Law and Human Behavior, Vol. 28, No. 4, August 2004 (© 2004)

Assessing the Risk of Domestic Violence Reoffending: A Validation Study[1]

Kirk R. Williams[2,4] **and Amy Barry Houghton**[3]

*Little research has been conducted to validate available instruments for assessing the risk of domestic violence reoffending, especially research using some form of prospective design. This study uses a prospective design to determine the reliability and validity of the Domestic Violence Screening Instrument (DVSI). The analysis is based on a sample of 1,465 male domestic violence offenders selected consecutively over a 9-month period. Data on reoffending were collected in a 6-month follow-up period from a subsample of the victims (*N = *125) of these perpetrators and from official records for all perpetrators during an 18-month follow-up period. The empirical results suggest that the DVSI was administered reliably, and they provide significant evidence of the concurrent, discriminant, and predictive validity of this instrument. Implications for further research and utilization of the DVSI are discussed.*

KEY WORDS: Domestic violence risk assessment; domestic violence reoffending; validating assessments of risk.

A vivid historical movement in the United States has been the criminalization of violence in intimate settings (Fagan, 1996; Gosselin, 2000; National Institute of Justice & American Bar Foundation, 1998; Pleck, 1987). For example, the battered women's movement successfully campaigned to heighten public awareness about the dangers lurking "behind closed doors" and clearly had a substantial influence on recasting battering and other forms of abuse as "crimes of domestic violence" (Gosselin, 2000; Mignon, Larson, & Holmes, 2002; Schecter, 1982). Such efforts were fortified by a political climate welcoming the nascent research evidence that legal interventions might curtail domestic violence reoffending (Binder & Meeker, 1992a, 1992b; Garner,

[1]This research was part of the larger Colorado Domestic Violence Risk Reduction Project.
[2]Department of Sociology and Family Research Laboratory, University of New Hampshire, Durham, New Hampshire.
[3]Legal Advocacy Fund, NW Washington, DC.
[4]To whom correspondence should be addressed at Department of Sociology and Family Research Laboratory, University of New Hampshire, Durham, New Hampshire 03824; e-mail: krwncy@att.net.

Fagan, & Maxwell, 1995; Garner & Maxwell, 2000; Maxwell, Garner, & Fagan, 2001; Sherman, 1992; Sherman & Berk, 1984; Sherman & Cohn, 1989; Williams & Conniff, 2002).

The wave of changes in criminal justice policy swamped the system with domestic violence cases. Officials were faced with making difficult decisions about case-processing issues, ranging from arrest and prosecution to court disposition, case management, and treatment. Challenging questions arose: What should guide decision making? And, how can decisions be reached that ensure public safety by protecting victims, supervising and treating offenders, and yet maximizing the efficient use of scarce criminal justice resources? Attempts to answer those questions inevitably lead to the issue of assessing risk, that is, determining which offenders are more likely to repeat their violence in the future and/or escalate its severity, perhaps even to lethal proportions.

The most appropriate method for assessing the risk of future violence has been debated extensively by those involved in prediction research, especially studies bearing on mental disorder and violent behavior (Grisso & Appelbaum, 1992, 1993; Litwack, 1993; Webster, 1990), with the primary focus being on the ethics, utility, and predictive validity of clinical judgment versus actuarial risk assessment instruments (Litwack & Schlesinger, 1999; Monahan, 1981, 1996, 1997; Monahan et al., 2001; Quinsey, Harris, Rice, & Cormier, 1998). Some suggest that the two methods can be combined (Swets, Dawes, & Monahan, 2000). The debate has extended to the most appropriate methods of analyzing data bearing on risk and their relation to the central criterion variable—violent behavior (Monahan et al., 2001; Monahan & Steadman, 1994; Rice & Harris, 1995; Swets et al., 2000). This debate has also emerged in discussions of the most appropriate methods of assessing the risk of domestic violence (e.g., Campbell, 1995; Weisz, Tolman, & Saunders, 2000).

The debate has been productive for sharpening guidelines in the development and use of risk assessment instruments, but prediction studies of violence may not completely resolve the debate because of the increasing fusion of clinical and statistical data. The distinction between clinical and actuarial risk assessments is clear in terms of how they are developed, with the former derived primarily from clinical training and experience, with actuarial techniques involving statistical procedures documenting independent relations between risk factors and repeated violence (Swets et al., 2000). However, the distinction is becoming increasingly blurred (Litwack & Schlesinger, 1999), especially in terms of the kinds of information included in risk assessment instruments and how they are used to assess risk. For example, rather than relying on strictly "objective" quantitative data, "subjective" clinically based observations have been incorporated into some of the most widely recognized actuarial assessment instruments (Swets et al., 2000), such as the Violence Risk Appraisal Guide (Harris, Rice, & Quinsey, 1993) or the iterative classification tree models developed in the MacArthur Study of Mental Disorder and Violence (Monahan et al., 2001). One of the risk assessment instruments analyzed in the present study, the Spousal Assault Risk Assessment (SARA) Guide (Kropp & Hart, 2000), also blends quantitative and clinically based data. Further, rather than determining clinical decisions through a mechanical, formulaic application of probabilistic

estimates of risk generated by actuarial assessments, such estimates are often merely one additional piece of information used to guide clinical decision making. Thus, although informed by actuarial techniques, clinical judgment persists in the assessment of risk (Litwack & Schlesinger, 1999).

The present study was not conducted under the pretense of resolving the debate, at least as applied to risk assessment and domestic violence reoffending. It cannot do so because the instruments analyzed do not maintain clear and sharp distinctions between clinical and actuarial risk assessment, either in their content or their use. Nonetheless, the research findings reported have relevance for the debate, as discussed below. The position taken here is that the technique or combination of techniques constituting the most valid and reliable method for assessing risk is ultimately an empirical question. Unfortunately, that question has not been adequately answered through well-designed validation studies of risk assessment and domestic violence reoffending. Granted, instruments for assessing "dangerousness" (i.e., the potential lethality of violence in intimate settings) or the risk of domestic violence reoffending have been developed and utilized in the field (Campbell, 2000; Dutton & Kropp, 2000; Gondolf, 2002; Roehl & Guertin, 1998; Saunders & Hamill, 2002). However, empirical studies designed to validate such instruments have been scarce. Research on the Dangerousness Guide (Campbell, 1995), the SARA (Kropp & Hart, 2000), and "survivors' predictions" (Weisz et al., 2000) are exceptions, but no empirical studies have utilized prospective designs to determine the predictive accuracy of domestic violence risk assessment instruments (e.g., Campbell, 2000). The present study addresses this gap in the research literature by using a prospective design to validate the Domestic Violence Screening Instrument (DVSI), developed in the Colorado Domestic Violence Risk Reduction Project.

RESEARCH PROCEDURES

The discussion of research procedures is organized in three sections: (1) the development of the Domestic Violence Screening Instrument (DVSI); (2) the nature and use of the Spousal Assault Risk Assessment (SARA) Guide; and (3) the selection of pilot sites, the sample of participants, and data collection methods.

Measurement: The DVSI

The Colorado Department of Probation Services (DPS), with input from domestic violence researchers and others in the community created the DVSI. To determine the items for inclusion, DPS staff initially conducted analyses of data collected on more than 9,000 domestic violence cases sentenced to probation between 1994 and 1996 and evaluated through a 34-item clinical assessment guide used previously in Colorado. This analysis identified the most common social and behavioral characteristics of perpetrators having a history of repeated intimate partner violence. Additional support for the empirical connection between those characteristics and

repeated intimate partner violence was found through a literature review of empirical studies documenting the statistical relations between the identified social and behavioral characteristics and the prevalence and incidence of intimate partner violence (a list of the published studies supporting the inclusion of items on the DVSI is available upon request from the authors). The behavioral items bear on previous family and nonfamily violence arrests, convictions, imposition of court orders, and noncompliance with court and/or probation orders. Social characteristics were limited to employment status and recent separation between the victim and perpetrator. The initial analyses and the literature review were also used to determine the scoring categories, with items having stronger associations with intimate partner violence given greater weight. Once the DVSI instrument was created, DPS convened focus groups with judges, probation officers, prosecutors, defense attorneys, law enforcement personnel, and representatives of the victim community (public and private advocates) to determine whether representatives from these constituencies felt the DVSI had adequate content and face validity. The instrument was revised and finalized, with some new items added upon recommendation from participants in the focus groups, specifically, the presence of weapons or children during the domestic violence incident. The focus groups also assisted in corroborating the scoring categories for each item based on the initial analyses and the literature review. See Appendix for a copy of the DVSI.

Colorado has a mandatory arrest provision (CRS 18-6-803.6), and the statute further requires that every domestic violence offender must remain in custody until he or she appears before a judicial officer. Hence, pressures exist for the expeditious processing of cases. Accordingly, the DVSI was designed to be a short statistical tool based on a quick criminal history review that could be made available to prosecutors, judges, and probation officers soon after an offender is arrested.

The 12 items on the DVSI can be completed by a review of state and national databases as well as prior court and probation records. The project also used information collected by Pre-Trial Services Programs and defendants' bond applications. Records reviewed included police reports, NCIC (National Crime Information Center), CCIC (Colorado Crime Information Center), Blackstone (a database used and maintained by Colorado Prosecutors), and the court and probation databases. Administration of the DVSI did not include interviews with defendants in order to avoid constitutional concerns about communication with the defendant prior to adjudication.

Probation Officers (POs) were trained by the DPS Project Manager on the administration of the DVSI prior to commencing data collection. A specific set of coding instructions for each of the items included in the DVSI was used in the training. The POs conducted the record reviews at the time of arrest throughout the project to acquire the data needed to complete the DVSI. Each item on this instrument has a set of response categories ranging from 0 to 2 or 0 to 3, depending on the item. The higher categories were created for those referring to restraining order history (previous restraining orders, previous violations of restraining orders, and a restraining order in place at the time of the presenting offense), presence of weapons or children during the incident, and other community supervision at the time of the incident. The evaluator uses the coding instructions to score each item. Once all

12 items are scored, s/he sums the scores across all 12 items, and the higher the total score the higher the risk for reoffending, noncompliance with court, and probation orders, and thus, the higher the risk to victims (see Appendix). The theoretical range of total scores on the DVSI is from 0 to 30. The DVSI evaluations were completed with the sample of 1,465 men in the four pilot judicial districts. Once the records were reviewed, the 12 items on the DVSI forms were scored as described above, with the actual range of total scores being 0 to 26, with a mean of 6.5 and a standard deviation of 4.8. Twenty-five percent of the men scored nine or higher on the DVSI.

Measurement: The SARA

The SARA was developed and validated by Kropp and Hart (2000). It is designed to assist professionals formulate appropriate case management strategies for domestic violence offenders and thus was used as a secondary assessment instrument in this project. The SARA is a combination of static (fixed and unchangeable) and dynamic (variable) factors related to the risk of reoffending. It was administered postadjudication, with the administration involving an interview with the defendant, an interview with the victim, and collection of collateral material (e.g., record reviews). DPS staff attended a training workshop in Denver conducted by the authors of the SARA (Kropp & Hart, 2000), and subsequently, POs were training by the DPS Project Manager. Like the DVSI, POs administered the SARA throughout the project for all cases remaining under state supervision.

The instrument provides numerical coding of 20 items similar to a statistical prediction tool, but it also provides for clinical judgement by allowing the evaluator to identify "critical items" and to provide a subjective summary risk rating for reoffending against the victim as well as some other relative or person known to the perpetrator. Further, some of the items included in the SARA require some degree of clinical expertise for accurate completion (e.g., recent psychotic or manic symptoms, personality disorder with anger, impulsivity, or behavioral instability). Hence, the SARA is based on a combination of statistical methods and clinical judgement and thus is referred to as a "structured professional" guide (Kropp & Hart, 2000).

The instrument serves as a measure to determine the concurrent and discriminant validity of the DVSI (Carmines & Zeller, 1979; Neuman, 1994). However, a SARA assessment was not conducted on the entire 1,465 offenders in the sample. The analysis using the total SARA scores was conduced with the 434 cases for which requisite data were available, typically cases kept under state supervision and subjected to a secondary screening. An examination of this subsample, compared to those not receiving a SARA assessment, suggested no significant evidence of sample bias regarding the key variables used in this analysis. Specifically, a logistic regression was conducted with a dummy-dependent variable scored 0 if no SARA assessment was done and one if such an assessment was completed. The results of the regression revealed no significant differences between these two subsamples in terms of DVSI risk scores, domestic violence offending, or other criminal offending prior or subsequent to adjudication.

Pilot Sites, Sample of Subjects, and Data Collection Methods

Colorado is divided into 22 judicial districts, and the project was implemented in four pilot judicial districts that varied in their economics, size, and cultural composition. Approximately 9,000 cases had been filed annually in the pilot districts, representing a majority of the case fillings in the state of Colorado. Each district has a different process for handling cases, different resources for offender management, and different community partners.

The study included two samples of participants. The first is 1,465 male offenders arrested for domestic violence offenses committed against female partners. The focus is on male offenders in this pilot study because they are disproportionately represented among cases coming into the system and thus appear to be the most relevant for initially informing judicial decision making. Extending this research to women perpetrators is a logical next step. These men were included in the sample by consecutively selecting all domestic violence offenders arrested during the 9-month period from July 1997 through March 1998 in the four pilot judicial districts. A review of all criminal records during an 18-month follow-up period after court disposition was completed for the entire sample of 1,465 men. This procedure for collecting data on reoffending avoids the problem of sample attrition that plagues prospective research designs. Specific information was collected on each offender's history of domestic violence offending, DUI offending, restraining orders, and other criminal involvement. The sample has the following demographic characteristics: average age at sentencing was 32, racial and ethnic composition includes 49.5% Anglo, 5.6% African American, 43.7% Latino, and 1.2% other. Thirty-five percent of the men in this sample had previous domestic violence convictions and/or restraining order violations.

The second sample of participants consists of the female partners, who were available and willing to participate in the study, of the men arrested. They were identified at the time of arrest when the male offenders were selected for inclusion in study. Because DPS honored constitutional concerns about communication with defendants prior to adjudication, thus their victims, and because the Human Research Committee at the University of Colorado – Boulder approved interviews with victims only by the independent research team at the university because of safety concerns for human participants, DPS could not assist in providing information about victims at the time of arrest. Consequently, various strategies were implemented to identify and recruit women victims by that research team. The eligible women victims were initially selected randomly for interviews conducted 6 months after court disposition. They were offered financial compensation to participate in the study; but regardless, locating them and soliciting their willingness to participate were extremely difficult tasks, resulting in a small sample size ($N = 125$). Nonetheless, data from the interviews are informative and vital for assessing the predictive validity of the DVSI. Furthermore, like the empirical examination to determine bias in the sample of men receiving a SARA assessment, an analysis was conducted with the total sample of men to identify any evidence of bias due to attrition. Specifically, logistic regression was conducted with a dummy-dependent variable coded 0 if no victim interview was done and 1 if such an interview was completed. The results of that regression revealed

no statistically significant differences between the interviewed and noninterviewed samples in terms of risk scores on the SARA and DVSI as well as domestic violence offending, domestic violence restraining orders, or other types of criminal offending prior and subsequent to adjudication. Moreover, the demographic characteristics of the men connected with these women were virtually identical to the total sample of men (see above).

Victim interviews were conducted with assurances of safety and confidentiality for the women. The Human Research Committee at the University of Colorado, Boulder approved the data collection protocol for human participants' protection. Interviews were conducted by telephone, using a questionnaire that included questions from a nationally recognized survey of victims of intimate partner violence (e.g., Tjaden & Thoennes, 1998). Representatives of local battered women's programs trained interviewers in working with victims of domestic violence.

DATA ANALYSIS

The primary objective of the data analysis is to determine the reliability and validity of the DVSI. The analysis is organized into four sections. The first presents prevalence rates for each of the 12 items on the DVSI. The second section reports on a test that approximates interrater reliability, and the third section assesses the concurrent and discriminant validity of the DVSI. The last section reports the findings bearing on the predictive validity of this instrument, using 18-month follow-up data involving record checks of reoffending and 6-month follow-up interviews of women victims. The analysis on predictive validity also includes a comparison of the predictive accuracy of the DVSI and the SARA.

Descriptive Statistics on the DVSI

The prevalence of each item on the DVSI within the sample of 1,465 male offenders is presented in Table 1. It shows the percentage distribution of the 1,465 offenders scoring greater than 0 on any of the 12 items of the DVSI. On the basis of these percentages, the prevalence is highest for prior nondomestic violence conviction, children present, and prior arrest for assault, harassment, and menacing. The prevalence is lowest for a history of violating restraining orders, evidence of an object used as a weapon, and a restraining order at the time of offense. Because the POs completing the instrument relied upon various data sources, some jurisdictions had access to more data sources than did others. This resulted in missing data on some items for some of the men in the sample. The percent missing was relatively low for most items (1–2%) and was highest for those items that might not be adequately covered in a criminal history record check (e.g., 7.7% for child present during incident, 6.3% for employment status, and 10.5% for separation in last 6 months). Decreasing the amount of missing data and identifying the most complete and accurate data sources would likely increase the validity of the DVSI, and such issues should be addressed in future use of the instrument. However, serial mean substitution for missing information was used in the analyses below (e.g., Little & Rubin, 1987). This

Table 1. Percentage of 1,465 Male Perpetrators Scoring Above 0 on Each Item of the DVSI

12 DVSI items	% Scoring on the item
Prior non-DV convictions	66%
Prior assault, harassment, menacing	51%
Prior DV treatment	31%
Prior drug or alcohol treatment	39%
History of DV-related restraining orders	29%
History of DV restraining order violation	14%
Object used as weapon in commission of crime	23%
Children present during DV incident	49%
Current employment status (unemployed)	36%
Separation from victim in last 6 months	32%
Did victim have restraining order at time of offense	12%
Defendant under community supervision at time of offense	25%

procedure maximizes the use of data from the full sample. Moreover, comparing the results involving DVSI scores with missing data to those without missing data (i.e., serial mean substitution) revealed no significant differences in the results.

A Quasi-Interrater Reliability Assessment of the DVSI

The reliability of indices or scales measuring a psychological disposition or some other state or trait of an individual is often evaluated in terms of test–retest stability, parallel forms of a given measure, item homogeneity, or internal consistency (Carmines & Zeller, 1979; Goodwin, 2001; Neuman, 1994). However, risk assessment instruments are designed to estimate the likelihood of future behavior, in this case, domestic violence reoffending. Kropp and Hart (2000: p. 109) suggest that such test for reliability "are of limited use" in evaluating risk assessment instruments. They argue that "assuming all the items in a risk scale are predictive of future behavior, the optimal situation would be to have . . . maximal predictive validity with minimal redundancy" among the items in the scale. This point accurately describes the logic behind the development of statistical prediction rules in general (Swets et al., 2000). Regardless, Kropp and Hart (2000: p. 109) claim that interrater reliability is more important because "if raters cannot agree on the presence of individual risk factors or the implications that can be drawn from them, there is little point in conducting risk assessments."

The DVSI has comparable internal consistency ($\alpha = .71$) to that reported for the Part 1 (general violence) and Part 2 (spousal violence) total SARA scores ($\alpha = .66$ and .73, respectively). Unfortunately, the logistics of implementing the DVSI in the larger risk reduction project precluded collecting data on interrater reliability. Hence, an alternative analysis was done to approximate such an assessment. Specifically, the DVSI has two important features that lend themselves to a quasi-interrater reliability analysis. First, it is based on a criminal history review, and second, 8 of the 12 items bear on behavioral characteristics of offenders that can be identified in such a review. Given these two features of the DVSI, previous involvement in criminal and domestic violence offending should be significantly associated with DVSI scoring prior to court disposition. To test this assertion, a multivariate equation was estimated

Table 2. Prior Offending Predicting Total DVSI Scores for 1,465 Male Perpetrators

Prior offending	*B*	*SE*	*β*
Prior DV arrests	0.682	.058	.285*
Prior restraining orders	2.033	.244	.187*
Prior DUI arrests	0.341	.081	.097*
Other prior arrests	0.138	.012	.274*

Note. $R^2 = .283$.
$*p < .05$.

that included previous violation of restraining orders, previous domestic violence convictions, previous DUI convictions, and other previous arrests. Assuming record checks were reliably administered at the time the risk assessment was completed, the estimated effects should significantly predict DVSI scores.

The logic here is that the researchers collected criminal history data from official records on each of the 1,465 men in the sample. The POs should have used such data when the DVSI risk assessments were administered. Hence, two independent evaluators (the POs and the researchers) used the same data sources, and thus extracted scores should be correlated. Although admittedly an unorthodox test of interrater reliability, the logic underlying the test is compatible with its conceptual meaning: "the extent to which scores obtained from two or more raters (scorers, judges, observers) are consistent (Goodwin's, 2001: p. 15).

Ordinary Least Squares (OLS) regression was used to estimate the multivariate equation, with the results presented in Table 2. Observe that the prior offending variables significantly predict DVSI scores, although the R^2 is not particularly high (.28). That should be expected because this analysis does not include all the items in the DVSI. Regardless, the beta coefficients indicate that prior domestic violence offending has the strongest estimated effect, followed by total prior criminal offending, prior restraining orders, and prior DUI arrests. These statistically significant estimated effects, particular for domestic violence violations, suggest that the DVSI was administered reliably in the criminal history record checks.

Concurrent and Discriminant Validity of the DVSI

Determining the concurrent validity of the DVSI requires an external criterion measure that has acceptable psychometric properties; that is, previous research has empirically established its own validity. As described above, the SARA is essentially the only risk assessment instrument for repeat offending that has been validated (Kropp & Hart, 2000); therefore, it is used in the present analysis. Concurrent validation involves comparing SARA scoring of male perpetrators with DVSI scoring of the same perpetrators to determine the level of agreement between the two measures. The greater the agreement in assessing risk of repeat domestic violence offending using these two measures, the greater is the concurrent validity of the DVSI (see, Carmines & Zeller, 1979; Neuman, 1994). Total SARA scores were used in the analysis, with the actual range being 0–33, with a mean of 9.4 and a standard deviation of 6.2. Twenty-five percent of the men in the SARA subsample scored 13 or higher

on this instrument. The correlation between the DVSI and the SARA is $r = .539$, suggesting relatively strong evidence of concurrent validity. Further evidence of concurrent validity is found by comparing the DVSI to the perceived imminent risk of violence to the intimate partner. Recall that the evaluator performs this subjective risk rating once the SARA has been completed. The correlation between this measure and the DVSI total score is similar to that between the SARA total score and the DVSI, $r = .567$.

Discriminant validity also requires an external criterion measure that has been validated, but in this instance, one does not expect agreement between the two measures (see, Neuman, 1994). The lack of agreement can be manifested in either no empirical relationship or perhaps an inverse relationship, assuming the two instruments are measuring opposite psychological dispositions, states, or traits of individual perpetrators. To assess the discriminant validity of the DVSI, the total score for this measure was correlated with the perceived imminent risk of violence to other persons, which is the second subjective risk rating on the SARA. The DVSI was originally developed as an instrument to assess the risk of repeat partner violence. Hence, the assumption here is that intimate partner violence is different from violence against others, even if they are related to or known by the perpetrator; therefore, the two measures should not be as highly correlated as the association between the DVSI and the imminent risk of partner violence or the total SARA score.

This assumption is supported by the analysis. The correlation between the DVSI and the imminent risk of violence to other persons is relatively weak ($r = .152$). This evidence of discriminant validity was corroborated further by an additional analysis using the Level of Supervision Inventory (LSI), an instrument developed to assess the risk criminal offending in general, not partner violence or other forms of domestic violence (Bonta & Andrews, 1993). The correlation between the LSI and the DVSI is also relatively weak ($r = .169$) and much lower that the association between the SARA measures and the DVSI. The LSI was administered only to a relatively small subsample of the 1,465 men included in this analysis ($N = 165$), but no evidence was found of any systematic difference between this subsample and the total sample in terms of SARA and DVSI risk scores or the prevalence of any type of offending prior to or after court disposition. Combined with the findings concerning the imminent risk of violence to other persons, these findings support the discriminant validity of the DVSI.

Predictive Validity of the DVSI

Reviewing official records related to arrests and restraining orders allowed a comparison of the DVSI scores with behaviors recorded in those records during the 18-month follow-up period. Three types of data were collected: arrests for violations of domestic violence restraining orders, arrests for domestic violence reoffending, and arrests for other types of criminal offending. Restraining order violations and arrests for partner violence were combined to form a single-partner violence reoffending outcome measure, with 29% of the men in the sample engaging in such reoffending. A total reoffending measure was also used, including subsequent arrests for any type of offending during the 18-month follow-up period. The total reoffending prevalence in the sample during this period was 53%.

Because violence toward partners and others does not always result in an arrest, questions can be raised about the adequacy of official records for information on reoffending. However, they are likely to capture the more serious incidents of domestic violence (or other crimes), and they are the most cost-effective source of information on reoffending. Nonetheless, the potential inadequacy of coverage by official records on recidivism provided the rationale for seeking follow-up reports from the victims of domestic violence perpetrated by the sample of men analyzed here.

The immediate empirical question for this analysis is whether the DVSI is significantly associated with these patterns of reoffending. Finding such associations corroborates the predictive validity of the DVSI (see, Carmines & Zeller, 1979; Neuman, 1994). Like most distributions of serious crime and violence, all the behavioral outcome measures are positively skewed, with a high concentration of cases at 0, and the remainder of the distribution dropping quickly as the frequency of reoffending increases. Given such distributions, prevalence rates were calculated by dichotomizing these measures as 0 for no reoffending and 1 for any reoffending. Hence, the results reported below determine how well the total DVSI score predicts the presence versus the absence of reoffending, not the frequency of such behavior. It should be mentioned, however, that similar results were found using the frequency measures, but the results using the prevalence rates are reported below to ensure that they are not impacted by extreme outlying or other influential cases in the continuous distributions of reoffending.

Official Records During the 18-Month Follow-Up

The predictive accuracy of the DVSI was determined by conducting a Receiver Operating Characteristic (ROC) analysis. Swets et al. (2000) have provided a thorough discussion of the history of ROC analysis and a description of the "statistical machinery" upon which it is based. This summary draws from their work. ROC analysis has become the preferred analytical technique for prediction studies, especially those involving the prediction of violence (Monahan et al., 2001; Rice & Harris, 1995). It compares predicted outcomes and actual outcomes for various decision thresholds or cutpoints on a prediction scale, in this case, the DVSI. "True positives" or sensitivity and "true negatives" or specificity represent agreement or correct classification between predicted and actual outcomes. "False positives" (violence was predicted but did not occur) and "false negatives" (violence was not predicted but did occur) represent disagreement or misclassification. ROC curves are constructed by plotting true positives (sensitivity) against false positives (1-specificity) for decision thresholds varying from very stringent (no predictions of violence are rendered) to very lenient (all cases are predicted to be violent). The area under the curve (AUC) measures the accuracy of prediction. AUC will equal .50 when the probability of true positives is virtually the same as false positives across all decision-making thresholds; and thus, the risk assessment instrument does no better than chance in predicting behavioral outcomes. AUC will equal 1.0 when predicted and actual outcomes are in complete agreement, meaning the accuracy of prediction is perfect. The analysis presented below reports the AUC coefficient for each behavioral outcome measure and whether that coefficient is statistically significant from .50 (chance

or the "line of no information"). Bivariate Person's Product Moment Correlations between the total DVSI or SARA scores and the continuous behavioral outcome measures, despite their extremely skewed distributions, are also presented in parentheses after the AUC coefficients as additional descriptive information. The analysis reported below was conducted using the AccuROC for Windows software (Vida, 2001).

The initial analysis estimated the accuracy with which the total DVSI score predicts the dichotomized outcome measures for domestic violence and total reoffending. For domestic violence reoffending, AUC = .61 ($r = .18$, $p = .00$) and is significantly different from .50 ($p = .000$, two-tailed test), suggesting that the DVSI provides a significant improvement over chance in predicting domestic violence reoffending during the 18-month follow-up period. For total reoffending, the area under the curve is slightly higher, AUC = .65 ($r = .21$, $p < .00$), and also significantly different from .50 ($p = .000$, two-tailed test). In short, although prediction is by no means perfect, the relations between the DVSI and the 18-month follow-up behavioral measures suggest at least a moderate degree of predictive accuracy and certainly an improvement over chance in the use of this risk assessment instrument.

Victim Reports During the 6-Month Follow-Up

As described above, 125 victims were interviewed in depth over the phone 6 months after the offender's sentencing. Although limited to a small subsample, the victim survey provides information on offender behavior during the probation period from someone well acquainted with the offender. The victim interviews capture information on a detailed battery of items covering a variety of behavioral aspects occurring in the relationship between the offender and victim, including controlling behaviors, threats, and physical violence. Such information is not typically available in official arrest records or probation files, and thus victim reports complement the results reported above, based on those records. Indices were constructed for three forms of abusive behaviors. The index of control ($\alpha = .85$) includes the following items: frightening the victim by following her; keeping her from sleeping, using her income, working, quitting a job, talking on the phone, or spending time with friends and family; taking money from the victim; stopping her from going someplace she wanted to go; or pressuring her into sex. Threatening behaviors were subdivided into an index of threats ($\alpha = .79$) and an index of severe threats ($\alpha = .63$). The former includes swearing, screaming, or insulting the victim; accusations of infidelity; humiliation; throwing, smashing, hitting, or kicking something; destroying property; and hurting pets. The index of severe threats includes threatening to hit, attack, or harm the victim; threatening to use force to gain sexual access; threatening to kill; threatening to take away children or harm them; and threatening to kill, attack, or harm pets. Similarly, physically violent behaviors were subdivided into an index of violence ($\alpha = .84$), including pushing, grabbing, shoving; slapping; kicking, biting, hitting; pulling hair; burning or scalding; squeezing the neck; hitting with an object; threatening with a gun or knife; and using a gun or knife. Very severe violence ($\alpha = .77$) includes choking or trying to drown, using physical force to engage in a sexual act against the victim's will, and trying to kill the victim.

The results of the victim survey convey the continuing potential for domestic violence after sentencing. The reports by victims on the reoffending of their partners show that 35% of them used some type of physical force during the 6-month follow-up period, as measured by a modified version of the physical aggression items on the Conflict Tactics Scale listed above (Straus, 1979, 1990). The degree and nature of the physical violence varied among the offenders, but over one-third again used some sort of violence against their previous victim.

Other forms of aggression or maltreatment, although not physically violent, were more common. According to victims, 80% of offenders engaged in threatening and/or verbally abusive behaviors, 65% engaged in controlling behaviors, and almost 60% of the victims expressed safety concerns for themselves and/or their children. These percentages, including the reported violent reoffending, exceed the percent arrested for violations of domestic violence restraining orders and domestic violence reoffending, as well as total recidivism reported in official records.

ROC analysis again was used to determine the predictive accuracy of the DVSI concerning the perpetration of these controlling, threatening, or physically violent behaviors by the arrested male offenders during the 6-month follow-up period. Once again, correlations are also reported. An interesting pattern emerged in estimating the relations between the DVSI and the indices constructed. The predictive accuracy for the DVSI concerning the index of controlling behaviors was not significantly greater than .50 (AUC = .58, $p = .14$, two-tailed test; and $r = .13$), and that also was the case for the predictive accuracy of the DVSI regarding the index of less threatening behaviors (AUC = .56, $p = .26$, two-tailed test; and $r = .09$) as well as the index of less serious physically violent behaviors (AUC = .49, $p = .92$, two-tailed test; and $r = .09$). However, accuracy of prediction was substantial and significantly greater than .50 for the indices of severe threatening behaviors (AUC = .68, $p = .001$, two-tailed test; and $r = .22$, $p < .05$) and very severe physical violence (AUC = .65, $p = .041$, two-tailed test; and $r = .18$, $p < .05$). These findings suggest that the DVSI has appreciable predictive accuracy for the more serious forms of threatening and physically violent behaviors, but not the less serious forms of these behaviors or other controlling behaviors reported by the women victims of the men in this sample.

SUMMARY AND ADDITIONAL ANALYSIS

The results of the analysis based on approximating an assessment of interrater reliability discussed above suggest that the DVSI was administered reliably prior to court disposition for the sample of 1,465 male domestic violence offenders. The findings also provide significant evidence of both the concurrent and discriminant validity of this risk assessment instrument, using the SARA and the LSI as external criterion measures. Finally, the predictive validity of the DVSI is empirically supported by findings from the analysis of two samples, different methodologies, and two follow-up periods: The 1,465 sample of male offenders, using official record checks 18-months after sentencing, and the 125 subsample of women victims interviewed 6 months after sentencing. The results bearing on the predictive validity of the DVSI are particularly relevant because no risk assessment instrument has been previously evaluated using a prospective design, allowing such a validation. Furthermore,

the findings here suggest that the DVSI not only predicts subsequent domestic violence reoffending, but it predicts more severe forms of such repeated violence, as indicated by the severity of the behavioral items in the victim interviews. Despite the relevance of these empirical results, an additional analysis was conducted having relevance for the issue mentioned at the outset, namely, the distinction between clinical and actuarial risk assessment.

Recall that the DVSI was explicitly designed as a strictly statistically based instrument that could be administered quickly and easily to inform court dispositions. The conditions of its use, as described above, dictated its design. The demands of the setting are certainly important considerations in adopting any instrument for assessing the risk of domestic violence reoffending in the processing of cases. However, these demands should be balanced against other considerations that maximize the predictive validity of the risk assessment instrument adopted.

Intuitively, one would assume that assessments informed by multiple sources of information, including offender and victim interviews, criminal history reviews through official records, and the clinical judgment of evaluators, would have greater predictive validity than those based on a single source, especially if it is restricted to official records. One would further assume that risk assessments informed by a greater number of factors associated with domestic violence offending would enhance its predictive validity. The SARA is an instrument having these two characteristics: multiple sources of information, items requiring clinical expertise for completion, subjective clinical assessment of imminent risk of future violence, and more numerous factors linked to domestic violence offending, at least compared the DVSI. Hence, empirical evidence bearing on these assumptions can be garnered from an analysis that compares the predictive outcomes of the SARA with those of the DVSI.

Such an analysis was conducted with the sample of 434 male offenders for which SARA assessments were available; that is, predictions of reoffending during the 18-month follow-up period using data drawn from official record checks were estimated with this sample for both the SARA and the DVSI. As noted above, the analysis of sample selection bias yielded no significant evidence that this subsample differed significantly from the total sample in risk scores or offending patterns pre- and postadjudication. ROC analysis again was conducted, but in this case, AUC coefficients are calculated for each of the risk measures used in the analysis, and a chi-square test is performed to determine whether those coefficients are statistically significant from each other (Vida, 2001).

Three risk measures are used in the analysis, two of which are the DVSI and SARA total scores. The third risk measure is a weighted version of the SARA. Specifically, a cross-product is computed between the total SARA score and the perceived imminent risk of violence toward the partner, thereby incorporating this summary clinical assessment into the overall SARA risk score. The imminent risk rating cannot by analyzed separately because AccuROC requires a minimum of five cutpoints (Vida, 2001), and the imminent risk rating is only a three-category variable (low, medium, and high risk). Table 3 presents the results of the ROC analysis. Observe that the AUC coefficients for domestic violence reoffending during the 18-month follow-up period range from .60 to .65 ($r = .21 - .32$, $p < .00$), with the range for total reoffending being from .68 to .71 ($r = .31 - .36$, $p < .00$), and all coefficients are

Table 3. The Predictive Accuracy (AUC) of Three Risk Assessment Measures

Behavioral measures	DVSI	SARA	Weighted SARA
DV reoffending	.60*	.65*	.65*
Total reoffending	.68*	.70*	.71*

Note. DV reoffending: $\chi^2 = 3.02$, $df = 2$, $p = .22$; Total reoffending: $\chi^2 = 3.47$, $df = 2$, $p = .18$.
* $p < .000$.

significantly greater than .50, reflecting a significant improvement over chance in the prediction of these forms of reoffending. However, notice that chi-square tests for both behavioral outcome measures suggest that the predictive accuracy of all three risk assessments are virtually the same, meaning their respective AUC coefficients are not statistically significant from each other.

No definitive conclusions can be drawn from the results of this analysis concerning the relative importance of statistical prediction versus clinical judgment. However, it should be noted that the total SARA score, which is a sum of the scores across the 20 individual items and thus is more "statistical" in nature, performs as well as the weighted total SARA score, which incorporates clinical judgment by weighting the total numeric SARA score by the evaluator's professional assessment of the imminent risk of repeated violence to the perpetrator's partner. Moreover, the results support the assertion that at least in terms of predictive validity, the DVSI, a purely statistical tool, performs as well as the SARA or the weighted SARA. These findings support the assumptions stated above and underscore the importance of conducting further research into the relative merits of statistical prediction and clinical judgment in assessing the risk of reoffending. Whatever the ultimate blending (or not) of these two approaches, a balance must be achieved between the demands of the risk assessment setting and the predictive accuracy of the instrument used for assessments of risk.

CONCLUSION

Given the historical movement to criminalize domestic violence in the United States, the findings of this validation study are timely and surely promising for those contemplating risk-based decision making in the criminal justice processing of domestic violence cases. The findings also support the contention that the validity and utility of risk assessments will be enhanced by adequate coverage of the most appropriate factors on the risk assessment instrument, perhaps including those requiring clinical expertise, and comprehensive sources of information for identifying the presence and intensity of those factors. Choices of the most appropriate risk assessment instrument will be challenging in weighing the relative benefits of predictive validity versus the exigencies of assessment settings. Nonetheless, the challenge is moot without continued research on the reliability and validity of risk assessment instruments and their suitability for those charged with the responsibility of supervising and treating domestic violence offenders, protecting victims, and maintaining accountability to the general public.

APPENDIX: DOMESTIC VIOLENCE SCREENING INSTRUMENT

Defendant Last		Defendant First		DOB (M/D/Y)	/ /	Case No.		Judicial District	
Offense Date	/ /	Current Offense		Screener Name		Date Completed	/ /		

1. Prior non-domestic violence convictions?	☐ Unknown	0 ☐ None	1 ☐ 2 or fewer	2 ☐ 3 or More		
2. Prior arrests for assault, harassment or menacing?	☐ Unknown	0 ☐ None	1 ☐ Once	2 ☐ 2 or More		
3. Prior domestic violence treatment?	☐ Unknown	0 ☐ None	1 ☐ Once	2 ☐ 2 or More		
4. Prior drug or alcohol treatment?	☐ Unknown	0 ☐ None	1 ☐ Once	2 ☐ 2 or More		
5. Any history of domestic violence related restraining orders?	☐ Unknown	0 ☐ None	1 ☐ Once		3 ☐ 2 or More	
6. Any history of violatio n(s) of domestic violence restraining orders?	☐ Unknown	0 ☐ None	1 ☐ In the Past	2 ☐ Current Offense	3 ☐ Past & Current	
7. Any evidence of object used a weapon in commission of a crime?	☐ Unknown	0 ☐ None	1 ☐ Prior Offense	2 ☐ Current Offense	3 ☐ Past & Current	
8. Were children present during the domestic violence incident?	☐ Unknown	0 ☐ None	1 ☐ Prior Offense	2 ☐ Current Offense	3 ☐ Past & Current	
9. Current employment status.	☐ Unknown	0 ☐ None		2 ☐ Unemployed		
10. Has victim separated from defendant within last 6 months?	☐ Unknown	0 ☐ None		2 ☐ Yes		
11. Did victim have a restraining order in place against defendant at time of offense?	☐ Unknown	0 ☐ None			3 ☐ Yes	
12. Was defendant under any form of community supervision at time of offense?	☐ Unknown	0 ☐ None			3 ☐ Yes	
			+	+	+	=

Sentencing Date	/ /	Final Disposition	☐ Acquittal	☐ Conviction	☐ Deferred	☐ Dismissal	☐ Diversion	Victim Last		Victim First	
Dispo. Reason									Recommend SARA	☐ Yes	☐ No

ACKNOWLEDGMENTS

This research was supported by a grant to the Colorado State Court Administrators Office (SCAO), Division of Probation Services by the Department of Justice, Violence Against Women Grants Office, Office of Justice Programs (Award # 97-WE-VX-0009). We express our deep appreciation to Vern Fogg and Eric Philp in Probation Services for their support in this study and to Fred Pampel, Jay Watterworth, Katie Irwin, Linda Ramos, and Janet Mickish for their assistance in data collection and analysis at various stages of this research. We also thank the numerous people in the judicial department and the domestic violence community in Colorado whose dedication, tenacity, and compassion for those involved in domestic violence made the Domestic Violence Risk Reduction Project a successful effort in the state of Colorado.

REFERENCES

Binder, A., & Meeker, J. (1992a). The development of social attitudes toward spousal abuse. In E. S. Buzawa & C. G. Buzawa (Eds.), *Domestic violence: The changing criminal justice response* (pp. 3–19). Westport, CT: Auburn House.

Binder, A., & Meeker, J. (1992b). Arrest as a method to control spouse abuse. In E. S. Buzawa & C. G. Buzawa (Eds.), *Domestic violence: The changing criminal justice response* (pp. 129–140) Westport, CT: Auburn House.

Bonta, J., & Andrews, D. (1993). The Level of Supervision Inventory: An overview. *International Association of Residential Community Corrections Journal*, *5,* 6–26.

Campbell, J. C. (1995). *Assessing dangerousness: Violence by sexual offenders, batterers, and child abusers.* Thousand Oaks, CA: Sage.

Campbell, J. C. (2000). *Intimate partner violence risk assessment instruments: A prospective validation field experiment.* Washington, DC: National Institute of Justice.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Dutton, D. G., & Kropp, P. R. (2000). A review of domestic violence risk instruments. *Trauma, Violence, and Abuse*, *1,* 171–181.

Fagan, J. A. (1996). *The criminalization of domestic violence: Promises and limits.* Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Garner, J. H., Fagan, J. A., & Maxwell, C. D. (1995). Published findings from the Spouse Assault Replication Program: A critical review. *Journal of Quantitative Criminology*, *11,* 3–28.

Garner, J. H., & Maxwell, C. D. (2000). What are the lessons of the police arrest studies? *Journal of Aggression, Maltreatment and Trauma*, *4,* 83–114.

Gondolf, E. W. (2002). *Batterer intervention systems: Issues, outcomes, and recommendations.* Thousand Oaks, CA: Sage.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5,* 13–34.

Gosselin, D. K. (2000). *Heavy hands: An introduction to the crimes of domestic violence.* Upper Saddle River, NJ: Prentice Hall.

Grisso, T., & Appelbaum, P. S. (1992). Is it unethical to offer predictions of future violence? *Law and Human Behavior*, *16,* 621–633.

Grisso, T., & Appelbaum, P. S. (1993). Structuring the debate about ethical predictions of future violence. *Law and Human Behavior*, *17,* 482–485.

Harris, G., Rice, M., & Quinsey, V. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior*, *20,* 315–335.

Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) Guide: Reliability and validity in adult male offenders. *Law and Human Behavior*, *24,* 101–118.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Litwack, T. R. (1993). On the ethics of dangerousness assessments. *Law and Human Behavior*, *17,* 479–482.

Litwack, T. R., & Schlesinger, L. B. (1999). Dangerousness risk assessments: Research, legal, and clinical considerations. In A. Hess & I. Weiner (Eds.), *The handbook of forensic psychology* (pp. 171–217). New York: Wiley.

Maxwell, C. D., Garner, J. H., & Fagan, J. A. (2001). *The effects of arrest on intimate partner violence: New evidence from the Spouse Assault Replication Program.* Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Mignon, S. I., Larson, C. J., & Holmes, W. M. (2002). *Family abuse: Consequences, theories, and responses.* Boston: Allyn and Bacon.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques.* Beverly Hills, CA: Sage.

Monahan, J. (1996). Violence prediction: The past twenty and the next twenty years. *Criminal Justice and Behavior*, *23,* 107–119.

Monahan, J. (1997). Actuarial support for the clinical assessment of violence risk. *International Review of Psychiatry*, *9,* 167–169.

Monahan, J., & Steadman, H. J. (Eds.). (1994). *Violence and mental disorder.* Chicago: University of Chicago Press.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Clark Robbins, P., Mulvey, E. P., et al. (2001). *Rethinking risk assessment.* Oxford: Oxford University Press.

National Institute of Justice & American Bar Foundation. (1998). *Legal interventions in family violence: Research findings and policy implications.*Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Neuman, W. L. (1994). *Social research methods.* Boston: Allyn and Bacon.

Pleck, E. (1987). *Domestic tyranny: The making of social policy.* New York: Oxford University Press.

Quinsey, V., Harris, G. T., Rice, M. E., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk.* Washington, DC: American Psychological Association.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63,* 737–748.

Roehl, J., & Guertin, K. (1998). *Current use of dangerousness assessments in sentencing domestic violence offenders.* Pacific Grove, CA: State Justice Institute.

Saunders, D. G., & Hamill, R. (2002). *Offender interventions to end violence against women: A research synthesis for practitioners.* Paper presented at the Committee on Law and Justice, National Research Council, National Academy of Sciences workshop on expanding the federal research agenda on violence against women, Washington, DC.

Schecter, S. (1982). *Women and male violence: The visions and struggles of the battered women's movement.* Boston: South End Press.

Sherman, L. W. (1992). The influence or criminology on criminal law: Evaluating arrests for misdemeanor domestic violence. *The Journal of Criminal Law and Criminology*, *83,* 1–45.

Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review*, *49,* 261–272.

Sherman, L. W., & Cohn, E. G. (1989). The impact of research on legal policy: The Minneapolis Domestic Violence Experiment. *Law and Society Review*, *23,* 117–144.

Straus, M. A. (1979). Measuring intrafamily conflict and violence: The Conflict Tactics Scales. *Journal of Marriage and the Family*, *36,* 13–29.

Straus, M. A. (1990). The Conflict Tactics Scale and its critics: An evaluation of new data on validity and reliability. In M. A. Straus & R. J. Gelles (Eds.), *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families* (pp. 49–73). New Brunswick, NJ: Transaction Publishing.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1,* 1–26.

Tjaden, P., & Thoennes, N. (1998, November). *Prevalence, incidence, and consequences of violence against women: Findings from the National Violence Against Women Survey* (Research brief). Washington, DC: National Institute of Justice and Centers for Disease Control and Prevention.

Vida, S. (2001). AccuROC for Windows 95/98/NT (Version 2.4). Montreal: Accumetric Corporation, Department of Psychiatry, McGill University Health Centre, Montreal General Hospital.

Webster, C. D. (1990). Prediction of dangerousness polemic. *Canadian Journal of Criminology*, *32,* 191–196.

Weisz, A. N., Tolman, R. M., & Saunders, D. G. (2000). Assessing the risk of severe domestic violence: The importance of survivors' predictions. *Journal of Interpersonal Violence*, *15,* 75–90.

Williams, K. R., & Conniff, E. (2002). *Legal sanctions and the violent victimization of women.* Paper presented at the Committee on Law and Justice, National Research Council, National Academy of Sciences workshop on expanding the federal research agenda on violence against women, Washington, DC.