Document Title:	Intimate Partner Violence Risk Assessment Validation Study: The RAVE Study Practitioner Summary and Recommendations: Validation of Tools for Assessing Risk from Violent Intimate Partners
Author(s):	Janice Roehl, Ph.D.; Chris O’Sullivan, Ph.D.; Daniel Webster, ScD; Jacquelyn Campbell, Ph.D.
Document No.:	209732
Date Received:	May 2005 (Revised December 2005)
Award Number:	2000-WT-VX-0011

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Intimate Partner Violence Risk Assessment Validation Study: The RAVE Study

Practitioner Summary and Recommendations:
Validation of Tools for Assessing Risk From Violent Intimate Partners

Janice Roehl, Ph.D.; Chris O'Sullivan, Ph.D.; Daniel Webster, Sc.D.; and Jacquelyn Campbell, Ph.D.

Background and Purpose of the Risk Assessment Study

There is an increasing demand for accurate risk assessment in the field of domestic violence. This demand is the result of a dramatic transformation over the past two decades in the response to intimate partner violence across all sectors, including the criminal justice system, social and advocacy services, health care, and public opinion. Increased use of criminal justice remedies has necessitated a sort of triage in case processing by law enforcement, prosecutors, and courts. Hotlines, emergency shelters, and advocacy and counseling programs are now available in almost every community. Emergency medical and prenatal settings are increasingly assessing for domestic violence and offering assistance to victims. As public awareness of domestic violence as a crime rather than a private family matter has grown, it has altered the landscape and increased the demands on all systems.

To respond to this increased demand for services, agencies dealing with victims and offenders have adopted a number of mechanisms to identify high-risk cases in order to direct scarce resources and intensive services to those most in need. There is also the need for abused victims to be aware of the level of danger the abuser presents to them. The central purpose of this study was to assess the accuracy of several different approaches to predicting risk of future harm or lethality in domestic violence cases.

This project was conducted under an NIJ grant awarded to Johns Hopkins University, with subcontracts to Safe Horizon and the Justice Research Center. The primary staff were:

Johns Hopkins University
Jacquelyn C. Campbell, Ph.D., R.N., Principal Investigator
Daniel Webster, Sc.D., Co-Investigator
Patricia Mahoney, M.A., Project Director

Safe Horizon, New York
Chris O'Sullivan, Ph.D., Co-Principal Investigator
Melissa White, M.S.W., Project Director
Johanne Eliacin, M.A., Project Director

Justice Research Center, California
Janice Roehl, Ph.D., Co-Principal Investigator
Kristen Guertin, M.A., Project Director
Kate Semple, Project Director

Four methods were tested: Danger Assessment (DA), DV-MOSAIC*, Domestic Violence Screening Instrument (DVSI), and Kingston Screening Instrument for Domestic Violence (K-SID). These four methods vary greatly in length and complexity and were designed for different purposes and settings. They were selected for the risk assessment study because agencies and service providers around the country currently use them, yet little is known about whether and how well they accurately assess the likelihood of future violence. Table 1 presents basic information about the length, content, administration, and primary intended use of each of the risk assessment methods tested.

In addition to the four risk assessment methods, we tested the predictive accuracy of (1) the victim's own assessment of the likelihood that her partner or ex-partner would physically abuse or seriously harm her over the course of the next year and (2) other risk factors drawn from the literature and other assessment tools (results not presented here).

* DV-MOSAIC is not actually designed to be used as a questionnaire; moreover, it is not intended to serve as a predictive instrument. It is, however, an approach to investigation of domestic violence cases for immediate threat assessment that is of great interest to law enforcement. For this study, we derived a questionnaire based on the factors and areas of inquiry in DV-MOSAIC, with the cooperation of Gavin de Becker & Associates.

Methodology

To test the risk assessment methods, we interviewed domestic violence victims two times, with a baseline interview including risk assessment and a followup interview 6 months to a year later. We also gathered arrest information on the offender for at least a year after the baseline interview. We compared the scores on the risk methods at the baseline interview to the following outcomes: physical assault during the followup period, severe assault, stalking and threats, and arrests. To measure the frequency and severity of reassault, we asked the victims questions from the revised Conflict Tactics Scale (CTS2) to measure physical, sexual, and psychological violence; questions from the WEB Scale to measure emotional abuse and controlling behaviors; and questions from HARASS to measure stalking and harassment.

Table 1: Description of four risk assessment methods

Method	Description	Administration	Primary intended uses
DA (Campbell, 1986, 1995; Campbell et al., 2003)	Review of past year with a calendar[1] to document severity and frequency of battering and 20 yes/no questions about risk factors. Scoring: -3–37 and four risk categories (variable, increased, severe, and extreme danger).	Interview with the victim, usually by victim advocate.	Assess risk of extreme dangerousness and lethal violence for victim education, awareness, safety planning, and service provision.
DV-MOSAIC (De Becker & Associates, 2000)	Computer-assisted method that includes 46 multiple response items about risk and protective factors. Scoring: Program computes risk scores of 1–10 and a missing data (IQ) score.	Criminal justice professional enters responses after victim interviews, perhaps after offender and other interviews; reviews of criminal records and police reports.[2]	Assess immediate, short-term threat of severe or lethal domestic violence situations for victim awareness, safety planning, further investigation, and criminal justice responses.
DVSI (Williams and Houghton, 2004)	Twelve questions given 0–3 points, primarily related to offender's criminal history, employment, and several other risk factors. Scoring: Risk scores of 0–30, and two risk categories (not high risk and high risk).	Probation or other court officer completes instrument based on offender's criminal record and interview.	Assess risk of recidivism/reassault for supervision, probation/parole, and other offender-related decisions.
K-SID (Gelles, 1998)	Ten questions about risk factors, each with two or three response categories, and an offender's poverty status scale. Scoring: Risk scores of 0–10 and four risk categories (low, moderate, high, or very high).	Offender and victim interviews and review of police reports by probation or other court officer.	Assess risk of recidivism/reassault for offender charging and supervision decisions; set conditions for release, probation, and protective orders.
Victim's perception of risk (Goodman et al., 2000; Heckert and Gondolf, 2004; Weisz et al., 2000)	Two questions about victim's perception of the likelihood that she will be physically assaulted or seriously hurt by abuser in the next year. Scoring: Victim rates likelihood on a scale of 1–10.		

[1] The calendar portion of DA was not used in this study. The CTS2 questions obtained severity of abusive tactic information.

[2] The DV-MOSAIC "domains of inquiry" were reformatted by the investigators as a victim interview.

Baseline interviews were conducted by highly trained interviewers with 1,307 battered women recruited from five different populations and settings: Women seeking protection orders against their male partners in New York City (NYC) Family Courts (n = 628), female victims in 911 domestic violence calls to the Los Angeles Sheriff's Department (n = 400), women in shelters in NYC (n = 177) and Los Angeles (n = 58), women seeking emergency care from NYC hospitals (n = 28), and domestic violence clients of Safe Horizon's community programs (n = 11). Two-thirds of the women were interviewed in person for the baseline interview, and one-third were interviewed by phone.

Each participant was randomly administered two of the four risk assessment methods (DA or DV-MOSAIC, and K-SID or DVSI), questions related to her own perception of risk, and the additional risk factor questions; questions from CTS2, WEB, and HARASS; and questions about past injuries. Each woman also answered questions about her own and her partner's demographic characteristics, current and past relationship with the abuser, past protective actions, the offender's arrest and/or incarceration, victim services (e.g., safety planning, counseling, shelter, legal assistance), and going into hiding.

Between 6 and 12 months after the baseline interview, the women were recontacted and asked to participate in a followup telephone interview. Followup interviews were successfully completed with 782 women, 60 percent of the original sample. The followup interview focused on any abuse experienced between the baseline and followup period and any preventive actions taken or interventions occurring during that period. The criminal records of all 1,307 offenders were checked for any violent offenses since the baseline interview.

This study presented the researchers with a number of methodological and practical challenges, particularly recruitment and retention (see full final report, Roehl et al., 2005).

Participant Characteristics

The demographic characteristics of the participants are presented in table 2, broken down by those who completed both baseline and followup interviews (T2) and those who completed the baseline interview only (T1). There were a few significant differences in demographics between those who participated in followup interviews and those who did not (primarily because they could not be reached by phone): Women who were employed, women who identified themselves as Latinas or Hispanics, and women who identified themselves as homemakers were significantly more likely to be reached at followup. Overall, the final sample of 782 women was primarily non-white, with 38 percent foreign born. About a third had some college education or a college degree, almost half were employed, and most (69 percent) were no longer involved or living with the offender.

Table 2: Characteristics of participants

Characteristic	Participants with baseline and followup interviews (n = 782)	Participants with only baseline interviews (n = 525)
Racial/Ethnic group**		
African descent/Black	209 (27%)	154 (29%)
Latina/Hispanic	444 (57%)	250 (48%)
European descent/White	72 (9%)	58 (11%)
Other racial/ethnic groups	55 (7%)	62 (12%)
Foreign born	295 (38%)	202 (39%)
Education		
Less than high school	262 (34%)	183 (35%)
High school diploma /GED	260 (33%)	152 (29%)
Some college or voc. school	186 (24%)	148 (28%)
B.A./B.S. or college degree	74 (10%)	41 (8%)
Employment status/situation**		
Employed full time	251 (32%)	143 (27%)
Employed part time	117 (15%)	63 (12%)
Homemaker**	155 (20%)	72 (14%)
Looking for work	98 (13%)	67 (13%)
Unemployed***	252 (32%)	227 (43%)
Student	83 (11%)	58 (11%)
Offender's relationship to victim		
Spouse/Common law spouse	230 (29%)	138 (27%)
Ex-spouse/Ex-common law spouse	47 (6%)	43 (8%)
Estranged spouse	117 (15%)	72 (14%)
Boyfriend	77 (9%)	31 (6%)
Ex-boyfriend	311 (40%)	230 (45%)
Cohabitation at baseline		
Cohabitating	180 (23%)	88 (17%)
Involved but not cohabitating	32 (4%)	17 (3%)
On again, off again	29 (4%)	17 (3%)
Not involved or cohabitating	541 (69%)	403 (77%)

**p < .01, ***p < .001

The only statistically significant difference between those retained in the study at T2 and those who could not be recontacted was that those who could not be recontacted more frequently experienced severe physical abuse at T1 (table 3).

Abuse Inflicted and Protective Actions Taken During Followup Period

A third of the 782 women who participated in the followup interview reported being physically assaulted by their partners or ex-partners between the baseline and followup interviews, a proportion similar to that in other studies (Hilton et al., 2004; Williams and Houghton, 2004).

Table 3: Average risk assessment scores, self-perceived risk, and frequency of severe physical abuse score at baseline

Baseline risk assessment or abuse measure (range of actual scores)	Participants with baseline & followup interviews (n = 782)	Participants with only baseline interviews (n = 525)
DA point score (range: -1–37)	15.02	15.91
DV-MOSAIC rating (range: 3–9)	6.85	6.97
DVSI point score (range: 0–28)	8.60	8.65
K-SID risk score (range: 0–10)	1.09	1.12
Likelihood partner will physically abuse me in the next year(range:1–10)	5.01	5.36
Likelihood partner will seriously hurt me in the next year (range:1–10)	4.63	5.05
Frequency of severe physical abuse, from CTS2 (range: 0–42)*	6.82	8.41

*p < .05

As shown in table 4, the women who experienced reassault during the followup period were evenly divided among those who experienced a "low" level of physical abuse such as having their arm twisted, their hair pulled, or being pushed or shoved; those who experienced moderate to high physical abuse, such as being punched, kicked, choked, or beaten up; and those who experienced very high, potentially lethal abuse. The eight categories listed below will be used later in this report as the primary outcomes for assessing the predictive accuracy of risk assessment scores obtained at baseline.

Table 4: Abuse experienced during followup period

Form of abuse*	Representative items in category	n	(%)
None		125	(16%)
Verbal	Called names, insulted.	48	(6%)
Psychological/ Harassment	Controlling behavior.	240	(31%)
Stalking/Threats	Stalking/Threats to harm.	126	(16%)
Physical abuse: Low	Twisted arm/hair, grabbed, pushed/shoved, caused sprain, bruise, small cut.	80	(10%)
Physical abuse: Medium	Punched, kicked, caused physical pain that still hurt the next day.	26	(3%)
Physical abuse: High	Choked, burned, beat up. Serious injury inflicted (e.g., blacked out due to blow to head, broken bone).	49	(6%)
Physical abuse: Very high	Used gun/ knife, tried to kill. Life-threatening injury (e.g., lost consciousness due to choking).	88	(11%)

*Participants are categorized by the highest level of abuse they reported.

A third of the 782 women retained in the study were psychologically abused and/or harassed by the offenders during the followup period. Almost one in 11 (8.8 percent) reported that her abuser tried to kill her during the period of 6 months to a year after the baseline interview. When verbal abuse was included, only 16 percent were totally free of all forms of abuse by their intimate partner or former partner during the followup period.

The review of criminal records in New York and California for the 1,307 original offenders showed that arrests for criminal offenses committed during the followup period were infrequent. Only 6 percent of the offenders were arrested for a domestic violence crime, and 11 percent were arrested for another violent crime. Including the women who reported being stalked and/or threatened with harm, the total percentage of women either physically assaulted or stalked during the followup period was 46, considerably more than the 17 percent of the offenders arrested.

Table 5: Protective actions limiting contact with abuser during followup period, by risk method and level of risk at baseline (frequency [n] and percentage)

	Number in each risk level	Victim lived in hiding		Victim went to DV shelter		Victim left town		No contact, Voluntary		Abuser jailed	
Method and level of risk*	n	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
DA											
Low (Variable, 1–7)	68	5	(7.4)	0	(0)	2	(0)	9	(13.2)	7	(10.3)
Moderate (Increased, 8–13)	99	29	(29.3)	8	(8.1)	9	(9.1)	40	(40.4)	15	(15.2)
High (Severe, 14–17)	80	22	(27.5)	9	(11.3)	4	(5.0)	35	(43.8)	13	(16.2)
Very high (Extreme, 18–36)	153	62	(40.5)	23	(15.0)	14	(9.2)	80	(52.3)	44	28.8)
Total	400	p < .001		p = .006		p = .284		p < .001		p = .002	
DV-MOSAIC											
Low (3–4)	23	0	(0)	0	(0)	0	(0)	5	(21.7)	3	(13.0)
Moderate (5–7)	225	56	(24.9)	18	(8.0)	13	(5.8)	94	(42.0)	28	(12.4)
High (8–10)	134	60	(44.8)	26	(19.4)	13	(9.7)	71	(53.0)	36	(26.9)
Total	382	p < .001		p = .001		p = .148		p = .010		p = .001	
DVSI											
Not high risk (0–7)	176	45	(25.6)	12	(6.8)	11	(6.3)	55	(31.3)	18	(10.2)
High risk (8–28)	212	73	(34.4)	33	(15.6)	10	(4.7)	114	(53.8)	54	(25.5)
Total	388	p = .059		p = .007		p = .506		p < .001		p < .001	
K-SID											
Low (0–3)	153	42	(27.5)	15	(9.8)	10	(6.5)	52	(34.2)	14	(9.2)
Moderate (4–6)	136	39	(29.5)	11	(8.1)	10	(7.4)	59	(43.7)	27	(19.9)
High (7–8)	12	2	(16.7)	1	(8.3)	1	(8.3)	6	(50.0)	7	(58.3)
Very high (9–10)	90	34	(37.8)	12	(13.3)	8	(8.9)	48	(53.3)	25	(27.8)
Total	391	p = .241		p = .635		p = .925		p = .030		p < .001	
Victim's perception of risk											
Low (1–4)	368	87	(23.6)	35	(9.5)	18	(4.5)	129	(35.1)	74	(20.7)
Moderate (5)	87	30	(34.5)	9	(10.3)	10	(11.5)	40	(46.0)	12	(14.0)
High (6–10)	321	117	(36.4)	41	(12.8)	26	(8.1)	160	(49.8)	60	(19.4)
Total	776	p < .001		p = .385		p = .054		p < .001		p = .361	

*Cutoff scores for each risk level are those used by the developer, with the exception of DV-MOSAIC which is intended to be a continuous scale. We standardized the terms for the levels of risk across instruments. The terms "low, moderate," and so forth are ours and not necessarily those suggested by the developer as described in table 1.

We found that many women participating in the study took significant steps to protect themselves from further abuse after the baseline interview (see table 5). In general, women who scored at a higher risk at baseline were more likely to take protective actions during the followup period. We found that, except for K-SID, the high-risk categories were associated with increased efforts on the part of victims to escape their abusive ex-partners. For example, women who scored in the highest risk category (extreme danger) of DA were over five times more likely than women in the lowest risk category (variable danger) to go someplace where their abuser could not find them (41 percent in the extreme danger category versus 7 percent in the variable danger category). Victims at high risk based on their DV-MOSAIC score were twice as likely as women who scored at lower levels of risk to go someplace where their abusers could not find them (45 percent versus 23 percent). Abusers who scored in the highest risk categories of all the methods except K-SID and victim assessment of risk were two to three times more likely to be in jail at followup. It should be noted that the victims may have incorporated into their perception of risk of harm in the next year the protective actions and interventions that were already in place at baseline (e.g., very dangerous abuser incarcerated, victim in shelter).

Tests of Predictive Accuracy of Risk Assessment Methods

The four risk assessment methods and other risk factors were assessed using a combination of simple and advanced statistics. In this practitioner summary, statistical tests and their meanings will be explained in practical terms; the reader is urged to see the full report for additional information.

We determined that 27 women had no possible contact with their abusers during the followup period, either because he was incarcerated or out of the country or because she was in a shelter or otherwise out of his reach the entire time. The analyses of the predictive accuracy of the methods reported in this summary exclude those 27 cases.

Evaluating a risk assessment instrument is not as straightforward a task as one might think. We performed many types of analyses using different statistics and several different ways of categorizing outcome measures (see final report for full descriptions of all analyses). Outcome measures were the victims' reports of abuse and the offenders' arrests for domestic violence during the followup period. Most tests were conducted using two categories of reassault: "Any assault," which is any physical violence reported by the victim, and "severe assault," defined as high or very high physical abuse on the levels in table 4. We also used both the 8-point severity of abuse scale shown in table 4 and CTS2 scores to assess outcomes. Some analyses controlled for actions taken by the victim or the system that reduced the possibility of reassault, such as the victim going to a shelter or the perpetrator being incarcerated. Depending on the requirements of the test statistic, risk scores were entered either as continuous variables (e.g., running from 1 to 10 or 0 to 30) or categories (e.g., low, moderate, high, very high). There were two measures of victim's perception of future risk at baseline—one that asked about the likelihood that she would be physically abused by her partner/ex-partner in the coming year and another that asked about the likelihood that he would seriously hurt her in the next year. When the outcome of interest was any physical abuse during the followup period, the comparison was made with the victim's perceived risk of any assault, and when we analyzed outcomes that took into account the severity of reassault, we used the item about the victim's assessment of the likelihood he would seriously hurt her.

Correlations Between Risk Assessment Scores and Subsequent Abuse

A central requirement of a risk assessment instrument is that the risk scores should be related to subsequent violence—the higher the risk score, the more likely there will be future violence or (for those methods purporting to assess the likelihood of lethal violence) the more likely the violence will be severe. The correlations are the first of many tests we performed, and the results are presented in table 6. This test used the continuous scores on the risk assessment methods and correlated them with the outcomes. A correlation can range from 0 to 1.00, where 1.00 represents a perfect relationship between risk score and outcomes. After the 27 cases with no potential victim-partner contact during the followup period were excluded, DA had the strongest correlation with the severity scale outcome; DV-MOSAIC and the victim's risk assessment tied for the second highest; DVSI was next; and K-SID was the lowest. While all the correlations were statistically significant (i.e., the association between the score and the severity of subsequent abuse was stronger than would be found by chance), they are low. We also tested the correlation of the risk scores with the outcome of severe physical abuse as measured by the CTS2. Only

Table 6: Correlation between risk scores and subsequent severity of abuse

Risk method	8-point severity scale	Severe physical abuse (CTS2)
DA	.38*	.25*
DV-MOSAIC	.22*	.07
DVSI	.20*	.17*
K-SID	.13*	.02
Victim's perception of risk	.22***	.15***

*Significant at p < .05, ***significant at p < .001.

DA, victim’s assessment of risk, and DVSI scores were significantly correlated with CTS2 severe abuse scores, but again these correlations are low.

Table 7: Method and level of risk of any subsequent assault and severe assault during the followup period

Method and level of risk	No. in each risk level n	No. and % experiencing any assault n (%)	No. and % experiencing severe assault n (%)
DA			
Low (Variable, 1–7)	67	11 (16.4)	2 (3.0)
Moderate (Increased, 8–13)	98	28 (28.6)	16 (16.3)
High (Severe, 14–17)	79	30 (38.0)	19 (24.1)
Very high (Extreme, 18–36)	144	63 (43.8)	47 (29.9)
Total	388	p < .001	p <.001
DV-MOSAIC			
Low (3–4)	23	4 (17.4)	1 (4.3)
Moderate (5–7)	219	67 (30.6)	31 (14.2)
High (8–10)	126	40 (31.7)	27 (21.4)
Total	368	p = .317	p = .060
DVSI			
Not high risk (0–7)	179	59 (33.0)	25 (13.8)
High risk (8–28)	194	67 (34.5)	50 (25.3)
Total	373	p = .748	p < .01
K-SID			
Low (0–3)	151	40 (26.5)	21 (13.9)
Moderate (4–6)	131	40 (30.5)	25 (19.1)
High (7–8)	11	5 (45.5)	2 (18.2)
Very high (9–10)	90	32 (35.5)	16 (17.8)
Total	383	p = .336	p = .688
Victim’s perception of risk			
Low (1–4)	313	73 (23.3)	45 (12.7)
Moderate (5)	93	34 (36.6)	18 (20.7)
High (6–10)	343	134 (39.1)	75 (24.8)
Total	749	p < .001	p < .001

For the second test, we examined the associations between the levels of risk of each method (e.g., low risk, moderate risk) with the outcomes of “any abuse” and “severe abuse” (see table 7). Only the four levels of DA and the three levels of the victim’s assessment of risk were significantly associated with the outcome of any assault. For the outcome of severe assault, we found significant associations between outcomes and the four levels of risk on DA, the three levels of the victim’s assessment, and the two DVSI levels of risk. The DV-MOSAIC levels approached a statistically significant correlation (p = .06) with severe reassault. K-SID, with four risk levels, showed no significant association with reassault or with severe assault.

Four Quadrant Model

True positives, false positives, true negatives, and false negatives are terms that are very important in evaluating risk assessment methods. The levels on DVSI, which has just a two-level risk scale (not high risk and high risk), are the easiest way to show the meaning of these statistical terms, illustrated in figure 1 (Green and Swets, 1966).

Figure 1: Four quadrant risk model, with DVSI figures for illustration

	True	False
Positive (violence predicted)	True positives: Violence occurred in 67 of the 194 cases rated high risk (34.5%).	False positives: Violence did not occur in 127 of the 194 cases rated high risk (65.5%).
Negative (violence not predicted)	True negatives: Violence did not occur in 120 of the 179 cases rated low risk (67.0%).	False negatives: Violence occurred in 59 of the 170 cases rated low risk (33.0%).

True positives are the cases that are predicted to be high risk at baseline and experience violence during the followup period. In figure 1 (using data drawn from table 7), we see that 194 cases were assessed to be at high risk by DVSI, and violence occurred in 67 of them. The true positive rate is 34.5 percent (67/194). The other 127 cases assessed as high risk did not have any further violence—a false positive rate of 65.5 percent (127/194).

Table 8: PPV levels for any assault for different risk categories

Method and level of risk	Proportion reassaulted based on victim report	Proportion reassaulted based on victim report and CJ records
DA		
Moderate (Increased)	.377	.386
High (Severe)	.417	.426
Very high (Extreme)	.438	.444
DV-MOSAIC		
Moderate	.310	.342
High	.317	.325
DVSI		
High	.345	.366
K-SID		
Moderate	.332	.349
High	.366	.366
Very high	.360	.356
Victim's perception of risk		
Moderate	.385	.397
High risk	.391	.397

True and false negatives involve cases assessed *not* to be high risk. Looking at DVSI again, 59 of the 179 cases rated not high risk had subsequent violence, a false negative rate of 33 percent. DVSI accurately assessed that violence was unlikely in 67 percent of the cases (120 out of 179) in which violence did not occur during the followup period (true negatives).

Positive Predictive Value

When there are more than two categories of risk, positive predictive value (PPV) indicates the proportion of cases at or above a given risk or cutoff level that experience reassault during the followup period. For example, extrapolating from table 7, if the moderate (increased) risk category of DA was the cutoff point, with all cases scoring at the moderate level or above considered at risk, the PPV is 37.7 percent. (Reassaults occurred in 121 of the 321 cases judged at moderate or higher risk by DA).

Generally, PPV levels were fairly low, with DA at the very high (extreme) level having the highest PPV of the methods (.444) when arrest data are included along with victim reports (see table 8). Second best was the victim's own rating of risk. The rest of the PPVs are clustered in the .31–.40 range. (Note that 31 percent of all the victims were reassaulted during the followup period, a typical finding of studies of intimate partner violence; therefore a risk assessment needs to give us a better idea of which women are going to be reassaulted to be informative.) Including the arrest data in outcomes along with the victim's reports of abuse slightly increases the number of reassaults detected, and adding the criminal justice data to the outcomes increased PPV at least slightly for all methods.

Sensitivity and Specificity

Figure 2: Measures of predictive power, definitions and formulas

Measure of predictive power	Definition	Formula*
Sensitivity	Proportion of cases (women assaulted during followup) correctly identified as high risk	TP/ (TP + FN)
Specificity	Proportion of noncases (women not assaulted during followup) correctly identified as not high risk	TN/ (TN + FP)
Positive predictive value (PPV)	Proportion of those identified as high risk who become cases (reassaulted)	TP/ (TP + FP)

*TP = true positive, FN = false negative, TN = true negative, FP = false positive.

The best risk assessment method maximizes both true positives *and* true negatives. Sensitivity is a measure that refers to the proportion of women who experienced an assault during the followup period who were correctly predicted to be at high (or increased) risk (i.e., sensitivity is the number of true positives divided by the number of true positives plus false negatives). Specificity refers to the percentage of women who were not assaulted during the followup period who were correctly identified as not at high risk (i.e., specificity is the number of true negatives divided by the number of true negatives plus false positives). The “best” all-around method will have high sensitivity and high specificity; that is, it will correctly identify both the high risk *and* low risk cases. Figure 2 gives definitions and formulas for sensitivity, specificity, and PPV.

Sensitivity and specificity are affected by “where one draws the line,” or what the cutoff score is to designate a case high risk or not. If an instrument measures risk from 1 to 10, for example, putting the line at “2” will capture most cases with further violence. Thus, at a risk level of 2, it will have a high level of sensitivity, but it is also likely to have a low level of specificity—it will identify many cases as high risk that do not experience further violence. Practitioners may be more interested in one than the other. A domestic violence advocate, for example, may be primarily interested in sensitivity, so that the maximum number of victims who could be at risk may be forewarned. A judge, however, may be at least as concerned with specificity in making a ruling on bail, sentencing, or probation supervision, to minimize violation of offender’s rights. However, putting the cutoff point low to maximize sensitivity has consequences for service providers as well: As more clients are judged high risk, sensitivity will approach 100 percent and specificity will approach 0 percent, but the risk assessment will have little utility in directing resources and intensifying services for the highest risk clients.

Table 9 presents the sensitivity and specificity of the four methods, comparing baseline risk scores to any physical or sexual assault during the followup period. To illustrate again the points discussed above, setting the cutoff point of DA at the moderate risk (increased) level yields the best sensitivity score (.917), meaning that scores of moderate, high, or very high risk correctly identified almost 92 percent of the risky cases. Yet DA at that level also has a low specificity score (.219), meaning it correctly classified only 22 percent of the women not reassaulted. Plugging in the data from table 7 and using the formula in figure 2, one can see that 56 of 67 women were correctly identified as low risk (no reassault occurred) and 200 women were identified as at moderate risk or above but were not reassaulted (70 of the 98 women judged at moderate risk were not reassaulted, 49 of those judged at high risk were not reassaulted, and 81 of the 144 women judged at very high risk were not reassaulted). Thus, specificity is 56/(56+200) or .22.

Table 9: Sensitivity and specificity of method's risk categories and any subsequent physical or sexual assault

Method and level of risk	Cases at risk correctly identified (sensitivity)	Cases not at risk correctly identified (specificity)	Most accurate overall (sensitivity + specificity)
DA			
Moderate (Increased)	.917	.219	1.136
High (Severe)	.704	.492	1.196
Very high (Extreme)	.477	.684	1.161
DV-MOSAIC			
Moderate	.826	.074	.900
High	.360	.680	1.040
DVSI			
High	.532	.486	1.018
K-SID			
Moderate	.658	.417	1.075
High	.316	.759	1.075
Very high	.274	.782	1.056
Victim's perception of risk			
Moderate	.697	.472	1.169
High risk	.556	.589	1.145

As previously discussed, raising either sensitivity or specificity by changing the cutoff score of an instrument tends to lower the other measure, with the ideal instrument exhibiting high sensitivity *and* specificity. Combining the sensitivity and specificity scores (in the last column in table 9) shows that the high (severe) level of DA, of all the methods' risk categories, is the most accurate risk category, followed by the victim's own risk assessment and other levels of DA.

Specificity (accurately predicting that the abuser will *not* be violent) increases on all instruments as the levels of risk predicted increase. This pattern is expected and desirable. That is, men who were assessed to be unlikely to commit violence were least likely to commit any violence and especially severe violence. In predicting any assault, specificity is highest for K-SID at the very high risk level (.78) and at the high risk level (.76). As K-SID was primarily designed for use by probation officers, maximizing specificity at the expense of sensitivity may be deemed appropriate. Higher levels of specificity are achieved by categorizing more participants at low risk and fewer at high risk, while the reverse is true for achieving higher levels of sensitivity. The best balance between specificity and sensitivity is achieved by DA and victim's self-perception, although both methods incorrectly categorized 40–43 percent of the cases.

Sensitivity was somewhat higher for predicting *severe* violence outcomes at all risk levels of all the methods, with the exception of the highest two levels of K-SID and the victim's prediction. DA and DV-MOSAIC were designed to identify high risk, potentially lethal cases; therefore, it makes sense that they would more accurately predict severe violence. Except for two highest levels of K-SID, sensitivity + specificity scores for all other risk levels are higher for predicting severe assault than predicting any assault.

Receiver-Operator Characteristic (ROC) Curve Analysis

The ROC curve analysis is considered to be one of the most important means of examining the predictive accuracy of any approach to risk assessment. The area under the ROC curve is a statistic that summarizes the predictive accuracy (sensitivity and specificity) of a measure with each unit (or point) change in a score (such as going from a 9 to a 10). The ROC curve creates a series of successive cut-points for each unit and measures sensitivity and specificity at each. The area under the ROC curve is a summary statistic of all these sensitivities and specificities. If the risk scores do not enhance prediction of future abuse, the area under the ROC curve is not significantly greater than .50, the value where the instrument provides no predictive information.

We found that self-protective actions taken by the victims were independently associated with reassault and thereby introduced a confounding variable into assessing the accuracy of the predictions. In other words, the prediction of violence might have been accurate, but the intended victim or the system took steps that circumvented the abuser's ability to inflict harm. To take these protective actions that reduced risk into account, we included the following factors in the ROC analysis: (1) length of time the victim was potentially at risk of assault during followup, (2) whether the victim avoided contact with the abuser during the followup period by mutual choice, (3) whether the victim went to a shelter, (4) whether the victim received counseling, (5) whether the victim changed the locks on her doors, and (6) whether the abuser was incarcerated. Table 10 presents the summary ROC curve analyses with and without the criminal justice outcome data and with and without taking into consideration the protective actions.

Table 10: Comparative areas under the ROC curve with and without CJ data, and with and without controlling for protective actions taken

Method	Any reassault (w/CJ data) (n = 1307)	Severe reassault (w/CJ data) (n = 1307)	Any reassault w/o CJ data (n = 782)	Severe reassault w/o CJ data (n = 782)	Any reassault w/o CJ data controlling for protective actions (n = 782)	Severe reassault w/o CJ data controlling for protective actions (n = 782)
DA	.613 ***	.628 ***	.635 ***	.670 ***	.674 ***	.687 ***
DV-MOSAIC	.474	.525	.513	.589*	.583 *	.647 ***
DVSI	.487	.567	.508	.597 **	.595 *	.616 **
K-SID	.511	.523	.516	.514	.606 ***	.622 **
Victim's perception of risk	.572 **	.551 *	.599 ***	.610 ***	.619 ***	.619 ***

* p ≤ .05, **p < .01, ***p < .001

Although DA performs the best across the different tests including (columns 1 and 2) or excluding criminal justice data (columns 4–6) and controlling for (columns 5–6) or not controlling for (columns 1–4) protective actions, the other methods perform differently under the different conditions, with the second best method changing across columns. Some explanation is in order of why the different tests show such different results. Including the criminal justice data appears to reduce the predictive accuracy of all the methods and the victim's assessments. The reason is that arrest data were available for all the participants, even those not reached at followup. For those not reached at followup, only arrest data were available—but arrest data drastically underestimate the reassault rate compared with victim reports. (In our data, only 18 percent of the reassaults that victims reported were captured in the criminal justice data.) Therefore, tests using criminal justice as the only measure of reassault for 40 percent of the participants will inflate the false positive rate. That is, assault will be predicted by the method but will not be detected in the arrest data.

Controlling for protective actions improves the predictive ability of all the methods by decreasing the false positive rate. That is, when a method predicts risk but an action is taken that reduces or eliminates the possibility of reassault, the method appears to be overestimating risk (false positive). By taking into account or controlling for such protective actions, the ROC curves better reflect the real predictive accuracy. The accuracy of victims' assessments also improved when controlling for protective actions, but they improved the least of all the methods. The reason for this may be that victims were taking into account their protective action plans when they made their assessment, but this is hard to know for sure.

Overall, DA and victims' estimates were consistently better than chance, with DA performing somewhat better than victims' estimates. As noted above, adjusting for the protective actions taken increased the predictive accuracy of all of the approaches. When controlling for the protective actions taken, all the methods predicted any assault and severe assault better than chance. All the approaches predicted severe assault better than they predicted any abuse, especially DV-MOSAIC. When controlling for protective actions, all the risk assessment methods predicted severe assault better than the victim's own assessment.

Wald Statistic

To assess the methods further, we tested their ability to accurately assess the probability of three different types and levels of outcomes: (1) stalking and/or threats but no physical or sexual abuse, (2) minor or moderate physical or sexual abuse, and (3) severe physical or sexual abuse. We used regression analyses to control for protective actions undertaken during the followup period and assessed each method's ability to predict the three categories of abuse using Wald statistics derived from statistical models. Wald statistics gauge the strength of the association between the risk score derived by a particular risk assessment method, after controlling for protective actions, and are not dependent on the units of measurement, which vary for each method.

Table 11: Results of predictive models controlling for frequency of event and protection actions taken (measured by the Wald statistic)

Method	Prediction of stalking or threatening	Prediction of minor or moderate physical or sexual abuse	Prediction of severe abuse
DA	15.14***	9.58**	27.64***
DV-MOSAIC	16.27***	.89	4.80*
DVSI	13.38***	.13	8.61**
K-SID	.03	2.06	.34
Victim's perception of risk	7.91***	0.12	20.01***

*p < 05. **p < .01, ***p < .001

DA produced the best overall predictive model, as compared with the other methods and victims' predictions (table 11). DA had the only statistically significant Wald statistic for predicting all types/levels of abuse. DA, victim's assessment, DV-MOSAIC, and DVSI had statistically significant predictive capabilities for stalking/threatening and severe abuse. DV-MOSAIC had the highest accuracy for predicting stalking and threatening. The DA's full model of prediction of severe abuse was several times greater than those of the other methods, as was the victim's perception of risk.

Conclusions and Recommendations

This study is the largest prospective test of predictive accuracy of several of the most widely used risk assessment approaches in the field of domestic violence. While the methodology was by no means perfect, it had several important strengths: More than one risk assessment method tested, participants recruited from multiple settings (family court, law enforcement, shelters, health care, and advocacy) in more than one locale (Los Angeles County and New York City), both victim reports and criminal justice records used as outcome measures, collaboration with criminal justice and domestic violence service agencies, victim self-protective steps taken into account, and the substantial diversity in victim characteristics.

On the down side, the retention rate of 60 percent is not ideal. While most of the differences between those who remained in the study and those who did not were not significant, those who could not be retained in the study had higher severity of abuse at baseline than did study participants whom we reinterviewed at followup. Another limitation of the study is its primary reliance on victim interviews for the risk assessments. DA was designed to rely on victim interviews alone, but DVSI, K-SID, and DV-MOSAIC were designed to draw on criminal justice records and information from or about offenders as well as victim interviews. With these additional sources of information, it is possible that their predictive abilities would be increased.

Summary of Findings

The primary findings are:

1. The participants in the study were all women seeking help for violence from their intimate male partners. Ninety percent were non-white or Hispanic. Half were not working, and a third did not have high school degrees. Half were married to the abuser at some point, and the other half were abused by a current or ex-boyfriend; the majority were no longer living with or involved with the offender. They were a severely abused population, with 43 percent experiencing a severe act of violence (as measured by the CTS2) three or more times in the 6 months prior to the baseline interview.

2. In 91 percent of the cases in which we obtained a followup interview, some sort of action was taken after the baseline interview that could have reduced the risk of subsequent abuse. Many of these actions were taken by victims (e.g., avoiding contact with the abusive partner, going to a shelter, getting a protective order, changing locks). Other actions were taken by the criminal justice system (e.g., arrest, incarceration). The entire sample was recruited from points where victims were already receiving services and most were implementing protective actions at baseline. The largest single group in the study was recruited from family court, for example, where the women were interviewed immediately after being granted a protection order. The next largest group had law enforcement involvement, and the third largest group was recruited from shelters. Those with higher risk scores—including victim's assessment of risk—were more likely to take additional protective steps.

3. Despite the protective actions, 31 percent of the women were physically abused between the baseline and followup interviews, a time period of 5 months to more than a year. More than half of the women who experienced any violence during the followup period (56 percent) were severely abused—choked, burned, beaten up, or otherwise seriously hurt—and 36 percent experienced potentially lethal abuse. An additional third of the women endured psychological abuse and/or harassment with no physical abuse, and 16 percent were stalked and/or threatened.

4. In spite of these high rates of repeat assault, only 6 percent of the perpetrators were arrested for domestic violence and an additional 11 percent arrested for other violent crimes. This low rate of reported assaults shows the underestimation of repeat domestic violence when only criminal justice records are used.

5. All four of the risk assessment methods tested were found to be significantly related to subsequent severity of abuse, but not very highly related. After controlling for the protective actions taken, all predicted any assault and severe assault significantly better than chance. DA and the victim's self-rated level of risk had the highest correlations with subsequent abuse, although these correlations were low. DV-MOSAIC, DVSI, and K-SID followed, with their order varying depending on the statistic used. When protective actions were taken into account in statistical tests, all approaches except K-SID had more predictive accuracy than the victim's perception of risk. Again, the predictive capabilities of DV-MOSAIC, DVSI, and K-SID may improve when risk assessment includes criminal justice system information beyond what the victim can provide, as they were designed.

6. The risk assessment methods correctly classified most of the women who were indeed reassaulted (i.e., they showed high sensitivity). This bodes well for the use of risk assessments for victim safety, yet, depending on the method used, from 16 to 33 percent of the women predicted to be at quite low risk subsequently experienced violence. This caveat applies to the victim's own prediction—23 percent of those who rated their risk of being physically abused as low experienced reassault. The false negatives for severe assault (where severe violence was not predicted but was inflicted) were lower, at 3–14 percent. Thirteen percent of the victims who rated their risk of serious physical harm low experienced severe assaults during the followup period. At these rates, false negatives are a serious concern.

7. The risk assessment methods (including victims' predictions) also had a high rate of predicting reassault for women who did *not* experience assaults during the followup period (i.e., the methods had low specificity). Even at the highest levels of predicted risk, all of the methods had a fairly high proportion of these false positives. Low specificity is more of a concern for offender rights than victim safety, but they may have negative effects on the victim as well. She may be unnecessarily fearful or make major changes in her life that may not be necessary. It also conflicts with the goal of providing the most intensive services to those most at risk.

8. A concern of practitioners on the front lines, pressed for time in a crisis situation, is the length of risk assessments. The methods tested varied greatly in length, but on the whole, brevity or length did not correspond with accuracy. DA performed better than the other methods, producing the best overall predictive model. However, on many of the tests, two questions asked of the victim were the second best predictors after the 20 questions covered by DA.

9. Taking into account the impact of protective actions on outcomes improved the methods' accuracy in predicting reassault and even more in predicting severe assault. When taking protective actions that might have prevented assault into account in statistical tests, all the methods performed better than victims' assessments in predicting severe assault, although all methods and the victim's assessment performed significantly better than chance.

10. DV-MOSAIC performed best in predicting subsequent stalking or threats, with DA and DVSI performing well also. DVSI and DV-MOSAIC also show promise for predicting severe abuse, as did

the victim's assessment of risk. K-SID was the most accurate in correctly identifying cases *not* at high risk; i.e., it had the highest levels of specificity.

11. Additional research is needed on all the risk assessment methods, particularly comparing their predictive capabilities against those of a knowledgeable expert (e.g., domestic violence advocate, police officer, investigator), or in combination with expert and victim judgment. To be useful, the risk assessment methods must perform far better than chance. They must also perform better than or enhance expert judgment in order for expert practitioners to find them worth doing.

12. Victims are fairly good predictors of their own risk, yet not accurate enough to depend on alone for risk assessment. Victim self assessments, unlike most instruments, may incorporate both risk and protective factors. Further research in this area may improve risk assessments by integrating victim perceptions. Victim self-assessments may also prove useful as a one-item "screen" for risk, to be followed by more formal risk assessment methods.

13. Additional analysis of the current data is needed to examine separate risk factors more closely. Across the four methods and other items that were not redundant, more than 100 risk factors were included in the interviews. Additional item analysis will shed light on which individual risk factors are most predictive and which might lead to new instruments tailored to different settings and purposes.

Recommendations

While the systematic risk assessment approaches were shown to be better than chance and improved on the victim's own predictions, they are far from being perfectly accurate, and the study did not address whether any of the methods are better than experienced practitioners. To capture most of the higher risk cases, they cast a wide net, at the cost of also capturing a fairly high proportion of cases that are not at risk of further violence. The false negative rates, although not very high, are a great concern for victim safety. The false positives are high and may lead to violations of the rights of offenders and misallocation of resources. Better, more accurate predictive methods would reduce both these problems. In the sexual assault and mental health fields, formal methods of risk assessment have been found to be significantly better than expert judgment, and a combination of formal methods and expert judgment is deemed to be the best approach (Pinard and Pagani, 2000; Hanson and Morton-Bourgon, 2004). Instead of looking at expert judgment versus instruments as an either/or choice, the best approach is probably to gather as much information from as many sources as possible given the time available and circumstances of the assessment. The ideal would be a well-validated instrument specific to domestic violence in the hands of a practitioner who is expert in domestic violence by virtue of training and experience, who listens to a victim who is expert in her particular situation, and who has access to other sources of information. The bottom line purpose for all risk assessment for practitioners is prevention. The risk assessment should be used as a guide to develop effective interventions to be implemented by the system and/or by the victim.

Without further research, we cannot unequivocally recommend a particular approach for use in assessing risk in domestic violence cases. We advise practitioners to:

1. Carefully ask the victim her perception of her risk and take heed of her judgment.

2. Continue to assess risk with all means available, including the expert judgment and clinical wisdom of practitioners (their knowledge of domestic violence and the offender's criminal record); a formal

method with some evidence of predictive accuracy like those tested here; and the victim's own assessment.

3. Where victim safety is your greatest concern, use lower risk categories on formal methods to identify cases for intervention. Where offender fairness and/or scarce system resources are your greatest concern, use higher risk categories to identify cases for sanctioning or intensive services.

4. Be vigilant about potential harm to both victims and perpetrators, as the science of risk assessment is young.

References Cited

(See Final Report for Full Risk Assessment Reference List)

Campbell, J.C. “Nursing Assessment of Risk of Homicide for Battered Women.” *Advances in Nursing Science* 8 (4) (1986): 36–51.

Campbell, J.C. *Assessing Dangerousness*. Newbury Park: Sage, 1995.

Campbell, J.C., D. Webster, J. Koziol-McLain, C.R. Block, D. Campbell, M. Curry, F. Gary, N. Glass, J. McFarlane, C. Sachs, P.W. Sharps, Y. Ulrich, and S. Wilt. “Assessing risk factors for intimate partner homicide.” *National Institute of Justice Journal* 250 (2003): 14–19.

De Becker, G. *The Gift of Fear.* Boston: Little, Brown & Co., 1997.

De Becker, G. & Associates. *Domestic Violence Method (DV MOSAIC),* 2000. http://www.mosaicsystem.com/dv.htm.

Gelles, R., and R. Tolman. *The Kingston Screening Instrument for Domestic Violence (K-SID).* Providence: University of Rhode Island, 1998.

Goodman, L., M.A. Dutton, and L. Bennett. “Predicting Repeat Abuse Among Arrested Batterers: Use of the Danger Assessment Scale in the Criminal Justice System.” *Journal of Interpersonal Violence* 10 (2000): 63–74.

Green, D.M., and J.A. Swets. *Signal Detection Theory and Psychophysics.* New York: Wiley, 1966.

Hanson, R., and K. Morton-Bourgon. *Predictors of Sexual Recidivism: An Updated Meta-Analysis* (User Report 2004–02). Ottawa: Public Safety and Emergency Preparedness Canada, 2004. http://www.psepc.gc.ca/publications/corrections/pdf/200402_e.pdf.

Heckert, D.A., and E.W. Gondolf. “Battered Women's Perceptions of Risk Versus Risk Factors and Instruments in Predicting Repeat Reassault.” *Journal of Interpersonal Violence 19* (7) (2004): 778–800.

Hilton, N.Z., G.T. Harris, M.E. Rice, C. Lang, C.A. Cormier, and K.J. Lines. “A Brief Actuarial Assessment for the Prediction of Wife Assault Recidivism: The Ontario Domestic Assault Risk Assessment.” *Psychological Assessment* 16 (3) (2004): 267–275.

Kropp, P.R. “Some Questions Regarding Spousal Assault Risk Assessment.” *Violence Against Women* 10 (6) (2004): 676–697.

Pinard, G.F., and P. Pagani, eds. *Clinical Assessment of Dangerousness: Empirical Contributions.* New York: Cambridge University Press, 2000.

Roehl, J., C. O’Sullivan, D. Webster, and J.C. Campbell. *Intimate Partner Violence Risk Assessment Validation Study, Final Report.* Washington, DC: National Institute of Justice, 2005. http://www.ncjrs.org/pdffiles1/nij/grants/209731.pdfT.

Weisz, A., R. Tolman, and D.G. Saunders. “Assessing the Risk of Severe Domestic Violence.” *Journal of Interpersonal Violence* 15 (1) (2000): 75–90.

Williams, K.R., and A.B. Houghton, “Assessing the Risk of Domestic Violence Re-offending: A Validation Study.” *Law and Human Behavior* (2004): 437–455.