so that youth could be followed to assess reoffending. These 166 youth include nearly all youth who had been admitted to the program between 1994 and 2005. The mean age of youth at admission was 15.31 years (SD = 1.51) and the majority of youth's index offenses were perpetrated against youth who were at least 3 years younger than themselves (79.5%, n = 132). Further descriptive information regarding the sample is reported in Table 1.

Procedures

Graduate student raters assessed risk of reoffending for each participant using two risk assessment instruments, the Structured Assessment of Violence Risk in Youth (SAVRY; Borum et al., 2003) and the Juvenile Sex Offender Assessment Protocol-II (J-SOAP-II; Prentky & Righthand, 2003), both of which are described below. These ratings were completed based on reviews of comprehensive psychological file information (averaging 800–1500 pages in length) which included psychological evaluations and on-going psychological assessments (e.g. Behavioral Assessment System for Children, BASC; Reynolds & Kamphaus, 1992), psychiatric reports, medical and psychopharamalogical evaluations, social work reports (including annual clinical interviews with the youths' guardian(s) and caseworker(s)), treatment plans, therapy progress notes, teacher assessments and school records, criminal information, and legal reports.

Prior to commencing the study, the graduate student raters underwent a thorough training procedure that involved both didactic and applied procedures. Raters read the SAVRY and J-SOAP-II manuals and were assigned several key readings on risk assessment as well as the specific instruments. After training, raters

Table 1. Descriptive information on sample

Characteristic	N or M	(% or SD)
Age at admission (years)	15.39	(1.51)
Age at discharge (years)	16.14	(1.52)
Race		
Non-Hispanic Caucasian	138	(83.1%)
African-American	14	(8.4%)
Hispanic	8	(4.8%)
American Indian or Alaskan Native	2	(1.2%)
Type of index offense*		
Genital penetration	61	(36.7%)
Anal penetration	59	(35.5%)
Digital penetration	21	(12.7%)
Oral–genital contact	82	(49.4%)
Fondling	101	(60.8%)
Exhibitionism	23	(13.9%)
Age of victim		
3 or more years younger than perpetrator	132	(79.5%)
Similar age (2 yrs younger to 2 yrs older)	19	(11.4%)
3 or more years older than perpetrator	9	(5.3%)
Sex of victims		
Males only	40	(21.4%)
Females only	76	(45.8%)
Both males and females	47	(28.3%)

* Does not add up to 100% as some youth had multiple index offenses.

independently completed five practice files, which were then discussed in order to examine discrepancies. Specifically, raters discussed discrepancies on individual items on the SAVRY and J-SOAP-II; neither total scores nor instrument-informed clinical judgments were discussed as part of these meetings.

Subsequent juvenile and adult arrest and legal processing records for each youth were obtained from state law enforcement sources. Taking into account the possibility of dispositional bargaining, we measured charges, rather than convictions (see Doren, 1998). Recidivism was also measured through unofficial means, including treatment records. For the present research, reoffenses were divided into sexual felonies, nonsexual violent felonies, and any violent offense. Length of follow-up ranged from 280 days to 12.01 years, with an average of 6.58 years (SD = 3.49). Raters were blind to outcome regarding whether the youth reoffended.

Measures

Juvenile Sex Offender Assessment Protocol-II (J-SOAP-II)

The Juvenile Sex Offender Assessment Protocol-II (J-SOAP-II; Prentky & Righthand, 2003) is a 28-item ''checklist whose purpose is to aid in the systematic review of risk factors that have been identified in the professional literature as being associated with sexual and criminal offending'' (p. 1). It is intended for use for males between the ages of 12 and 18 years who have a history of sexual offending or sexually coercive behavior. It is important to note that the J-SOAP-II does not distinguish between risk of sexual reoffense and risk of general, criminal reoffense. The instrument contains four subscales: Sexual Drive/Preoccupation, Impulsive/ Anti-

social Behavior, Intervention, and Community Stability/Adjustment. Items on the J-SOAP-II are rated on a three-point scale (absent, possibly present, clearly present), with a higher score representing greater risk. While total scores are obtained by summing the items on the four scales, at the present time the authors state that cut-off scores should not be used, and the J-SOAP-II should function as an ''empirically informed guide'' (Prentky & Righthand, 2003, p. 8). While there is limited information on the psychometric properties of the J-SOAP-II, recent research provides preliminary evidence for its concurrent validity, construct validity, and inter-rater reliability (Parks & Bard, 2006). In a sample of sexually abusive urban minority youth, Martinez and colleagues (2007) found the J-SOAP-II total score was correlated with treatment compliance and was able to predict both general and sexual reoffending. In addition, the Sexual Drive/Preoccupation scale (Hecker et al., 2002) as well as the Impulsive/Antisocial Behavior scale (Parks & Bard, 2006) have been shown to significantly predict sexual recidivism. Further, a modified scale consisting of eight of the nine J-SOAP-II Impulsive/Antisocial Behavior items has been shown to significantly predict general rearrest (Waite et al., 2005), and a total score consisting of 26 of the 28 J-SOAP-II items has been able to predict sexual reoffense in both preadolescents and adolescents (Prentky, 2006). Other studies, however, have reported difficulties in predicting sexual reoffending with this instrument (Caldwell, Vitacco, & Ziemke, 2008; Viljoen et al., 2008).

The Structured Assessment of Violence Risk in Youth (SAVRY)

The Structured Assessment of Violence Risk in Youth (SAVRY; Borum et al., 2003) is a guide for assessing risk for general violence in adolescents. Although it does not specifically assess risk for sexual offending, it includes sexual violence in its definition of violence. The SAVRY is based on the structured professional judgment model of risk assessment and is composed of 24 risk items, as well as six protective factors. These items make up three sets of risk factors: Historical Risk Factors, Social and Contextual Risk Factors, and Individual Risk Factors. Each risk factor is coded as ''high,'' meaning a youth is high risk on this item, ''moderate,'' or ''low.'' The SAVRY also assesses whether six protective factors are ''present'' or ''absent.'' Consistent with the structured professional judgment model, clinicians are to use item scores on risk and protective factors in order to make a professional judgment about a youth's risk for violence.

Research indicates the instrument has adequate psychometric properties (Borum, Bartel, & Forth, 2003; Catchpole & Gretton, 2003; McEachran, 2001; Parks & Bard, 2006). In terms of predictive validity, the SAVRY has been shown to significantly predict both general and violent recidivism (Catchpole & Gretton, 2003; Dolan & Rennie, 2008). Welsh, Schmidt, McKinnon, Chattha, and Meyers (2008) found that in a sample of juvenile offenders the SAVRY was able to predict both general and violent reoffending above and beyond that of the Psychopathy Checklist: Youth Version (PCL-YV; Forth, Kosson, & Hare, 2003) and the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge, Andrews, & Leschied, 2002). In addition, the SAVRY has also been found to significantly predict physical violence, rule violations and verbal threats in a sample of youth in residential treatment (Lodewijks, Doreleijers, de Ruiter, & Borum, 2008). Finally, in a sample of sexually abusive adolescents, Viljoen and colleagues (2008) found a SAVRY total score was able to predict nonsexual recidivism.

Instrument-Informed Clinical Judgments

After completing both the SAVRY and J-SOAP-II, raters made a clinical judgment as to whether each youth was at low, moderate, or high risk of engaging in post-release sexual violence and nonsexual violence. Raters were instructed that, while they could use the SAVRY and J-SOAP-II to guide their judgments, they could place as little or as much weight on the instruments as they wanted.

Confidence Ratings

For each category (sexual violence risk, nonsexual violence risk, overall violence risk), raters expressed their confidence in the estimate of violence risk on a Lik-erttype rating scale that ranged from 1 (''not at all confident'') to 10 (''extremely confident''). It is important to note that risk estimates and confidence judgments were made independently. That is, raters could make a high-risk judgment with low confidence, or a low-risk judgment with high confidence. Consistent with Douglas and Ogloff (2003), we performed a median split in order to form two confidence groups for each type of violence (overall violence, sexual violence, and non-sexual violence). For each type of violence, the median confidence rating was 7. Therefore, ratings of 7 and below were categorized as ''low confidence,'' and ratings of 8 and above were classified as being made with ''high confidence.''

Data Analysis

Univariate Analyses

Spearman–Brown correlations were used in order to examine the strength of the relationship between instrument-informed clinical judgments and violent crime at follow-up. In addition, point-biserial correlations were used in order to illustrate the strength of the relationship between instrument-informed clinical judgments and violent recidivism across confidence groups (low, high).

Receiver Operating Characteristic Analyses

Although correlations permit comparison with other published studies, this index of accuracy is confounded with the base rate of the criterion (e.g. post-release sexual or non-sexual violence; Mossman, 1994). Thus, receiver operating characteristic (ROC) analyses were also performed to evaluate the ability of instrument-informed clinical judgments to predict recidivism, as well as to evaluate the predictive validity of these judgments across confidence groups. ROC statistics have been recommended to evaluate the accuracy of recidivism predictions because they are less influenced by the base rate of violence, easily interpreted, and can be graphically represented, thus permitting visual comparisons. By examining the area underneath the curve (AUC), ROC analysis describes how likely the prediction was to be greater than chance. A perfectly accurate prediction would yield an AUC of 1.0, while an AUC of .5 indicates a chance prediction, and an AUC of .0 indicates a perfectly negative prediction. AUC's between .5 and 1.0 signify a better than chance predictive accuracy.

Logistic Regression Analyses

When examining the factors most predictive of instrument-informed clinical judgments of risk, we chose to examine both instrument total scores and instrument subscale scores. We believe an examination of subscale scores will provide a more detailed picture of factors that influence clinical judgment. For the purposes of our analyses, instrument-informed clinical judgments of risk were split into two categories: low (which includes low and moderate ratings of risk), and high (which just includes ratings of high risk).

Binary logistic regression analyses, based on hierarchical methods, were used to examine the relative contribution of the J-SOAP-II and the SAVRY total scores and subscales to instrument-informed clinical judgments of risk. Separate analyses were performed with instrument total scores, and then with instrument subscale scores. For each analysis, either the J-SOAP-II and SAVRY total scores or their subscales were entered in a forward stepwise fashion with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates. The criterion for entry was set at 0.05 and the criterion for removal was set at 0.10. The stepwise procedure was used because of the exploratory nature of the analysis.

Results

Inter-Rater Reliability

For a random sample of 21.9% of the cases (n = 37), intraclass correlation coefficients for single raters (ICC1s) were calculated. Using a two-way random effect model (McGraw & Wong, 1996), the ICC1s for instrument-informed clinical judgments of nonsexual violence and sexual violence fell in the ''excellent'' range (Cicchetti & Sparrow, 1981). Specifically, the ICC1s for nonsexual violence and sexual violence were .78 and .79, respectively. In addition, J-SOAP-II and SAVRY total scores were found to be highly correlated with instrument-informed clinical judgments (see Table 2).

Table 2. Correlations between instrument total scores and instrument-informed clinical judgments

Instrument total scores	Clinical judgments after completion of the J-SOAP and SAVRY	
	Sexual violence	Nonsexual violence
J-SOAP-II	.71**	.61**
SAVRY	.68**	.63**

** $p < .01$.

Post-Release Offenses

For the purpose of our analyses, we looked at three categories of post-release charges: sexual assault, serious nonsexual violent offenses (felonies), and any violent offense (sexual assault, nonsexual violent felonies, and nonsexual violent misdemeanors). Felony and misdemeanor nonsexual violent offenses included assault, aggravated assault, strangulation, attempted homicide, and robbery. At the end of follow-up (M = 80.01 months, SD = 42 months), 8.4% (n = 14) of the youth were charged with sexual felonies, 10.2% (n = 17) were charged with nonsexual violent offenses, and 20.5% (n = 34) had been charged with any violent offense.

Risk and Confidence Classification

The percentage of adolescents classified as low, moderate, and high risk after completing the risk assessment instruments is presented in Table 3. Raters most commonly rated youth as being moderate risk on each of the three outcomes. In addition, raters were more likely to give high confidence ratings when rating a youth as high risk, compared with when they rated a youth as low or moderate risk.

Are Clinical Judgments Made After Completing Risk Assessment Instruments Accurate?

Spearman–Brown correlations and ROC analyses were conducted to investigate our first question, whether instrument-informed clinical judgments were able to predict recidivism in the sample of youth. Results of Spearman–Brown correlations between risk predictions and follow-up variables were all non-significant. In addition, the ROC AUCs indicated a near-zero relationship between risk classification and violent recidivism, including sexual recidivism. Raters were no more accurate than chance in predicting either sexual recidivism or nonsexual violent recidivism. These results are reported in Table 4.

Table 3. Number of youth classified with low, moderate, and high confidence

Risk rating	Violence category		
	Any violence N (%)	Sexual violence N (%)	Nonsexual violence N (%)
Low risk	43 (25.4)	35 (20.7)	70 (41.4)
Low confidence	19 (11.2)	20 (11.8)	38 (22.5)
High confidence	23 (8.9)	14 (8.3)	31 (18.3)
Moderate risk	93 (55.0)	90 (53.3)	72 (42.6)
Low confidence	61 (36.1)	65 (38.5)	47 (27.8)
High confidence	31 (18.3)	25 (14.8)	24 (14.2)
High risk	33 (19.5)	44 (26.0)	27 (15.4)
Low confidence	8 (4.7)	21 (50.0)	11 (40.7)
High confidence	25 (14.8)	21 (50.0)	16 (59.3)

Low and high confidence groups may not add up to 100% due to missing data.

Table 4. Accuracy of instrument-informed clinical judgments of risk

Structured risk rating	Violence category		
	Any violence	Sexual violence	Nonsexual violence
Spearman–Brown correlations			
Risk prediction	.08	–.06	.09
Areas under ROC curve			
Risk prediction	.53	.44	.58

No data reported are significant.

Are Clinical Judgments Made with a High Degree of Confidence More Accurate?

The next sets of analyses were conducted to evaluate our second question, whether predictions made with a high degree of confidence will be more accurate than those made with a lower degree of confidence. First, point-biserial correlations were used in order to illustrate the strength of the relationship between structured professional judgments of risk for violence across confidence groups (low, high). Although correlations permit comparison with other published studies, ROC analyses were also performed, as they are less sensitive to the base rate of recidivism. As shown in Table 5, correlations were near zero and nonsignificant across confidence groups. Likewise, the AUC values for judgments of risk were routinely near chance (.42–.65) across confidence groups. That is, a high degree of rater confidence was not associated with more accurate judgments of risk.

Which Instruments are Most Predictive of Clinical Judgments of Risk?

Binary logistic regression analyses were used to investigate our third question, whether certain instruments or instrument subscales are associated with raters' instrument-informed clinical judgments of risk. When instrument total scores

Table 5. Impact of confidence on instrument-informed clinical judgments of risk

Structured risk rating	Violence category		
	Any violence	Sexual violence	Nonsexual violence
Point-biserial correlation			
Confidence ≤ median	–.05	.05	–.08
Confidence > median	.16	.04	–.14
Areas under ROC curve			
Confidence ≤ median	.47	.46	.56
Confidence > median	.62	.42	.65

No data reported are significant. Point-biserial correlations were used since the three risk categories (low, moderate, high) are continuously rated, and violence outcomes (recidivated or not) are dichotomously rated.

were examined, only the total score on the SAVRY significantly explained raters' judgments of nonsexual violence risk (Wald's χ^2 = 7.01, β = .19, p < .01, OR = 1.18), while only the total score on the J-SOAP-II significantly explained judgments of risk for future sexual violence (Wald's χ^2 = 10.21, β = .16, p = .001, OR = 1.18). Both models accounted for 27.3% of the variance in instrument-informed clinical judgments of risk.

When subscales from the J-SOAP-II and SAVRY were entered into a binary logistic regression, only the Individual/Clinical Factors subscale from the SAVRY (Wald's χ^2 = 19.67, β = .48, p < .001, OR = 1.67) and the Sexual Drive/Preoccupation from the J-SOAP-II (Wald's χ^2 = 13.60, β = .29, p < .001, OR = 1.33) made independent contributions to the model explaining rater's ratings of risk for sexually reoffending and accounted for 43.4% of the variance. With regard to clinical ratings of risk for nonsexual violence, only the Impulsive/Antisocial Behavior subscale of the J-SOAP-II (Wald's χ^2 = 6.58, β = .41, p < .01, OR = .41) and the Individual/ Clinical Factors subscale from the SAVRY (Wald's χ^2 = 8.15, β = .28, p < .01, OR = 1.51) made significant contributions to the risk ratings, and explained 35.8% of the variance.

Discussion

The current research is the first known study to examine the accuracy of instrument-informed clinical judgments of risk in a sample of youth that have sexually offended. While studies have examined the predictive validity of single instruments, little is known about how clinicians combine information from various instruments in order to make risk judgments. Secondary aims of the current study included examining both the role of rater confidence on the accuracy of risk assessment judgments, and the instrument(s) and instrument subscales most predictive of raters' clinical risk judgments.

In our study, raters made judgments regarding risk of reoffending after completing a comprehensive file review, as well as two widely used youth violence risk assessment measures, the SAVRY and the J-SOAP-II. A subsample of youth was coded by two raters in order to permit inter-rater reliability. Perhaps surprisingly, inter-rater reliability of instrument-informed clinical judgments was excellent. This suggests that even after the completion of two risk assessment instruments and the opportunity to consider case-specific factors raters are coming to similar risk conclusions.

Results of our first question revealed that instrument-informed clinical judgments did not significantly predict *sexual* recidivism. This is perhaps not surprising, given that, using the same sample, total scores on both the J-SOAP-II and the SAVRY were unable to predict future sexual violence (Viljoen et al., 2008). However, this finding also suggests that raters' ability to consider idiographic, case-specific factors did not make up for the limitations of these instruments. This finding is troublesome, as mental health professionals are consistently asked to make such risk judgments.

Furthermore, instrument-informed clinical judgments did not predict future *nonsexual* violent offending, although total scores on the J-SOAP-II and SAVRY did (Viljoen et al., 2008). This finding is surprising particularly since these judg-

ments were made after using effective instruments. One possible explanation is that clinical biases or beliefs might lead raters to place too much weight on certain subscales. While this might suggest that a more structured approach is needed, this finding must be interpreted with caution. A number of empirical studies have demonstrated that structured clinical judgments based on empirically validated instruments show a high degree of accuracy in predicting future violence (e.g. Worling, 2004; Douglas, Ogloff, & Hart, 2003), and even add incremental validity to actuarial assessments (Douglas et al., 2003).

The results of our second question indicated that rater confidence was not associated with more accurate instrument-informed clinical judgments of risk. This finding is troublesome, as clinicians are often asked by the court to evaluate risk for reoffending in youth that have committed sexual offenses, and mental health practitioners who are highly confident in determinations of levels of risk may be extremely convincing to judges and courts. Such judgments of risk often inform the courts at various stages of the legal process, including placement decisions, and if and when the youth can return to his family or the community. In our study, raters' overall confidence was quite high for both sexual and nonsexual violence judgments. The use of instruments may, in fact, give raters a false sense of confidence. While further research is needed to examine this relationship, our findings, the first to examine this relationship in this population, suggest that the degree of clinician confidence should not influence judicial and court decisions with regard to youth that have committed a sexual offense.

Our third question was largely exploratory. In order to examine how raters use risk assessment instruments to make overall clinical judgments, we examined the total scores and subscales of the J-SOAP-II and the SAVRY. The total score on the JSOAP-II significantly contributed to rater judgments of sexual risk, while the total score on the SAVRY significantly contributed to rater judgments of nonsexual risk. While these data suggest that raters relied heavily on the instruments to derive overall risk judgments, a large amount of variance remained unaccounted for, suggesting that raters placed weight on factors outside of the instruments.

Further, as noted above, the Impulsive/Antisocial Behavior subscale of the JSOAP-II and the Individual/Clinical Factors subscale from the SAVRY made independent contributions to the model explaining raters' judgments of nonsexual risk, while the Sexual Drive/Preoccupation subscale of the J-SOAPII and the Individual/Clinical subscale of the SAVRY significantly corresponded to raters' sexual risk ratings. This suggests that raters are not necessarily utilizing all scales on these instruments, but instead focus on several scales they may consider particularly critical. However, the scales and factors on which raters focus may not necessarily have strong predictive validity. Rater attention to certain subscales may be due to inaccurate beliefs that certain factors or subscales are more predictive of future violence than are other factors or subscales or due to item salience. For example, rater attention to the J-SOAP-II Sexual Drive/ Preoccupation subscale may be due to the item content on this scale. Consistent with the salience or vividness heuristic (see Nisbet & Ross, 1980; Reyes, Thomson, & Bower, 1980), perhaps information tapped by this scale is more vivid, provocative, or intuitively linked to sexual offending, resulting in raters making judgments that are overdependent on specific salient risk items. Items on this scale include, for example, duration of sexual offending, number of victims, and hypersexuality.

On the other hand, with additional validation studies across different samples and settings, certain subscales on the J-SOAP-II and the SAVRY may indeed emerge as predictive, and thus rater attention to these scales may be warranted. For example, to date, although many believe that a lack of victim empathy is related to an increased risk of sexually reoffending, the empirical literature has been unable to validate this belief in adolescent populations. However, as Worling and Långström (2003) note, perhaps with additional research or with improved measurement techniques, empathy may be found to relate to sexual recidivism.

Limitations and Future Research

There are several limitations to the current study. First, consistent with the literature, our study yielded a low base rate of sexual reoffending (8.4% of the sample), which makes accurate prediction of reoffending difficult. Contributing to this low base rate problem may be (1) underreporting and (2) use of criminal records for follow-up. Sexual offenses are not always reported to law enforcement agencies; in fact, Marshall and Barbaree (1990) report that unofficial sources show a rate of recidivism 2.4 times greater than that of official, criminal records. Due to these difficulties (underreporting, nature of follow-up), it is likely that our reoffense rates are a conservative estimate.

Second, raters relied solely on file information to make violence risk decisions. While important strengths of the present study include the extremely extensive files and a greater period of follow-up than nearly every prospectively designed study to date, dynamic factors may be particularly difficult to assess without a clinical interview. There is a need for more prospective research on this population, particularly research that includes file reviews, clinical interviews, and contact with collateral sources, such as caregivers and teachers. Further research should also assess youth throughout treatment, such that dynamic factors and their contribution to risk for future violence can be examined.

Third, risk and confidence judgments were made by three Masters-level graduate student raters. While the nature of these coders (graduate students, not clinicians with professional qualifications in applied settings) should be taken into account when interpreting findings, there is reason to presume a reasonable degree of generalizability (see Douglas & Ogloff, 2003): (1) The raters in our study had clinical experience in forensic settings (at least 1,000 hours) conducting violence risk-assessments, (2) they received didactic and applied training on conducting risk assessments with adolescents (including reading and reviewing the relevant empirical literature), perhaps more so than the average mental health professionals in an applied setting, and (3) there is some evidence to suggest that judgments by individuals with varying levels of experience do not vary significantly (Andreason et al., 1982; Grove & Meehl, 1996). In addition, it is important to note that the raters in our study all had a Master's level training and may be similar to early career individuals with the same or similar degree (e.g. MSW) who often conduct forensic evaluations for the courts. Thus, while some generalizability was sacrificed by using a limited number of student raters, there is reason to believe the results may be generalizable to other settings in which adolescent violence risk assessments are conducted. Nonetheless, future research should ex-

amine instrument-informed clinical judgments and confidence levels according to level of experience.

Finally, because we were interested in how raters used risk assessment instruments, our examination of factors predictive of instrument-informed judgments of risk was limited to J-SOAP-II and SAVRY subscales. That is, we did not examine other clinical cues and case-specific factors that may have impacted raters' risk judgments. In addition to instrument subscales, it may be important to assess whether clinicians believe that certain case-specific factors, or ''broken leg cases,'' override risk assessment instruments. As originally termed by Meehl (1954) and described by Monahan (2006), ''broken leg cases'' refer to those circumstances, such as a broken leg, that may affect an individual's likelihood of reoffending. Similarly, it may be important to assess whether clinicians hold certain beliefs or assumptions regarding adolescents' risk for future violence, or violence risk assessment in general, which may in turn influence risk judgments.

Summary

Our study was the first to examine clinical judgments guided by several risk assessment instruments, and the factors most associated with these judgments. We found that after using several risk assessment measures raters were unable to accurately predict sexual and nonsexual violent recidivism. Rater confidence in judgments was not associated with greater predictive accuracy. Given the frequency with which mental health professionals are asked to make judgments regarding an adolescents' risk for future violence, it is imperative that future research examines the validity of instruments designed to predict recidivism. Based on our findings, further research should investigate the factors that most influence clinicians' violence risk judgments, as well as the factors that contribute to increased accuracy of these judgments. By examining how clinicians use and weigh risk assessment instruments, we may be better able to discern how well risk assessment technology has impacted clinical practice (Elbogen et al., 2005), and how we can best train those mental health professionals who are charged with performing violence risk assessments.

Acknowledgments

References

Andreason, N. C., McDonald-Scott, P., Grove, W. M., Keller, M. D., Shapiro, R. W., & Hirschfeld, R. M. A. (1982). Assessment of reliability in multi-center collaborative research using a videotape approach. American Journal of Psychiatry, 139, 876–882.

Barbaree, H. E., Hudson, S. M., & Seto, M. C. (1993). Sexual assault in society: The role of the youth offender. In H. E. Barbaree, W. L. Marshall, & S. M. Hudson (Eds.), The youth sex offender (pp. 1–24). New York: Guildford.

Barefoot v. Estelle, 463 US 880. (1983).

Borum, R. (1996). Improving the clinical practice of violence risk assessment. American Psychologist, 51, 945–956.

Borum, R., Bartel, P., & Forth, A. (2003). Manual for the Structured Assessment of Violence Risk in Youth, Version 1.1. University of South Florida.

Caldwell, M. F. (2002). What we do not know about youth sexual reoffense risk. Child Maltreatment, 7, 291–302.

Caldwell, M., Vitacco, M., & Ziemke, M. (2008, March). Will Adam Walsh work? A preliminary examination of SORNA as applied to juveniles. Paper presented at the annual meeting of the American Psychology–Law Society, Jacksonville, FL.

Catchpole, R., & Gretton, H. (2003). The predictive validity of risk assessment with violent young offenders: A 1 year examination of criminal outcome. Criminal Justice and Behavior, 30, 688–708.

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.

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty guidelines for forensic psychologists. Law and Human Behavior, 15, 655–665.

Dolan, M. C., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorder. Psychological Assessment, 20, 35–46.

Doren, D. (1998). Recidivism base rates, predictions of sexual offender recidivism, and the ''sexual predator'' commitment laws. Behavioral Sciences and the Law, 16, 97–114.

Douglas, K. S., & Ogloff, J. R. P. (2003). The impact of confidence on the accuracy of structured professional judgments and actuarial violence risk judgments in a sample of forensic psychiatric patients. Law and Human Behavior, 27, 573–587.

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. Psychiatric Services, 54, 1372–1379.

Douglas, K. S., & Webster, C. D. (1999). Assessing risk of violence in mentally and personality disordered individuals. In R. Roesch, & S. D. Hart, & J. R. Ogloff (Eds.), Psychology and law: The state of the discipline (pp. 175–239). New York: Plenum.

Elbogen, E. B., Huss, M. T., Tompkins, A. J., & Scalora, M. J. (2005). Clinical decision making about psychopathy and violence risk assessment in public sector mental health settings. Psychological Services, 2, 133–141.

Forth, A. F., Kosson, D. S., & Hare, R. D. (2003). The Psychopathy Checklist: Youth Version. Toronto: Multi-Health Systems.

Gretton, H. M., McBride, M., Hare, R. D., O'Shaughnessy, R., & Kumka, G. (2001). Psychopathy and recidivism in adolescent sex offenders. Criminal Justice and Behavior, 28, 427–449.

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

Hanson, R. K., & Thornton, D. (1999). Static 99: Improving actuarial risk assessments for sex offenders (User Report 1999-02). Ottawa: Department of the Solicitor General of Canada.

Hecker, J., Scoular, J., Righthand, S., & Nangle, D. (2002, October). Predictive validity of the J-SOAP over 10-plus years: Implications for risk assessment. Paper presented at the Annual Meeting of the Association for Treatment of Sexual Abusers, Montreal.

Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. Criminal Justice and Behaviour, 29, 380–396.

Hoge, R., Andrews, D. A., & Leschied, A. (2002). Youth Level of Service/Case Management Inventory: YLS/ CMI manual. Toronto: MultiHealth Systems.

Kent v. United States, 383 US 541. (1966).

Långström, N. (2002). Long-term follow-up of criminal recidivism in youth sex offenders: Temporal patterns and risk factors. Psychology, Crime and Law, 8, 51–58.

Lewis, A. H. O., & Webster, C. D. (2004). General instruments for risk assessment. Current Opinion in Psychiatry, 17, 401–406.

Lodewijks, H. P. B., Doreleijers, T. A. H., de Ruiter, C., & Borum, R. (2008) Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. International Journal of Law and Psychiatry, 31, 263–271.

Marshall, W. L., & Barbaree, H. E. (1990). Outcome of comprehensive cognitive-behavioral treatment programs. In Marshall, W. L., Laws, D. R., & Barbaree, H. E. (Eds.), Handbook of sexual assault: Issues, theories and treatment of the offender (pp. 363–385). New York: Plenum Press.

Martinez, R., Flores, J., & Rosenfeld, B. (2007). Validity of the Juvenile Sex Offender Assessment Protocol (J-SOAP-II) in a sample of urban minority youth. Criminal Justice and Behavior, 34, 1284–1295.

Mazur, T., & Michael, P. M. (1992). Outpatient treatment for adolescents with sexually inappropriate behavior: Program description and six month follow-up. Journal of Offenders Rehabilitation, 18, 191– 203.

McEachran A. (2001). The predictive validity of the PCL-YU and the SAVRY in a population of adolescent offenders. Unpublished master's thesis. Simon Fraser University, Burnaby, British Columbia.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30–46.

McNeil, D. E., Sandberg, D. A., & Binder, R. L. (1998). The relationship between confidence and accuracy in clinical assessment of psychiatric inpatients' potential for violence. Law and Human Behavior, 22, 655–669.

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

Menzies, R. J., Webster, C. D., McMain, S., Staley, S., & Scaglione, R. (1994). The dimensions of dangerousness revisited: Assessing forensic predictions about criminality and violence. Law and Human Behavior, 18, 1–28.

Monahan, J. (1992). Mental disorder and violence behavior: Perceptions and evidence. American Psychologist, 47, 511–521.

Monahan, J. (2006, April). Risk assessment. Workshop presented at the Nebraska Psychological Association Spring Conference, Omaha, NE.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. Journal of Consulting and Clinical Psychology, 62, 783–792.

N.C. Gen. Stat. §§14-208. 26(a)–(a1). (2005).

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice-Hall.

Otto, R. K. (2000). Assessing and managing violence risk in outpatient settings. Journal of Clinical Psychology, 56, 1239–1262.

Parks, G. A., & Bard, D. E. (2006). Risk factors for adolescent sex offender recidivism: Evaluation of predictive factors and comparison of three groups based upon victim type. Sexual Abuse: A Journal of Research and Treatment, 18, 319–342.

Prentky, R. (2006). Risk management of sexually abusive youth: A follow up study. Rockville, MD: National Institute of Justice/NCJRS; http://www.ncjrs.gov/pdffiles1/nij/grants/214261.pdf

Prentky, R., & Righthand, S. (2003). Youth Sex Offender Assessment Protocol-II (J-SOAP-II): Manual. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

Rabinowitz, J., & Garelik-Wyler, R. (1999). Accuracy and confidence in clinical assessment of psychiatric inpatients' risk of violence. International Journal of Law and Psychiatry, 22, 99–106.

Reyes, R. M., Thomson, W. C., & Bower, G. H. (1980). Judgmental biases resulting from different availabilities of arguments. Journal of Personality and Social Psychology, 2, 5–12.

Reynolds, C. R., & Kamphaus, R. W, (1992). Behavior Assessment System for Children. Circle Pines, MN: American Guidance Service.

Righthand, S., Prentky, R., Knight, R., Carpenter, E., Hecker, J. E., & Nangle, D. (2005). Factor structure and validation of the Youth Sex Offender Assessment Protocol (J-SOAP). Sexual Abuse: A Journal of Research and Treatment, 17, 13–30.

Rubenstein, M., Yeager, M. A., Goodstein, B. A., & Lewis, D. O. (1993). Sexually assaultive male juveniles: A follow-up. American Journal of Psychiatry, 150, 262–265.

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. Psychological Assessment, 17, 156–167.

Smith, W. R., & Monastersky, C. (1986). Assessing youth sex offenders grown up: Recidivism in young adulthood. Criminal Justice and Behavior, 25, 109–124.

Snyder, H. (2001). Youth Arrests 2000. Washington, DC: Office of Youth Justice and Delinquency Prevention.

Snyder, H. (2005). Youth Arrests 2003. Washington, DC: Office of Youth Justice and Delinquency Prevention.

Trivits, L. C., & Reppucci, N. D. (2002). Application of Megan's law to juveniles. American Psychologist, 57, 690–704.

Viljoen, J. L., Elkovitch, N., & Ullman, D. (2007). Assessing risk for violence in adolescents. In B. Jackson (Ed.), Learning forensic assessment (pp. 385–414). New York: Erlbaum.

Viljoen, J. L., Scalora, M., Cuadra, L., Bader, S., Chavez, V., Ullman, D., & Lawrence, L. (2008). Assessing risk for violence in adolescents who have sexually offended: A comparison of the J-SOAP-II, SAVRY, and J-SORRAT-II. Criminal Justice and Behavior, 35, 5–23.

Waite, D., Keller, A., McGarvey, E. R., Wieckowski, E., Pinkerton, R., & Brown, G. L. (2005). Juvenile sex offender re-arrest rates for sexual, violent nonsexual, and property crimes: A 10-year follow-up. Sexual Abuse: A Journal of Research and Treatment, 17, 313–331.

Webster, C., Douglas, K., Eaves, D., & Hart, S. (1997). HCR-20 assessing risk for violence: Version II. Burnaby, British Columbia: Mental Health, Law and Policy Institute, Simon Fraser University.

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

Worling, J. R., & Långström, N. (2003). Assessment of criminal recidivism risk with adolescents who have offended sexually: A review. Trauma, Violence and Abuse, 4, 341–362.

Worling, J. R. (2004). The Estimate of Risk of Adolescent Sexual Offense Recidivism (ERA-SOR): Preliminary psychometric data. Sexual Abuse: A Journal of Research and Treatment, 16, 235–254.