An Examination of the Interrater Reliability and Concurrent Validity of the Spousal Assault Risk Assessment Guide – Version 3 (SARA-V3)

by

Tara J. Ryan

B.S. (Magna Cum Laude), Creighton University, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Fall 2016

Approval

Name:	**Tara J. Ryan**
Degree:	**Master of Arts**
Title:	***An Examination of the Interrater Reliability and Concurrent Validity of the Spousal Assault Risk Assessment Guide – Version 3 (SARA-V3)***
Examining Committee:	**Chair:** Dr. Robert Ley Associate Professor

Stephen D. Hart
Senior Supervisor
Professor

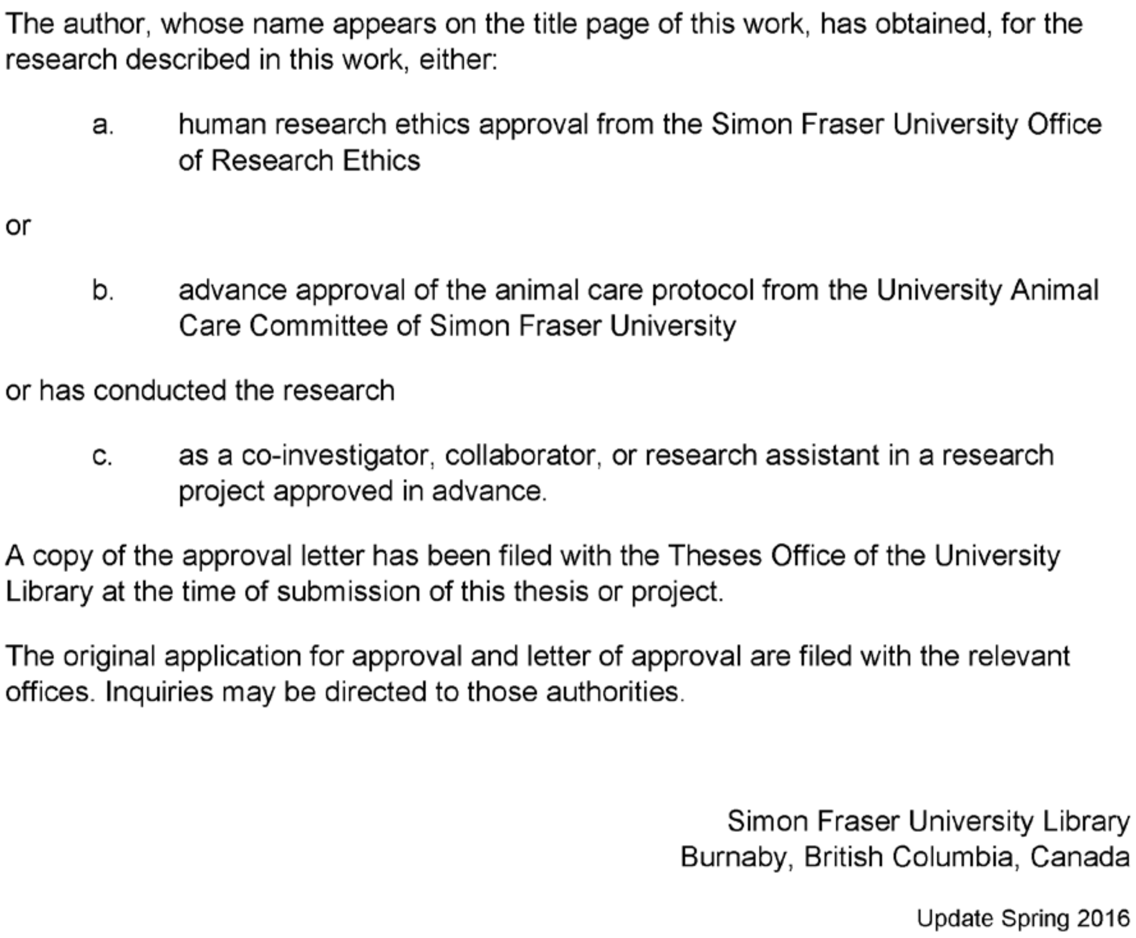
Kevin S. Douglas
Supervisor
Professor

P. Randall Kropp
Supervisor
Registered Psychologist

Mark Olver
External Examiner
Professor
Department of Psychology
University of Saskatchewan

Date Defended/Approved: December 08, 2016

Ethics Statement



The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The Spousal Assault Risk Assessment Guide-Version 2 (SARA-V2; Kropp, Hart, Webster, & Eaves, 1995, 1999, 2008) is one of the most widely used Intimate Partner Violence (IPV) risk assessment tools in the world. After over 20 years, the SARA has been updated to reflect advances in research related to IPV and risk assessment more generally. The purpose of this thesis is to examine the interrater reliability and concurrent validity of the most recent version of the Spousal Assault Risk Assessment-Version 3 (SARA-V3). A total of *N* = 97 closed IPV cases were used to rate SARA-V3. To examine interrater reliability, a second rater coded *n* = 30 of the same files using SARA-V3. Interrater reliability for individual risk factors, SARA-V3 numerical total scores, and summary risk ratings fell primarily in the moderate range and consistent with prior research. Other raters had previously coded the same files with SARA-V2 and a number of other IPV risk assessment tools, and these tools served as the basis for evaluating the concurrent validity of the SARA-V3. ICCs were mostly in the fair to good range indicating adequate interrater reliability. Correlations between SARA-V3 and other IPV risk assessments were medium to large indicating good concurrent validity. Overall, the interrater reliability and concurrent validity findings were in line with previous research on SARA-V2 and the other assessments of IPV risk. Limitations of this study and implications for future research and practice are discussed.

Keywords: Intimate partner violence; Spousal Assault Risk Assessment; SARA; violence risk assessment; SARA reliability; SARA validity

Dedication

Dedicated to my mom.

I cannot thank her enough for her continued sacrifices and unconditional support as I pursue my dreams and ambitions.

Acknowledgements

Many thanks to my senior supervisor, Dr. Stephen Hart, for his guidance, patience, infinite wisdom, and kindness during this process. Not many graduate students have the freedom to research and pursue opportunities almost completely of their choosing. I am lucky to say that Steve rarely says no and instead asks - how can we make this happen? Thank you, Steve.

I would also like to thank my committee members and mentors, Drs. Kevin Douglas and Randy Kropp, for their insight, supervision, and kindness not only in the process of completing this thesis but also throughout my time at SFU thus far.

I would not be in graduate school but for the mentorship and support of Dr. Matthew Huss. I cannot thank him enough for igniting my interest in forensic psychology (nearly a decade ago!). His continued support and advice as I pursue my graduate training has been invaluable. I often ask myself, “Am I working as hard as Dr. Huss would be?” The answer is often no, but it gives me something to strive for anyway.

Special thanks to my lab mate and friend, Sarah Coupland. This thesis would not have been possible without her many hours of work. It is always nice to gain a colleague and friend for life during grad school. Thanks Sarah!

Thank yous are owed to former Hart students - Jen Storey and Ashley Murray. Thanks for all of your prior work coding data that was included in this study and for answering a lot of questions via email. Also, I would like to thank Karen Whittemore and all of the staff at the Forensic Clinic in Surrey for being great friends of research as I conducted this study.

I would like to thank Dylan Gatner for his peer mentorship, often offered while jogging around the seawall. I cannot help but to believe that our many jogs are at least partially responsible for me getting this thesis done. Also, thanks to Nicole Muir, Jenny Pink, Kat O’Donnell, and Yan Lim for your friendship, advice, and insight as we wind our way through grad school.

To the best group of friends a gal could ask for - Lee, Kelly, Erin, Maddie, Casey, Dani, Corrie, Alycia, and Emily - thanks for always picking up right where we left off whenever I make it home and for all of the encouragement during the times in between.

Finally, an emphatic thanks to my mom and brother for being unconditionally supportive of my pursuits.

Table of Contents

List of Tables

Chapter 1. Introduction

The most common form of violence experienced by women throughout the world is intimate partner violence (United Nations Department of Public Information, 2009). Also known as spousal assault, wife assault, and domestic violence, intimate partner violence (IPV) is a serious problem. Ellsberg, Jansen, Heise, Watts, and Garcia-Moreno (2008) found that lifetime prevalence rates for physical and sexual IPV ranged from 15% to 71% worldwide. The World Health Organization (2012) examined the prevalence of IPV in Bangledesh, Ethiopia, Namibia, Peru, and Tanzania and found widespread IPV with 13% to 61% of women reporting lifetime prevalence of physical violence by a partner and 4% to 49% reporting severe physical violence throughout the lifetime. The data reveal that IPV is not bound by geography or culture.

Defined as the actual, attempted, or threatened physical harm of a current or former intimate partner (Kropp & Hart, 2015), IPV taxes multiple systems including criminal justice, health care, and social services. The most current data from Statistics Canada indicate that domestic disturbances use vast police resources in this country. In 2013, there were over 90,000 victims of police-reported IPV across Canada, corresponding to over 25% of all police-reported victims of violent offenses (Canadian Centre for Justice Statistics, 2015). Between 2005 and 2011, 57% of the more than 300,000 violent offenses tried in Canadian courts involved IPV. Furthermore, during that same time span over 90,000 IPV offenders were convicted and received a sentence that required supervision or incarceration. Based on interviews with victims via the National Crime Victimization Survey in the United States, there were over 634,000 reported incidents of IPV (defined as violence committed by current or former spouses, boyfriends, or girlfriends) in 2014. Of those incidents, over 265,000 were considered serious (e.g., sexual assault or aggravated assault) (Truman & Langton, 2015). Phone calls pertaining to domestic violence related incidents received by police within the United States outnumber any other type of emergency call (Klein, 2009).

In addition to the impacts on the criminal justice system, the negative health outcomes for victims of IPV have been well established and include increased relative risk prevalence rates for chronic disease, mental illness, injury, and drug use (Coker et al., 2002). In a review of the literature, Golding (1999) found that rates of depression, suicidality, PTSD, and alcohol and drug abuse are far higher among female victims of IPV than non-victims. Monetarily, IPV is not a cheap societal problem. It has been estimated that IPV costs the federal government in Canada between $1.5 billion and $15.7 billion annually (Bowlus, McKenna, Day, & Wright, 2003; Day, 1995; Zhang, Hoddenbagh, McDonald, & Scrim, 2012). The cost of IPV to employers in Canada—the consequence of distraction, absenteeism, and tardiness among other problems affecting productivity—is estimated to be more than $6,000 annually for each employee who is victimized. In the United States, similar research has estimated the total cost of IPV at more than $5.8 billion dollars annually, including productivity losses of $728 million and household productivity losses estimated at $131 million (Max, Rice, Finkelstein, Bardwell, & Leadbetter, 2004). Complicating the picture, treatment outcomes for IPV offenders are mixed. High recidivism rates reflect the chronicity of the problem and the difficulty in treating IPV offenders (Babcock, Green, & Robie, 2004; Feder & Wilson, 2005; Sartin, Hansen, & Huss, 2006).

Due to these various issues, one area that IPV research and practice should focus on is risk assessment and management of violence. These tasks are ultimately geared toward violence prevention. High prevalence and varied rates of recidivism following treatment, as well as the many probable negative outcomes for victims of IPV, have prompted the creation of a growing number of empirically informed risk assessments for offenders of IPV. These assessments are increasingly used to aid a variety of professionals in understanding and managing IPV risk.

Violence Risk Assessment for Intimate Partner Violence

The last few decades have resulted in tremendous advances with respect to understanding and assessing violence risk. One important area of work has been the development of structured risk assessment tools. Decades of research on clinical decision making has certainly influenced risk assessment. In Paul Meehl's seminal book, *Clinical vs. Statistical Prediction: A Theoretical Analysis and Review of the Evidence (1954),* he outlined a debate among psychologists: is unstructured clinical judgment reliable or should statistical (actuarial) methods instead be implemented in place of clinical judgment? Meehl examined the reliability of clinical judgment (i.e., diagnostic) by reviewing 20 studies. In half of these studies, clinical judgment and actuarial judgment were equally accurate. In the other half, actuarial methods outperformed clinical judgment. In only one study was clinical judgment more accurate than actuarial methods. Meehl firmly came down on the side of actuarial methods, arguing that clinical judgment was imprecise.

Over twenty years later, Cocozza and Steadman (1978) conducted a study related to future violence risk. Cocozza and Steadman found that unstructured opinions of dangerousness made by psychiatrists for men who had been charged with felony crimes were wildly inaccurate and mostly based on the seriousness of the alleged index offense. Cocozza and Steadman also found that psychiatrists rarely provided a detailed or rational explanation for their determination of dangerousness. In this study, defendants convicted of more serious crimes were much more likely to be found dangerous (83% of those charged with violent felonies compared to 50% of those charged with non-violent crimes). However, when psychiatrists provided explanations for their determination of dangerousness, just 11.5% cited the defendant’s charge. More concerning was that those who were designated as dangerous were arrested at a lower rate than those designated as not dangerous (49% versus 54%) as well as rearrested for violent offenses (14% versus 16%). Cocozza and Steadman made the bold claim that psychiatrists making dangerousness determinations for the court of law were clearly practicing outside their area of competency.

More recent empirical endeavours have found actuarial methods to be superior to that of unstructured clinical judgment or intuition. Grove, Zald, Lebow, Snitz, and Nelson (2000) conducted a meta-analysis of 136 studies ranging from 1944 to 1989 and found a small overall effect size in favor of actuarial methods (*M* = .12; in favor of statistical methods). In another meta-analysis spanning 60 years of published research, Ægisdóttir et al. (2006) found an overall small effect in favor of statistical methods, *N* = 41 studies and 48 effects; *d* = -.12 in favor of statistical methods, 95% CI [-.14, - .09]. Overall, these results reflected a 13% increase in the accuracy of statistical methods over clinical judgment. The authors concluded that statistical prediction methods are more accurate than unstructured clinical judgment or intuition and that the small effect size should not be ignored in practice. These findings, dating back over 60 years, provide evidence that actuarial methods, sometimes integrating statistical algorithms, have improved upon unstructured clinical judgment. Many in the field have extended these conclusions to the area of violence risk assessment. Others have suggested that the prediction of future violent behavior should *only* be algorithmic and that clinical judgment, being inferior, should be replaced by statistical actuarial methods (Quinsey, Harris, Rice, & Cormier, 1998). However, others in the field believe that the end result of a violence risk assessment is not only to *predict* future violence, but to rather to manage a person's potential for future violence and ultimately prevent the risk for future violence.

Douglas, Cox, and Webster (1999) outline a broader view of violence risk assessment in which prediction of future violence is a "necessary first step" (p. 155), but view risk management and prevention as equally important in the process of violence risk assessment. Douglas et al. acknowledge that there is empirical evidence that actuarial methods are superior to unstructured clinical judgment, specifically in terms of predictive validity. However, the authors argue that structured decision making based on empirical methods that are grounded in research ameliorate many of the critiques of the actuarial approach to violence risk assessment: static risk factors and disregard for idiosyncratic risk factors, lack of setting generalization, focus on prediction only, applying aggregate data to individual cases, et cetera. Abundant literature provides support for the notion that structured decision making methods, formally known as Structured Professional Judgment within the context of violence risk assessment, are a valid alternative to both actuarial and unstructured clinical judgment (Douglas, Hart, Groscup, & Litwack, 2013; Hart, 1998; Litwack, 2001; Liwack, Zapf, Groscup, & Hart, 2006).

The 1990s and early 2000s proved to be an especially productive time in the area of IPV risk assessment as several assessment tools were developed with varying degrees of published empirical validity (Dutton & Kropp, 2000). Ultimately, and in line with general violence risk assessment, there are currently various assessments that generally fall within one of the two aforementioned frameworks: actuarial risk assessment and Structured Professional Judgment (SPJ). As unstructured clinical judgment is generally agreed upon as unreliable and unsupported by research and practice guidelines, this approach will not be considered in this investigation.

Actuarial Violence Risk Assessment Instruments

Actuarial methods, according to Meehl (1954), are defined by a fixed algorithm or set of a priori decision making rules. Actuarial violence risk assessment instruments often rely on a statistical algorithm, however statistical procedures are not required for actuarial decision making. The main goal of most actuarial risk assessment instruments is to determine the probability of recidivism. After deriving an overall score from the instrument, the rater compares the score of the individual to risk categories or "bins" established by the construction sample resulting in an absolute risk probability of recidivism. It is assumed that actuarial risk assessment instruments can accurately estimate, based on group level data, individual risk to recidivate (Harris, Rice, Quinsey, & Cormier, 2015). Many of the individual items comprising actuarial risk assessment instruments tend to be static in nature. Due to this, it is often difficult to obtain a lower future score even if, for example, psychological treatment has improved risk level or if the offender has desisted in criminal behavior. Interrater reliability and concurrent validity for actuarial risk assessment instruments tend to be excellent. Various actuarial risk assessment instruments for different types of violence have been developed since the 1990s with moderate to good levels of predictive accuracy (Dvoskin & Heilbrun, 2001; Quinsey et al., 1998; Yang, Wong, & Coid, 2010). A number of actuarial risk assessment instruments have been developed and appear to be useful for IPV risk assessment including the Ontario Domestic Assault Risk Assessment (ODARA; Hilton, et al., 2004), the Domestic Violence Risk Appraisal Guide (DVRAG; Hilton, Harris, Rice, Houghton, & Eke, 2008) and the Danger Assessment (DA; Campbell, 1986; Campbell et al., 2003). See the Method section for a more detailed explanation of these risk assessment instruments.

Structured Professional Judgment (SPJ)

Emerging in the mid-1990s, the SPJ approach to violence risk assessment focuses on the prevention and management of future violence, rather than the calculation of an exact probability of likely future violence (Douglas & Kropp, 2002; Heilbrun, 1997). Within the SPJ model of risk assessment, Douglas and Kropp (2002) note that there is a focus on dynamic risk factors, or those that change over time, rather than static factors as in actuarial risk assessment. Additionally, ongoing assessment and monitoring and management of violence risk by focusing on dynamic risk factors is key in the SPJ framework. SPJ manuals and guides have evolved as over time to be more reflective of the complete risk assessment process, from the coding of risk factors to providing management recommendations. For example, manuals that were published early in the development of the SPJ approach provided risk factors and evaluators were directed to gather information and then to code the presence or absence of each risk factor and also consider and code for critical factors – a risk factor that compels the evaluator to determine an imminent risk for harm exists (Kropp, Hart, Webster, & Eaves, 2008). Presently, the SPJ approach to violence risk assessment has three general phases including identifying facts, making meaning of those facts, and then taking action and often these phases are divided into 6 or 7 specific steps within the risk assessment manual (Douglas et al., 2014).

As in the actuarial approach, most SPJ tools are developed based on the type of violence the assessment aims to evaluate. Although there are some risk factors common in most types of violence, it is usually the case that specific types of violence (i.e., stalking, IPV, sexual violence) have specific or unique risk factors that should be considered in order to effectively manage and prevent future possible violence. However, both approaches do include assessments for general violence. In most ways, the development of SPJ tools is starkly contrasted to that of actuarial risk assessments. In order to minimize sample dependence, risk factors on SPJ tools are derived based on a thorough literature review (Douglas et al., 2014) accompanied by consultation from subject matter experts (researchers, law enforcement, etc.). Often after the initial draft of the manual is written, it is piloted in the field and changes are then made based on feedback from users. To date, two IPV SPJ risk assessment tools have been developed including Spousal Assault Risk Assessment (SARA; Kropp, Hart, Webster, Eaves, 1994, 1995, 1999, 2008), and the Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp, Hart, & Belfrage, 2005, 2010).

The Spousal Assault Risk Assessment Guide (SARA)

The SARA was the first risk assessment tool developed and published within the SPJ theoretical framework. The initial SARA guide was published in 1994 with a second revision and edition published in 1995. Subsequent revisions in 1999 and 2008 included only minor changes to the text and are considered Version 2. From this point forward, the second version will be abbreviated as SARA-V2. The first stage of development was a literature review to establish meaningful risk factors associated with IPV (Kropp et al., 2008). From there, authors pared down relevant risk factors to 20 factors comprising Parts 1 and 2. See Table 1.1 for a breakdown of SARA-V2 risk factors. Part 1 includes risk factors related to criminal history and psychosocial adjustment and Part 2 has risk factors related to spousal assault history and the index offense. Since its inception, SARA-V2 has become one of the most widely used risk assessments for IPV around the world and the guide has been translated into numerous languages (Hanson, Helmus, & Bourgon, 2007; Kropp & Hart, 2015).

SARA-V2 administration procedures are outlined in the manual. First, users code each of the risk factors on a 3-point ordinal scale representing the user's judgment of the presence of the risk factor for the examinee (*Present, Possibly or partially present, Absent*). Users can omit risk factors when not enough information exists to code the factor. Second, users rate the presence of critical factors on a 2-point ordinal scale (*No, absent or Yes, present)*. Critical factors are those which the rater deems significant enough on their own conclude the examinee poses an imminent risk of harm. Finally, raters will code their summary risk judgments about the case related to two areas of risk: imminent risk of harm to spouse and imminent risk of harm to another person(s). Summary risk judgments are coded on a 3-point ordinal scale (*Low, Moderate, High)*. Generally, previous research has provided evidence that SARA-V2 has good to excellent interrater reliability, moderate to good predictive validity, and good to excellent concurrent validity when compared to other IPV risk assessment tools.

Table 1.1 Spousal Assault Risk Assessment Guide, Version 2 (SARA-V2): Parts and Risk Factors

SARA-V2 Part/Risk Factor
Part 1: Psychosocial Adjustment
1. Past assault of family members
2. Past assault of strangers or acquaintances
3. Past violation of conditional release or community supervision
4. Recent relationship problems
5. Recent employment problems
6. Victim of and/or witness to family violence as a child or adolescent
7. Recent substance abuse/ dependence
8. Recent suicidal or homicidal ideation/intent
9. Recent psychotic and/or manic symptoms
10. Personality disorder with angry, impulsivity, or behavioral instability
Part 2: History of Spousal Assault
11. Past physical assault
12. Past sexual assault/sexual jealousy
13. Past use of weapons
14. Recent escalation in frequency or severity of assault
15. Past violations of “no contact” orders
16. Extreme minimization or denial of spousal assault history
17. Attitudes that support or condone spousal assault
18. Severe and/or sexual assault
19. Use of weapons and/or credible threats of death
20. Violation of “no contact” order

Peer-Reviewed Research on the SARA-V2

SARA-V2 has been the focus of more empirical investigations than many other IPV risk assessment instruments and in general the research suggests that risk decisions made using SARA-V2 are reliable and valid. For a narrative review of SARA-V2, see Dutton and Kropp (2000) and for a meta-analysis see Helmus and Bourgon (2011). Several representative studies are summarized below.

Kropp and Hart (2000) conducted the first psychometric analysis of SARA-V2 in a field evaluation with a sample of N = 2,681 offenders. SARA-V2 ratings were made by probation officers, treatment staff including doctoral-level psychologists, counselors, and social workers, research assistants, and case managers. The sample consisted of both n = 1,615 offenders sentenced to probation (assumed to be lower risk) and n = 627 offenders sentenced to a custodial sentence (assumed to be higher risk). Although not standard administration procedure (i.e., SARA-V2 risk factors are not summed), continuous scores were derived based off of three separate calculations for the purposes of research: sum of the risk factors, number of risk factors coded as present, and number of risk factors rated as critical. Internal consistency for the total score was good (a = .78) and item homogeneity (mean inter-item correlations) for Part 1 was .16 and .21 for Part 2. There was good agreement among raters both at the risk factor level, *Mdn* ICC_1 = .60 with a range .45 to .86 and for summary risk ratings, *Mdn* ICC_1 = .63. Summed SARA-V2 risk factors, Part 1 and Part 2 summed risk factors were moderately correlated with the Screening Version of the Hare Psychopathy Checklist-Revised (PCL:SV; Hart, Cox, & Hare, 1995). There was a large, positive correlation between summed Part 1 risk factors and VRAG total scores, r = .50, $p \leq .001$. Upon offender follow-up for recidivism, SARA-V2 demonstrated moderate predictive validity, Area Under the Curve (AUC) = .70.

Grann and Wedin (2002) conducted a file review study (N = 88) in which SARA-V2 ratings were made by two raters. The offenders who comprised the sample had been convicted of spousal assault or homicide. The authors described the ratings as actuarial in nature – the raters did not assign any risk factors as “critical” (a step in the risk assessment process when using SARA-V2) and also summed the 20 risk factors. The researchers collected follow-up recidivism data on the cohort until December 31, 1995. A total of 25 men (28%) within the sample received further convictions for crimes that were IPV in nature. Authors noted that Risk Factors 3 (past violation of a conditional release or community supervision), 10 (personality disorder with anger, impulsivity, or behavioral instability), and 16 (extreme minimization or denial of spousal assault history) were particularly important to the risk of recidivism within the sample. The 5-year follow-up AUC for total SARA-V2 scores fared best with an AUC of .65 compared to shorter follow-up time frames. SARA-V2 demonstrated weaker predictive validity for one-year follow-up reconviction prediction, AUC = .59, compared to the PCL-R and VRAG, AUC = .71 and .75, respectively.

In a different type of validation study, Heckert and Gondolf (2004) were interested in if the perceptions of risk of IPV victims were more accurate or could add incrementally to the validity of risk ratings on several IPV risk assessment. The authors examined SARA-V2, the Kingston Screening Instrument for Domestic Violence (K-SID; Gelles & Tolman, 1998), and DA. SARA-V2 risk factors were summed in order to compare across tools and predictive validity analyses were conducted. The DA alone was the best predictor across tools, AUC = .70, compared to the SARA-V2, AUC = .64, and the K-SID total score, AUC = .57. When combined with women's perceptions of their own risk for violence, AUC values improved slightly for the DA and SARA-V2, AUC = .73 and AUC = .69, respectively.

Hilton et al. (2004) presented the construction methods and psychometric data for the Ontario Domestic Assault Risk Assessment (ODARA; Hilton et al., 2004). The ODARA was compared to other assessments for IPV, including SARA-V2. The ODARA and SARA-V2 summed risk factors were largely correlated, r = .60, p < .01. The researchers found SARA-V2 had fair predictive validity (AUC = .64) in the ODARA construction sample totalling N = 589, but poor predictive validity in the cross-validation sample, AUC = .54. All of the assessments analyzed in the cross-validation sample, including the ODARA, performed worse. The authors attributed this decline to sampling error rather than shrinkage.

Williams and Houghton (2004) conducted a prospective study of N = 1,465 men arrested for domestic violence offenses committed against their female partners in Colorado using the Domestic Violence Screening Instrument (DVSI; Williams & Houghton, 2004). The authors examined concurrent and predictive validity via comparisons with SARA-V2. DVSI total scores and SARA-V2 summed risk factors had a large association, r = .57. Based on official police records, a total of n = 776 (53%) of the sample reoffended during the 18-month follow-up period. The SARA-V2 predictive validity for IPV reoffending was fair, AUC = .65, and for any reoffending was good, AUC = .70. Predictive validity for the DVSI was lower for both IPV and general violence reoffending, AUC = .60 and AUC = .68, respectively.

Finally, Belfrage et al. (2012) conducted a prospective field study in which police officers rated SARA-V2. The study was conducted across three counties in Sweden and the sample was comprised of *N* = 429 male-to-female IPV offenders who were charged with a range of criminal acts including assault, unlawful threat, harassment, and breach of peace. Police officers completed SARA-V2 ratings for risk factors and summary risk ratings following the investigation of the incident of IPV. Officers were also tasked with establishing risk management plans such as victim safety planning. The initial ratings resulted in *n* = 201 (47%) of the offenders being rated at low risk, *n* = 169 (39%) were rated as moderate risk, and *n* = 59 (14%) were rated as high risk. Following the initial data collection, *n* = 93 (21%) offenders had subsequent contact with police over an 18-month follow-up period in which the SARA-V2 was scored again. This resulted in a statistically significant increase in SARA-V2 mean ratings from 11.48 upon first contact to 13.04 post-second contact. Total scores and summary risk ratings were correlated with the number of management strategies recommended (*r* = .40) and recidivism was higher with higher total SARA-V2 scores, AUC = .63. However, summary risk ratings were not as consistently associated with summary risk ratings, AUC = .57. Finally, the use of more management strategies in cases that received a high risk rating were associated with decreased recidivism.

Spousal Assault Risk Assessment Guide—Version 3 (SARA-V3)

Spousal Assault Risk Assessment Guide-Version 3 (SARA-V3; Kropp & Hart, 2015) is the most recent iteration of the guide. SARA-V3 was developed according to the standard SPJ process as outlined by Douglas et al. (2014). The updates to the guide reflect advancements in the violence risk assessment process. More specifically the newest version includes guidance for risk formulation, risk scenario planning, and risk management planning. SARA-V3 also includes victim vulnerability factors to aid in victim safety planning. Generally, SARA-V3 authors adopted the newer developments in SPJ theory and the administration procedures are similar to those in the latest versions of the Historical-Clinical-Risk Management 20, Version 3 (HCR-20 V3; Douglas, Hart, Webster, & Belfrage, 2013) and Guidelines for Stalking Assessment and Management (SAM; Kropp, Hart, & Lyon, 2008).

The guide also gives guidance for evaluators to take action to recommend management strategies in a more structured manner than previous versions. SARA-V3 domains and risk factors are presented in Table 1.2. Compared to SARA-V2, SARA-V3 has undergone several changes including the addition of new risk factors, reorganization of existing factors, and it is comprised of three domains as opposed to two parts including risk factors as they relate to the Nature of IPV (N domain factors), Perpetrator risk factors (P domain factors), and Victim Vulnerability Factors (V domain factors). None of the risk factors from SARA-V2 were eliminated and remain in SARA-V3, were folded into other risk factors, or may have been renamed. SARA-V3 has additional steps that assist evaluators in making meaning of facts and risk ratings.

There are six primary steps to administering the SARA-V3 risk assessment guide. After the initial step of information gathering, Step 2 requires the user to rate the presence of the individual risk factors on a 3-point ordinal scale (*Present, Possible or partially present, Not present)* across the three domains. Where no information exists to code a risk factor it can be omitted. Ratings are coded both in the recent (anytime in the last 12 months prior to the evaluation) and past (anytime prior to the last 12 months). Three risk factors related to examinee problems with mental disorder and personality disorder and victim mental disorder that require a 2-point rating (*Definite, Provisional)* to indicate if a qualified professional has conducted a psychological evaluation and diagnosed these problems or if the rating is based off observation. In Step 3, users rate the relevance of risk factors on a 3-point ordinal scale (*Yes, Possible or partially, No)* to indicate if the risk factor is important for management planning considerations. Again, relevance can be omitted for ratings in which low or no information is available. In Step 4, users engage in risk scenario planning. Step 5 requires the user to recommend management plans in the areas of monitoring or surveillance, treatment/assessment, supervision/control, and victim safety planning. Finally, in Step 6 users rate their summary risk judgments in 4 areas: case prioritization, risk for serious physical harm, imminent violence, and other risks indicated on a 3-point ordinal scale (*Low or Routine, Moderate or Elevated, High or Urgent)* and also document a case review date.

Table 1.2 Spousal Assault Risk Assessment Guide, Version 3 (SARA-V3): Domains and Risk Factors

SARA-V3 Domain/Risk Factor
Nature of IPV
N1. Intimidation
N2. Threats
N3. Physical harm
N4. Sexual harm
N5. Severe IPV
N6. Chronic IPV
N7. Escalating IPV
N8. IPV-related supervision violations
Perpetrator Risk Factors
P1. Intimate relationships
P2. Non-intimate relationships
P3. Employment/finances
P4. Trauma/victimization
P5. General antisocial conduct
P6. Major mental disorder
P7. Personality disorder
P8. Substance use
P9. Violent/suicidal ideation
P10. Distorted thinking about IPV
Victim Vulnerability Factors
V1. Barriers to security
V2. Barriers to independence
V3. Interpersonal resources
V4. Community resources
V5. Attitudes or behavior
V6. Mental health

Note. IPV = intimate partner violence.

Current Study

After more than 20 years, the SARA has been updated to incorporate advancements in research related to IPV and violence risk assessment more generally. The purpose of the current study is to evaluate the reliability and validity of risk judgments made using SARA-V3. This is the first empirical evaluation of the SARA-V3. In this study, I will be comparing SARA-V3 to established measures for IPV within the SPJ framework including SARA-V2 and the Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp, Hart & Belfrage, 2005, 2010) and additionally to established actuarial risk assessment instruments including ODARA, DVRAG, and the DA.

Research Questions

Research Question 1. What is the distribution of risk ratings on SARA-V3?

Research Question 2. What is the association among ratings of risk factors on SARA-V3?

Research Question 3. What is the interrater reliability of risk ratings made using SARA-V3?

Research Question 4: What is the association between risk ratings made using SARA-V3 and of those made using other procedures to assess risk for IPV?

Chapter 2. Method

Overview

To conduct this study, I took advantage of an existing dataset of *N* = 100 coded closed case IPV offender files that were referred to an outpatient forensic clinic for assessment in British Columbia, Canada. Evaluations for these offenders were made during the presentencing stage of criminal proceedings. A total of *N* = 97 of the original 100 cases were used in the current study as three files were unavailable at the time of coding. The dates of original assessment for these closed files ranged from 2000 to 2009. The offenders within this sample were convicted of at least one offense committed against an intimate partner (e.g., assault, threatening, breach of no-contact order). The sample is best described as one of “moderate risk” as most of the offenders were released on bail while awaiting sentencing in the community. The original assessment for risk was conducted by a registered clinical psychologist who had specialized forensic training in graduate school. The psychologist used the SARA-V2 to guide decisions about risk in order to prepare a presentence report for the courts used to aid judges in sentencing determinations. In addition to SARA-V2 risk ratings that were completed by the registered psychologist, this dataset included a number of additional IPV risk assessments previously coded by two graduate students (JS and AM) who were enrolled in a forensic psychology PhD program at the time ratings were coded. The use of this existing data set permitted raters for the current study to code SARA-V3 blind to the ratings on the other IPV risk assessments.

One empirical paper evaluating the risk estimates of the DA has been published based on the dataset being used in the current study (Storey & Hart, 2013). In the study, JS coded all *N* = 100 files whereas AM coded a subset *n* = 23 files to evaluate interrater reliability. In general, administration and scoring of the DA (more detailed administration procedures are outlined in the Procedures section below) is typically based only on an interview with the IPV victim. Storey and Hart (2013) purposefully varied the recommended administration procedures in that each rater coded the DA twice: the first time the DA was coded based only on a victim interview notes (the interview was not conducted specific to scoring the DA). The second round of coding the DA was based on additional file information that included offender criminal history and an interview with the offender as well as the victim interview information. The additional round of coding that included more offender information resulted in changed DA scores for *n* = 14 offenders and a small, overall mean increase in total scores, *M* = .60. Although a small difference on average was found between these two methods of scoring the DA, DA scores based only on victim interview information systematically resulted in lower total scores (in one case a difference of 15 points) than when additional offender file information was also used. Study authors concluded that this was perhaps due to victim minimization or victim lack of knowledge of the offender's past. Storey and Hart concluded that the systematically lower scores are potentially problematic when using the DA for risk decisions. Due to these systematic differences, the authors suggested more guidance from the DA authors and that perhaps the DA should always be scored with any additional offender file information when possible. Storey and Hart also found that DA total scores tended to estimate higher risk levels than the other IPV risk assessments on the same cases. In this study, the DA was positively associated with other IPV risk assessment tools with moderate effect sizes.

Participants

All offenders in the sample were male. The mean age at the time of initial assessment was 36.04 (*SD* = 8.92) and a range of about 19 years old to approximately 61 years old. The ethnic breakdown of the sample was predominantly Caucasian (63%) and Indo Canadian (25%) with Asian and black men each comprising 3% of the sample, and First Nations men comprising 4%. Within the sample, most participants reported being married (10%), divorced or separated (40%), or single (38%). Many offenders were charged with more than one index offense and most commonly assault. A total of 41 other charge types included charges such as assault with a weapon, attempt murder, various breaking and entering charges, and various mischief charges.

Procedure

Two graduate students (TR and SC) who are enrolled in a clinical-forensic psychology program, each with over 120 hours of formal risk assessment coursework and training, completed SARA-V3 training. Training was administered by SARA-V3 co-author, Dr. P. Randall Kropp. TR coded *N* = 97 files while SC coded a subset of selected files, *n* = 30, for the purposes of interrater reliability analysis. In the prior research study, two doctoral graduate students, JS and AM, coded the B-SAFER, ODARA, DVRAG, and DA. SARA-V2 risk ratings were derived from the registered psychologist's report and notes based on her original risk assessment – coders did not rate SARA-V2. The previously coded ratings were used in this study for concurrent validity analysis. Of the 30 files coded for interrater reliability in the current study, *n* = 23 were the same files selected for interrater reliability analysis coded by AM.

Of the *n* = 30 interrater files, *n* = 23 files were used to establish consensus ratings. For these files, TR and SC met to discuss their individual ratings and where differences arose, they came to an agreement on the risk ratings at the individual risk factor level and for conclusory opinions. TR and SC coded the first five consensus files independently, but then met after each one to establish consensus ratings and in order to work out any coding problems. The consensus files thereafter were coded on average every 7-10 cases to ensure interrater agreement and similar coding methods. Raters deviated from SARA-V3 protocol in that the ratings were made based on a review of file information only. The file information was typically substantial and included the original offender interview and victim interview notes from the time that the risk assessment was completed in addition to police records related to the incident, and risk assessment reports that were prepared for the courts. For the purposes of this investigation, TR and SC remained blind to each other's risk ratings as well as to the risk ratings included in the presentence reports and to the risk ratings made on the B-SAFER, ODARA, DVRAG, and DA. TR and SC treated the interview data for the initial presentencing forensic assessment as the current date in order to establish proper timelines per SARA-V3 use guidelines.

Measures and Materials

SARA-V3

SARA-V3 is the most recent and third version of the SARA and encompasses three domains of IPV including the Nature of IPV (N domain, 8 risk factors), Perpetrator Risk

Factors (P domain, 10 risk factors), and Victim Vulnerability Factors (V domain, 6 factors). There are a total of six administration steps outlined in detail within the SARA-V3 manual: Step 1: information gathering; Step 2: Rate the presence of risk factors; Step 3: Rate the relevance of risk factors and formulate the case; Step 4: Scenario plan; Step 5: Determine risk management plans; Step 6: Determine conclusory risk judgments. Raters TR and SC completed Steps 1-3 and also Step 6 for the purposes of this study. Based on a review of file information, TR coded N = 97 SARA-V3 assessments and SC coded a subsample of n = 30 SARA-V3 assessments. A total of 30 omitted risk factor ratings across the N and P domains were coded as 0. It was much more difficult to code V domain factors in this sample and therefore only a limited number of V domain factors were able to be coded.

Ratings for N, P, and V risk factors were coded for two time frames: recent (anytime in the last 12 months prior to the evaluation) and in the past (anytime prior to the last 12 months). Factor presence coding was completed on a 3-point ordinal scale (*Present, Possible or partially present, Not present*). Relevance coding was also on a three-point ordinal scale (*Yes, Possible or partially relevant, Not relevant)*. For research purposes, this coding scheme was converted to numerical scores (*Present/Yes* = 2*, Possible or partially present/relevant* = 1*, Not present/relevant* = 0). For simplicity of analyses, “presence” ratings were created, essentially combining past and recent ratings for N and P factors. In other words, the maximum coded rating across time on each risk factor was used to create an ever present rating with a possible range of 0 to 36. Domain numerical scores were created by summing risk factor ratings across the respective N and P domains. A total numerical score was also created by summing N and P present risk factor ratings.

SARA-V2

SARA-V2 is the second version of the SARA and includes 20 risk factors organized into Parts 1 and 2. Administration procedures are outlined in detail in the guide include information gathering, coding the presence of risk factors, coding the presence of critical factors (i.e., risk factors that indicate the examinee poses imminent risk for violence), and making summary risk judgements. Risk factor presence coding is on a 3-point ordinal scale (*Present, Possible or partially present, Not present*). For research purposes, this coding scheme can be converted to numerical scores (*Present* = 2*, Possible or partially present* = 1*, Not present* = 0) and summed creating a range from 0 to 40. In this study, a total of *n* = 84 SARA-V2 assessments were completed by a registered clinical psychologist for the purposes of presentencing reports that were submitted the courts to aid in sentencing decisions. A total of 24 risk factor ratings across the cases were omitted and were coded as 0 for this analysis.

In the present study, SARA-V2 numerical total scores ranged from 5 to 36 with a mean of 20.06 (*SD* = 6.21). Parts 1 and 2 had means of 10.15 (*SD* = 3.89) and 9.90 (*SD* = 3.66), respectively. The distribution of case prioritization ratings were as follows: *Low*, *n* = 4 (5%), *Moderate, n* = 51 (61%) and *High, n* = 29 (34%). These risk judgments were recorded based on the risk ratings of a registered clinical psychologist and therefore interrater reliability analyses are not possible.

Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp, Hart & Belfrage, 2005, 2010)

The B-SAFER was adapted from the SARA-V2 and developed as a front-line IPV assessment tool for police. The B-SAFER comprises 15 risk factors organized into 3 Sections with 5 factors in each (Section I: Perpetrator Risk Factors: Intimate Partner Violence, Section II: Perpetrator Risk Factors: Psychosocial Adjustment, and Section III: Victim Vulnerability Factors. Administration procedures are outlined in the B-SAFER manual. After gathering case information in Step 1, users rate the presence of risk factors in Step 2 based on a 3-point ordinal scale (*Yes, Possible or partially present, No).* Users may omit risk factors when no information is available to code. In Step 3 users judge risk factor relevance and then select from a number of risk management strategies across four areas: monitoring/surveillance, assessment/treatment, control/supervision, and victim safety planning. Finally, in Step 4 the user determines summary risk ratings (*Low/routine, Moderate/elevated, High/urgent*) related to case prioritization, risk for life threatening violence, and imminent violence and also document likely victims (current/former intimate partner, family/friends of current/former intimate partner, other). Additionally, users document a case review date or reassessment of risk timeline.

In the present study, B-SAFER assessments coded by two graduate (JS and AM) students were available for all *n* = 97 cases. No B-SAFER risk factors were omitted during data collection. Ratings for risk factors were coded for two time frames: currently (anytime in the last 4 weeks prior to the evaluation) and in the past (anytime prior to the last 4 weeks in the person's life). Factor presence coding is on a 3-point ordinal scale (*Present, Possible or partially present, Not present*). For research purposes, this coding scheme can be converted to numerical scores (*Present/Yes* = 2*, Possible or partially present/relevant* = 1*, Not present/relevant* = 0) and then summed for a numerical total score ranging from 0 to 20. Maximum ratings across time on each risk factor were used to create an "ever" rating with a possible range of 0 to 30.

In the present study, B-SAFER numerical presence total scores ranged from 11 to 29 with a mean of 21.46 (*SD* = 3.68). Case prioritization ratings were as follows: *Low*, *n* = 11 (11%), *Moderate, n* = 57 (58%) and *High, n* = 29 (30%). Interrater reliability of the numerical total score in the subsample of *n* = 30 cases was evaluated using the Intraclass Correlation Coefficient (ICC). A full explanation of the ICC models used in this study with recommended interpretive guidelines are covered in the Results section. Interrater reliability of the numerical total scores was $ICC_{2,1}$ = .81 and $ICC_{2,2}$ = .89 and for the Case Prioritization summary risk rating was $ICC_{2,1}$ = .53 and $ICC_{2,2}$ = .70. For other studies of reliability and validity on the B-SAFER see Au et al. (2008), Belfrage and Strand (2012), Storey, Kropp, Hart, Belfrage, and Strand (2014).

Ontario Domestic Assault Risk Assessment (ODARA; Hilton et al., 2004)

The ODARA, an actuarial risk assessment instrument, was developed from an Ontario sample of IPV offenders (*N* = 589) who were followed-up over a period of 5 years (Hilton et al., 2004). The ODARA was designed for use by frontline workers (i.e., police officers) in emergent IPV incidences and therefore an in depth review of the offender's psychological and criminal history is not needed to score the assessment. Hilton and colleagues collected data on the construction sample in six areas (sociodemographic characteristics, domestic violence history, general criminal history, relationship characteristics, victim characteristics, index offense details) multivariate regression analyses were used to determine the combination of variables that resulted in the optimal prediction of recidivism. Recidivism was defined as a subsequent physical assault on an intimate partner. The risk assessment instrument includes 13 dichotomously scored items that are coded according to explicit rules and are all weighted equally. Items are summed for a total score ranging from 0 to 13 and can then be compared to cut-off scores organized into 7 risk categories or bins that each has an associated estimated recidivism probability: Bin 1 = 5%; Bin 2 = 10%; Bin 3 = 20%; Bin 4 = 27%; Bin 5 = 41%; Bin 6 = 59%; Bin 7 = 70%.

In the present study, ODARA assessments coded by two graduate (JS and AM) students were available for all *n* = 97 cases. A total of three item ratings were missing for the entire sample and were coded as 0. Total scores ranged from 0 to 10 with a mean of 6.63 (*SD* = 2.08). Interrater reliability in the subsample of *n* = 23 cases was $ICC_{2,1}$ = .81 and $ICC_{2,2}$ = .90. Interrater reliability for risk bins was $ICC_{2,1}$ = .89 and $ICC_{2,2}$ = .94. In terms of allocation across risk bins, this sample had the following distribution: Bin 1, *n* = 1 (1%); Bin 2, *n* = 1 (1%); Bin 3, *n* = 3 (3%); Bin 4, *n* = 2 (2%); Bin 5, *n* = 5 (5%); Bin 6, *n* = 38 (39%); and Bin 7, *n* = 47 (49%). For other studies on the ODARA, see Hilton et al. (2008) and Messing and Thaller (2012).

Domestic Violence Risk Appraisal Guide (DVRAG; Hilton et al., 2008)

The DVRAG is an actuarial IPV risk assessment tool that includes all 13 ODARA items, but also incorporates one additional item – the offender's score on the Psychopathy Checklist-Revised (PCL-R; Hare, 1991, 2003). Contrary to the ODARA, the DVRAG is not intended for use by frontline workers to make immediate decisions about violence risk (Hilton et al., 2009). Instead the DVRAG is intended to be coded with more thorough file information as it includes a PCL-R rating. Another difference between the DVRAG and the ODARA is that not all items on the DVRAG are coded dichotomously. Scores range from -10 to 37. Total scores are organized into 7 risk categories with associated probabilities of recidivism: Bin 1 = 2%; Bin 2 = 22%; Bin 3 = 43%; Bin 4 = 63%; Bin 5 = 81%; Bin 6 = 97%; Bin 7 = 100%. In the current study, DVRAG assessments coded by two graduate (JS and AM) students were available for all *n* = 97 cases. Instead of the PCL-R, raters coded the PCL:SV based on file information.

In the present study, DVRAG total scores ranged from -7 to 41 with a mean of 16.04 (*SD* = 10.26). Interrater reliability for total scores was $ICC_{2,1}$ = .70 and $ICC_{2,2}$ = .82. Interrater reliability for the risk bins was $ICC_{2,1}$ = .80 and $ICC_{2,2}$ = .88. Distribution of this sample across risk bins was as follows: Bin 1, *n* = 0 (0%); Bin 2, *n* = 1 (1%); Bin 3, *n* = 4 (4%); Bin 4, *n* = 10 (10%); Bin 5, *n* = 15 (15%); Bin 6, *n* = 48 (50%); and Bin 7, *n* = 19 (20%). For other studies on the DVRAG, see Hilton et al. (2008) and Rettenberger and Eher (2015).

The Danger Assessment (DA; Campbell, 1986; Campbell et al., 2003)

The DA was initially developed to predict the danger of male-to-female and female-to-male inflicted homicide in domestically violent relationships (Campbell, 1986). The assessment development started with a reviewed literature and consultation with battered women, women's shelter workers, as well as law enforcement and other experts on IPV (Campbell, 1986). The first version of the DA was a relatively short assessment that could be administered by either a healthcare professional, criminal justice professional, victims advocate, or could be completed by the victim. It was comprised of two parts – a calendar in which the victim would indicate incidences of battering in the past year and rate the incidences on a scale of one-to-five. The scale ratings were a combination of the severity of the violence as well as the type of violence used against the victim. The second part of the DA included 15 yes-no questions scored dichotomously (0 = *no,* 1 = *yes*) (Campbell, 1995). Scores were summed to produce a total score in which a high score was indicative of a greater risk for lethal violence, however no cut-off scores to estimate grouped level of risk were provided. The DA was initially validated on a sample of *N* = 79 female victims of IPV recruited through the community and women's shelters in two cities within the United States. Campbell and colleagues (2003) conducted a multi-site study in order systematically identify risk factors associated with femicide within the context of IPV with the purpose of updating the DA. The resulting DA left the calendar portion unchanged, however the some of the specific items answered by the victim changed. This is the current version of the DA and includes 20-items that are scored according to a weighted algorithm with possible scores ranging from -3 to 39. Campbell et al. (2003) also included cut-off scores and categories: *Variable Danger* (0-7), *Increased Danger* (8-13), *Severe Danger* (14-17), *Extreme Danger* (18+).

In the present study, DA assessments coded by two graduate (JS and AM) students were available for all *n* = 97 cases. There was a mean of 2 DA items missing per case; all missing items were coded as 0. In this sample, DA total scores ranged from 1 to 31. The DA total score mean was 16.46 (*SD* = 6.79). Rater agreement for total DA scores was $ICC_{2,1}$ = .88 and $ICC_{2,2}$ = .93. In terms of risk categories, *n* = 15 (15%) of cases fell in the *Variable Danger* category, *n* = 18 (19%) fell in the *Increased Danger* category, *n* = 15 (15%) fell in the *Severe Danger* category, and *n* = 49 (51%) fell in the *Extreme Danger* category. For other studies on the DA, see Campbell, (1986), Campbell et al. (2005), Goodman, Dutton, and Bennett (2000), Heckert and Gandolf (2004), and Hilton et al. (2004).

Chapter 3. Results

Research Question 1. What is the distribution of risk ratings on SARA-V3?

If a risk assessment guide is going to be useful, not everyone being assessed with the guide should receive the same rating. In other words, an adequate distribution of risk ratings is necessary. To examine this, I looked at the distribution in three ways: means and standard deviations, frequency across individual risk factors, and distribution of summary risk ratings. I expected SARA-V3 risk ratings to have adequate distribution across *Present, Possibly or partially present,* and *Not present* ratings. I also expected adequate distribution of *Low, Moderate,* and *High* summary risk ratings, but with more cases falling in the *Moderate* risk range than in *Low* or *High*. The distribution of individual risk factor presence and relevance ratings means and standard deviations are presented in Table 3.1. As you can see in the table, the pattern of location and dispersion across domains is acceptable. The relevance factors numerical total score mean was slightly less than the P domain, 12.59 (*SD* = 3.62). This is an indication that raters discriminated between mere presence of P risk factors and their relevance to the offender's risk level.

Table 3.1 Distribution of SARA-V3 Presence and Relevance Numerical Scores, Total and Domain

SARA-V3 Numerical Score	Presence		Relevance	
	M	(*SD*)	*M*	(*SD*)
Total (N+P)	24.88	(4.90)		
Nature of IPV	11.14	(2.62)		
Perpetrator Risk	13.73	(3.60)	12.59	(3.62)

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence.

The frequency of endorsement across presence (i.e., ever present) and relevance ratings are presented in Table 3.2. There was particularly high frequency of endorsement of presence ratings for factors N1, N2, N3, P1, P3, P8, and P10. Overall, relevance ratings tended to follow the same pattern of frequency of endorsement as presence ratings throughout the P domain. There was a paucity of information pertaining to specific circumstances of many victims and therefore many ratings of *Omit* were coded.

The spread of summary risk ratings across *Low, Moderate, High,* are presented in Table 3.3. As you can see in the table, a majority of the cases were designated as moderate risk in the Case Prioritization, Serious Physical Harm and Risk for Imminent Violence summary risk ratings. In terms of Case Prioritization, 57% of the cases were rated as moderate risk whereas high and low risk ratings comprised 30% and 13% of the cases, respectively. Risk for Serious Physical Harm and Risk for Imminent Violence summary risk ratings followed a similar pattern with slightly more in each group being rated as low risk. This result was again expected due to the general level of moderate risk offenders comprising the sample.

Overall, all N and P factors were endorsed at least some of the time providing evidence that each risk factor was relevant to IPV cases. Within this sample, some N and P factors were endorsed very frequently. Some of these were factors that are expected to occur often in cases of IPV including problems with intimate relationships and problems with distorted thoughts about IPV. Relevance ratings were scored as *Yes* and *Possible/Partially relevant* at a slightly lower rate overall compared to factor presence ratings. This provided some evidence that raters were appropriately differentiating presence and relevance. As expected, no factors were scored as relevant if the factor was not rated as present.

Table 3.2 Distribution of SARA-V3 Presence and Relevance Ratings for Individual Risk Factors

SARA-V3 Domain/Risk Factor	Presence				Relevance			
	Y	P	N	O	Y	P	N	O
Nature of IPV								
N1. Intimidation	91%	0%	9%	0%	--	--	--	--
N2. Threats	85%	1%	13%	1%	--	--	--	--
N3. Physical harm	93%	3%	4%	0%	--	--	--	--
N4. Sexual harm	17%	8%	74%	1%	--	--	--	--
N5. Severe IPV	41%	10%	47%	1%	--	--	--	--
N6. Chronic IPV	70%	20%	10%	0%	--	--	--	--
N7. Escalating IPV	68%	25%	7%	0%	--	--	--	--
N8. IPV-related sup…	60%	0%	40%	0%	--	--	--	--
Perpetrator Risk								
P1. Intimate rel…	100%	0%	0%	0%	100%	0%	0%	0%
P2. Non-intimate rel…	60%	17%	24%	0%	46%	25%	29%	0%
P3. Employment/fin…	89%	9%	2%	0%	71%	20%	9%	0%
P4. Trauma/victimization	35%	25%	39%	1%	11%	25%	63%	1%
P5. General ant…	42%	28%	30%	0%	41%	25%	34%	0%
P6. Major mental dis…	20%	29%	52%	0%	19%	23%	59%	0%
P7. Personality disorder	35%	29%	36%	0%	53%	11%	36%	0%
P8. Substance use	85%	8%	7%	0%	81%	8%	10%	0%
P9. Violent/suicidal ide…	39%	27%	34%	0%	28%	26%	46%	0%
P10. Distorted thinking…	96%	0%	4%	0%	95%	4%	1%	0%
Victim Vulnerability								
V1. Barriers to security	19%	28%	1%	53%	34%	11%	1%	54%
V2. Barriers to ind…	55%	2%	2%	41%	53%	3%	2%	42%
V3. Interpersonal res…	2%	1%	2%	95%	2%	1%	2%	95%
V4. Community res…	1%	0%	0%	99%	1%	0%	0%	99%
V5. Attitudes or behavior	30%	16%	16%	37%	31%	11%	19%	39%
V6. Mental health	3%	1%	0%	96%	1%	2%	1%	96%

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence; Y = *Present*; P = *Possibly present*; N = *Not present*; O = *Omit*: -- = rating not applicable. See Table 1.2 for complete names of risk factors.

Table 3.3 Distribution of SARA-V3 Summary Risk Ratings

SARA-V3 Summary Risk Rating	High	Moderate	Low
Case Prioritization	30%	57%	13%
Serious Physical Harm	14%	63%	23%
Imminent Violence	16%	56%	28%

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence.

Research Question 2. What is the association among ratings of risk factors on SARA-V3?

In my analysis of SARA-V3, I was guided by Slaney, Storey, and Barnes (2011). This paper outlined a logical method and guidelines for evaluating the evidence for the reliability and validity of psychological assessments. Slaney et al. (2011) offer a framework when testing validity. First, the internal validity of the test must be examined to determine if any risk factors are redundant. Only after evidence for internal validity has been established should reliability be examined.

The risk factors that comprise SARA-V3 were chosen because they are not redundant, or in other words, the risk factors were chosen for their specific and unique variance. For research purposes it is common to sum risk factors in order to derive a total numerical score. In order to justify deriving a total numerical score, the association among risk factors was examined. It was important that associations were neither too small or large. If any of the individual risk factors were too largely correlated with the others in the domain, this would indicate that the risk factor contributed minimal unique variance and therefore should potentially be excluded. If associations were too small, this would be a likely indication that not all of the important risk factors for IPV comprised SARA-V3. Given these considerations, I expected the association between risk factors to be in the small to medium ranges indicating a low level of redundancy, but adequate enough associations to compute numerical total scores. I examined three different types of associations among SARA-V3 risk factors to decide if it was reasonable to compute total numerical scores:

corrected item-total correlations (CITCs), Cronbach's alpha (*α*) and mean inter-item correlations (MICs). Due to low endorsement, analyses of the V domain were not possible.

CITC values were calculated for presence ratings for each factor in the N domain. CITCs for the presence and relevance ratings for each factor in the P domain were also calculated. All CITCs are presented in Table 3.4. As shown in the table, two patterns emerge. CITCs for the N domain presence ratings are lower than for presence ratings in the P domain. Within the N domain presence ratings, two CITC values resulted in near zero correlations (N4 and N8). Other than these two risk factors, CITC values ranged from .16 to .35. Additionally, most P factor presence and relevance CITC values fell above .30 and ranged from .15 to .59. Overall, associations between risk factors that comprise SARA-V3 are neither too small or large indicating individual risk factors add unique variance to the assessment of IPV.

Cronbach's *α* and MIC calculations are located in Table 3.5. Cronbach's *α* ranged from .43 to .71 across domain and total numerical scores and MIC values ranged from .10 to .21. P domain presence and relevance ratings as well as total numerical scores for combined N and P domain numerical total scores were more consistent than N domain numerical scores. Generally, there was more correspondence between the P factors on SARA-V3 and the risk factors on SARA-V2. As the original risk assessment for each offender was based on SARA-V2, it is possible that information to rate these factors was more prevalent in the file. Many SARA-V3 N domain risk factors may not have been inquired about during the original assessment because many of these risk factors are not represented on SARA-V2.

There was moderate to strong evidence that SARA-V3 risk factors are associated with small to medium correlations indicating risk factors have unique variance that is not redundant. Associations as measured by Cronbach's *α* and MICs provide more evidence that the risk factors comprising SARA-V3 represent a good sampling of the possible risk factors for measuring IPV. Taking all measures of association among risk factors into consideration, SARA-V3 demonstrated adequate evidence to support summing numerical total scores for latter analyses in this study.

Table 3.4 Corrected Item-Total Correlations (CITC) for SARA-V3 Presence and Relevance Numerical Scores for Individual Risk Factors, By Domain

SARA-V3 Domain/Risk Factor	Presence	Relevance
Nature of IPV		
N1. Intimidation	.24	
N2. Threats	.35	
N3. Physical harm	.21	
N4. Sexual harm	.07	
N5. Severe IPV	.16	
N6. Chronic IPV	.24	
N7. Escalating IPV	.23	
N8. IPV-related sup…	.09	
Perpetrator Risk		
P1. Intimate rel…	--	--
P2. Non-intimate rel…	.54	.59
P3. Employment/fin…	.31	.22
P4. Trauma/victimization	.44	.37
P5. General ant…	.45	.51
P6. Major mental dis…	.26	.27
P7. Personality disorder	.53	.52
P8. Substance use	.33	.22
P9. Violent/suicidal ide…	.37	.38
P10. Distorted thinking…	.15	.18

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence; -- = not calculated due to lack of variance. See Table 1.2 for complete names of risk factors.

Table 3.5 Cronbach's α and Mean Inter-item Correlation (MIC) for SARA-V3 Presence and Relevance Numerical Scores, Total and Domain

SARA-V3	Presence		Relevance	
Domain/Risk Factor	α	MIC	α	MIC
Total (N+P)	.66	.10		
Nature of IPV	.43	.10		
Perpetrator Risk	.71	.21	.69	.19

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence.

Research Question 3. What is the interrater reliability of risk ratings made using SARA-V3?

Interrater Reliability of SARA-V3 Steps 2, 3, and 6

If raters cannot agree on the presence of risk factors or on summary risk ratings for a case using SARA-V3, it is not a helpful guide for clinicians and others who make violence risk decisions. As such, interrater reliability was analyzed via ICCs in several different ways for adequate coverage on this topic. Fleiss and Shrout (1977) outlined the various ICC models when assessing interrater reliability. The first step in the decision process for model selection is determining if the research design is a one-way or two-way analysis of variance. The research design in this analysis included *n* = 30 interrater reliability cases in which two raters coded the same files and an equal number of files. This can be described as a Target X Judges two-way ANOVA (Fleiss & Shrout, 1977). Additionally, in this investigation I was interested in the random effects generated by judges - how ratings generalized across potential raters. For both of these reasons, ICC Model 2 was the most appropriate model of analysis in the current study. In addition to making a determination about the overall model for ICC analysis, the type of rater agreement must be specified. Two types of agreement exist: consistency and absolute agreement. Consistency ignores variance in agreement between raters and instead measures reliability within a rater's judgments. Absolute agreement estimates variance between raters, which was of great interest to me in this investigation. Therefore, within ICC Model 2, I used absolute agreement. Finally, I was interested in single rater ICC values (as denoted by $ICC_{2,1}$) and I am also interested in how the interrater reliability generalizes, so ICC $_{2,2}$ (average measures) were also examined. The quality of rater agreement was evaluated with the interpretive guidelines suggested by Landis and Koch (1977): < .00 = poor, .00 to .20 = slight, .21 to .40 = fair, .41 to .60 = moderate, .61 to .80 = substantial, .81 to 1.00 = almost perfect. I chose the Landis and Koch interpretive guidelines over others (e.g. Cicchetti & Sparrow, 1981) because they offer finer grain discriminations at lower levels of reliability.

I expected ICCs for individual risk factors and summary risk ratings to be smaller than for total numerical scores, but in the moderate range. Also, I expected rater agreement for total numerical scores to be in the substantial to almost perfect ranges and for averaged ICC values to be larger than single rater values. ICC values for individual presence and relevance risk factors are presented in Table 3.6 and 3.7 and several patterns emerge. First, single rater ICCs across the N and P domains for presence ratings fell in the moderate to almost perfect ranges. Additionally, two risk factors had small ICC values: N1 and P10. An examination of N1 showed that raters agreed in 73% of cases and of those, 70% were coded *Present*. In regard to P10, raters agreed on 90% of ratings, all of which were coded *Present*. A similar pattern was found with P1, which could not be calculated due to zero variance across ratings (97% of cases were coded *Present*). Two relevance ratings also had similar patterns in which raters agreed on *Present* relevance for 87% and 93% of cases for P1 and P10, respectively. Presence ratings for N1, P1, and P10 and relevance ratings for P1 and P10 had extreme endorsement frequencies and the high base rate of *Present* ratings and resulting low ICCs were due to a lack of variability – artifacts in this analysis. The lower bound limits of the confidence intervals for single rating ICCs for all other risk factors fell mostly within the moderate to substantial ranges with the exception of P3, P5, and P7, which fell in the fair range. All other upper bound limits fell in the substantial to almost perfect ranges. As expected, averaged rater ICC values generally fell in the substantial to almost perfect ranges. Coding most V factors was not possible due to a lack of file information and therefore no analyses were conducted.

Table 3.6 Interrater Reliability (ICC) of SARA-V3 Presence Numerical Scores for Individual Risk Factors

SARA-V3	Single Ratings		Averaged Ratings	
Domain/Risk Factor	ICC$_{(2,1)}$	95%CI	ICC$_{(2,2)}$	95%CI
Nature of IPV				
N1. Intimidation	.10	[-.28, .44]	.18	[-.78, .61]
N2. Threats	.83***	[.69, .91]	.91***	[.80, .96]
N3. Physical harm	.66***	[.40, .82]	.79***	[.57, .90]
N4. Sexual harm	.74***	[.51, .86]	.85***	[.68, .93]
N5. Severe IPV	.56***	[.25, .76]	.72***	[.41, .86]
N6. Chronic IPV	.55***	[.24, .76]	.71***	[.38, .86]
N7. Escalating IPV	.41**	[.07, .67]	.58**	[.14, .80]
N8. IPV-related sup…	.92***	[.83, .96]	.96***	[.91, .98]
Perpetrator Risk				
P1. Intimate rel…	--	--	--	--
P2. Non-intimate rel…	.73***	[.51, .86]	.85***	[.68, .93]
P3. Employment/fin…	.50**	[.18, .73]	.67**	[.30, .84]
P4. Trauma/victimization	.79***	[.60, .89]	.88***	[.75, .94]
P5. General ant…	.53***	[.22, .74]	.70***	[.36, .85]
P6. Major mental dis…	.72***	[.49, .86]	.84***	[.66, .92]
P7. Personality disorder	.54***	[.23, .75]	.70***	[.37, .86]
P8. Substance use	.86***	[.73, .93]	.92***	[.84, .96]
P9. Violent/suicidal ide…	.70***	[.46, .85]	.82***	[.63, .92]
P10. Distorted thinking…	-.04	[-.39, .33]	-.07	[-1.29, .49]

Note. *N* = 30. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence; ICC = intraclass correlation coefficient; -- = not calculated due to lack of variance. ICCs calculated using 2-way random effects model, absolute agreement. See Table 1.2 for complete names of risk factors.
* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001

Table 3.7 Interrater Reliability (ICC) of SARA-V3 Relevance Numerical Scores for Individual Risk Factors

SARA-V3 Domain/Risk Factor	Single Ratings		Averaged Ratings	
	$ICC_{(2,1)}$	95%CI	$ICC_{(2,2)}$	95%CI
Perpetrator Risk				
P1. Intimate rel…	--	--	--	--
P2. Non-intimate rel…	.73***	[.51, .86]	.85***	[.69, .93]
P3. Employment/fin…	.50**	[.18, .73]	.67**	[.30, .84]
P4. Trauma/victimization	.79***	[.60, .89]	.85***	[.68, .93]
P5. General ant…	.53***	[.22, .74]	.78***	[.54, .89]
P6. Major mental dis…	.72***	[.49, .86]	.85***	[.68, .93]
P7. Personality disorder	.54***	[.23, .75]	.67**	[.30, .85]
P8. Substance use	.86***	[.73, .93]	.89***	[.77, .95]
P9. Violent/suicidal ide…	.70***	[.46, .85]	.78***	[.54, .90]
P10. Distorted thinking…	-.04	[-.39, .33]	-.06	[-1.30 , .50]

Note. *N* = 30. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence; ICC = intraclass correlation coefficient; -- = not calculated due to lack of variance. ICCs calculated using 2-way random effects model, absolute agreement. See Table 1.2 for complete names of risk factors.
* $p \le .05$; ** $p \le .01$; *** $p \le .001$.

Full rater agreement results for summary risk ratings, completed in SARA-V3 Step 6, are presented in Table 3.8. As indicated by the table, summary risk ratings had variable interrater agreement. The primary summary risk rating, Case Prioritization and Risk for Imminent Violence had slightly less than expected rater agreement and fell in the fair range. Raters agreed on 60% of Case Prioritization ratings across cases and 43% of these were rated as *Moderate.* Similarly, raters agreed on 57% of Risk for Imminent Violence ratings. Rater agreement for Serious Physical Harm was in the substantial range, $ICC_{(2,1)}$ = .68. Confidence intervals across summary risk ratings for single rater ICCs were quite variable and ranged from lower bound intervals in the poor range to the moderate range while upper bound limits fell in the moderate to almost perfect ranges. Averaged rater confidence interval limits had a similar pattern of results.

At the individual risk factor level, there was evidence for moderate to substantial interrater reliability for most N and P risk factors comprising SARA-V3. Several risk factors including N1, P1, and P10 had low ICC values, but after a closer inspection of the data these lower values were a result of inadequate variance. Some summary risk ratings fell in lower ranges than expected, but this may be due to a restricted range and not enough low and high risk cases. Finally, confidence intervals tended to be less precise, however despite a relatively small interrater reliability subsample of *n* = 30 files, many of the lower bound limits for the confidence interval still fell in the moderate range.

Table 3.8 Interrater Reliability (ICC) of SARA-V3 Summary Risk Ratings

SARA-V3 Summary Risk Rating	Single Ratings		Averaged Ratings	
	$ICC_{(2,1)}$	95%CI	$ICC_{(2,2)}$	95%CI
Case Prioritization	.40*	[.05, .66]	.57**	[.09, .80]
Serious Physical Harm	.68***	[.44, .83]	.81***	[.61, .91]
Imminent Violence	.41*	[.07, .67]	.58**	[.13, .80]

Note. *N* = 30. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; ICC = intraclass correlation coefficient. ICCs calculated using 2-way random effects model, absolute agreement.

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

Interrater Reliability of Total Numerical Scores

Interrater reliability for presence and relevance numerical domain and total scores are presented in Table 3.9. As indicated by the table, single and averaged rater agreement for all domains and total numerical scores fell in the almost perfect range with the exception of the N domain, which fell in the substantial range. Interestingly, combined P domain risk factors had slightly greater rater agreement than combined N and P factors, $ICC_{2,1}$ = .89. All ICCs were significant at $p \leq .001$. Confidence intervals fell predominantly in the substantial range; however, the N domain single rater lower bound limit fell in the moderate range. Overall, across the SARA-V3 domains, rater agreement was reliable, falling in the substantial to near perfect ranges.

Table 3.9 Interrater Reliability (ICC) of SARA-V3 Presence and Relevance Numerical Scores, Total and Domain

SARA-V3	Single Ratings		Averaged Ratings	
Domain/Risk Factor	ICC$_{(2,1)}$	95%CI	ICC$_{(2,2)}$	95%CI
Presence				
Total (N+P)	.85***	[.71, .93]	.92***	[.83, .96]
Nature of IPV	.73***	[.51, .86]	.84***	[.68, .93]
Perpetrator Risk	.89***	[.78, .95]	.94***	[.88, .97]
Relevance				
Perpetrator Risk	.87***	[.75, .94]	.93***	[.86, .97]

Note. *N* = 30. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; IPV = intimate partner violence; ICC = intraclass correlation coefficient. ICCs calculated using 2-way random effects model, absolute agreement. * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

Prior research on SPJ risk assessments show a similar pattern of results in that total numerical scores have greater interrater reliability as indicated by ICCs than do categorical summary risk ratings. This study does not prove to be an exception to this frequent finding. Overall, numerical total scores had substantial to almost perfect rater agreement. The rater agreement for the summary risk ratings, specifically Case Prioritization and Risk for Imminent Violence, were less than what was expected. Although rater agreement for some summary risk ratings were in the fair range, none of the ratings for the cases differed by more than one category. In other words, there were no cases in which TR rated a case as *High* for Case Prioritization and SC rated the same case *Low* or vice versa*.* These mixed results could be the result of several things. As no ratings differed by more than one category, it is possible that the raters in this study disagreed about what type of case warrants a *Low* or *Moderate* versus *Moderate* or *High* rating of overall risk. Restricted range offered by a 3-category rating system inherently reduces variance and therefore ICCs. Additionally, a majority of the cases within the *n* = 30 interrater reliability sample were in the moderate range of risk, leading to a restricted range and a lack of overall variance within this subsample.

Research Question 4: What is the association between risk ratings made using SARA-V3 and of those made using other procedures to assess risk for IPV?

Another way of adjudicating SARA-V3 is determining the extent to which the guide corresponds to other well established risk assessments for IPV including SARA-V2, B-SAFER, ODARA, DVRAG, the DA. In this study, concurrent validity was tested via correlations between SARA-V3 presence numerical scores and the total scores of other measures of IPV risk. Understanding the strength and direction of associations between various violence risk assessments is important to establish concurrent validity. All correlations were interpreted according to Cohen (1988): small, r = .10; medium, r = .30; large, r =.50.

Associations between SARA-V3 and SARA-V2

In the examination of associations between SARA-V3 and V2, I expected large correlations between numerical total scores. I also expected the SARA-V3 P domain to to have a large association with Part 1 of SARA-V2 as many of the risk factors overlap. Additionally, I expected a small association between the P domain and Part 2 as many of the risk factors included in Part 2 are located in the N domain in SARA-V3. Associations between SARA-V3 and SARA-V2 total numerical scores are presented in Table 3.10. The expected pattern of associations occurred between the two versions of SARA. Correlations between SARA-V3 presence numerical scores and SARA-V2 numerical scores were large and positive. With the exception of the SARA-V3 P domain and SARA-V2 Part 2, all associations were statistically significant and most fell in the medium to large ranges. As expected, SARA-V3 P domain and SARA-V2 Part 2 were not correlated, r = .11. Many of the risk factors included in Part 2 match on well with N domain risk factors and this is evidenced by a large, positive association between the N domain and Part 2. As expected, the largest correlation was between SARA-V3 P domain and SARA-V2 Part 1, r = .78, p ≤ .001. Additionally, SARA-V3 total presence numerical scores and total SARA-V2 numerical scores were positively correlated, r = .66.

Table 3.10 Concurrent Validity of SARA-V3 Presence Numerical Scores: Correlation (*r*) with SARA-V2 Numerical Scores

SARA-V3 Presence Numerical Scores	SARA-V2 Numerical Scores		
	Total	Part 1	Part 2
Total (N+P)	.66***	.68***	.39***
Nature of IPV	.51***	.26*	.59***
Perpetrator Risk	.52***	.74***	.10

Note. *N* = 84. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; SARA-V2 = Spousal Assault Risk Assessment Guide, Version 2; IPV = intimate partner violence.
* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

Correlations between SARA-V3 and SARA-V2 summary risk ratings were also examined and are presented in Table 3.11. SARA-V3 has three summary risk ratings whereas SARA-V2 has one. Summary risk ratings between the two versions were moderately and positively correlated and all significant at *p* ≤ .01.

Table 3.11 Concurrent Validity of SARA-V3 Summary Risk Ratings: Correlation (*r*) with SARA-V2 Summary Risk Rating

SARA-V3 Summary Risk Rating	SARA-V2 Summary Risk Rating
Case Prioritization	.31**
Serious Physical Harm	.30**
Imminent Violence	.28**

Note. *N* = 84. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; SARA-V2 = Spousal Assault Risk Assessment Guide, Version 2; IPV = intimate partner violence.
* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

The results provide strong evidence of concurrent validity between the two most recent versions of SARA-V2. Both numerical domain and total scores for SARA-V3 and SARA-V2 had positive and moderate to large associations with the exception of SARA-V3 P domain and SARA-V2 Part 2. Although some Part 2 risk factors corresponded with the P domain, Part 2 risk factors tended to match-on better to N domain risk factors, which was demonstrated by the results.

Associations Between SARA-V3 and B-SAFER

SARA-V3 and B-SAFER are considered parallel forms for IPV risk assessment. The B-SAFER was derived from SARA-V2 and is most appropriate for frontline use, whereas SARA-V3 is best used for more extensive risk assessment procedures in which an interview can be conducted and file information can be reviewed. Based on this, I expected large associations between SARA-V3 and B-SAFER total numerical scores. Additionally, SARA-V3 P domain and B-SAFER Section II have several overlapping risk factors and therefore I expected the P domain and Section II to have a larger association than SARA-V3 P domain and Sections I and III. Due to the lack of correspondence between the N domain and Section II, I expected a small association between these total numerical scores. Correlations between SARA-V3 and B-SAFER presence numerical scores are presented in Table 3.12 and the results are in line with my expectations. SARA-V3 presence numerical total scores for combined N and P domains correlated positively with medium to large associations with all three Sections of the B-SAFER. The N domain correlated positively with Sections I and III, but there was no association with Section II, *r* = .02, *p* > .05. Additionally, the largest association was between the SARA-V3 P domain and B-SAFER Section II. All risk factors comprising Section II have equivalent risk factors in the P domain so this large and positive association is not surprising.

Table 3.12 Concurrent Validity of SARA-V3 Presence Numerical Scores: Correlation (*r*) with B-SAFER Presence Numerical Scores

SARA-V3 Presence Numerical Scores	B-SAFER Presence Numerical Scores		
	Section I	Section II	Section III
Total (N+P)	.46***	.52***	.34***
Nature of IPV	.57***	.02	.34***
Perpetrator Risk	.21*	.69***	.21*

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; B-SAFER = Brief Spousal Assault Form for the Evaluation of Risk; Section I = Perpetrator Risk Factors, IPV History; Section II = Perpetrator Risk Factors, Psychosocial Adjustment; Section III = Victim Vulnerability Factors; IPV = intimate partner violence.
* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

Concurrent validity between the SARA-V3 and B-SAFER was also analyzed at the level of summary risk ratings and results are presented in Table 3.13. As shown in the table, both SARA-V3 and B-SAFER have the same summary risk rating categories and expectedly all summary risk ratings across the two assessments correlated positively with medium to large associations.

In general, SARA-V3 and B-SAFER total numerical scores and summary risk ratings generally demonstrated excellent concurrent validity as evidenced by large and positive correlations. B-SAFER Section II and SARA-V3 N domain had no association. When considering the risk factors comprising Section II and the N domain this result is not unanticipated as Section II matches directly on to the P domain, which had a large association. In fact, this result perhaps provided some evidence for divergent validity.

Table 3.13 Concurrent Validity of SARA-V3 Summary Risk Ratings: Correlation (*r*) with B-SAFER Summary Risk Ratings

	B-SAFER Summary Risk Ratings		
SARA-V3 Summary Risk Ratings	Case Prioritization	Life-Threatening Violence	Imminent Violence
Case Prioritization	.49***	.47***	.39***
Serious Physical Harm	.51***	.59***	.37***
Imminent Violence	.36***	.31**	.46***

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; B-SAFER = Brief Spousal Assault Form for the Evaluation of Risk; IPV = intimate partner violence.
* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

Associations Among SARA-V3 and Actuarial Risk Assessment Instruments: ODARA, DVRAG, and DA

In my examination of associations between SARA-V3 and actuarial risk assessment instruments, I expected total numerical SARA-V3 scores to have large correlations with all actuarial risk assessment instruments, but likely smaller associations than with the SARA-V2 and B-SAFER total numerical scores. Associations between SARA-V3 and the actuarial risk assessment instruments are presented in Table 3.14. The results follow the pattern that I expected. The largest association was between the SARA-V3 P domain and the DVRAG. This association is likely due to the inclusion of the PCL-R on the DVRAG, which has overlap with some of the risk factors that comprise the P domain. All correlations between SARA-V3 and the actuarial instruments were positive and most had medium to large effect sizes. The N domain resulted in lower but generally moderate correlations. The smallest correlation was between the N domain and DVRAG.

These results largely provide evidence that there was a moderate to large positive association between SARA-V3 and actuarial risk assessment instruments used to assess IPV. The associations between SARA-V3 N domain and ODARA, DVRAG, and DA were smaller overall with than the P factor domain and combined N and P domains. Overall, these results evidence construct validity of SARA-V3 based on the moderate to large positive associations with other IPV risk assessments.

Table 3.14 Concurrent Validity of SARA-V3 Presence Numerical Scores: Correlation (*r*) with ODARA, DVRAG, and DA Total Scores

SARA-V3 Presence Numerical Scores	ODARA	DVRAG	DA
Total (N+P)	.45***	.57***	.45***
Nature of IPV	.30**	.22*	.51***
Perpetrator Risk	.40***	.62***	.25*

Note. *N* = 97. SARA-V3 = Spousal Assault Risk Assessment Guide, Version 3; ODARA = Ontario Domestic Assault Risk Assessment; DVRAG = Domestic Violence Risk Appraisal Guide; DA = Danger Assessment; IPV = intimate partner violence.

* *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

Chapter 4. Discussion

Given the prevalence and consequences of IPV, both to individuals as well as society, it is essential that mental health practitioners and other service providers working with offenders and victims do the best job possible assessing and managing IPV risk. Over the last two decades it became apparent that SARA-V2 is one tool that many use around the world to aid in the assessment and management of IPV risk (Hanson, Helmus, & Bourgon, 2007). The continued prevalence of SARA-V2 among practitioners who use risk assessments as well as advancements in risk assessment literature prompted a long overdue revision of the guide. SARA-V3 incorporates new information from years of research related to IPV risk. The guide also provides additional direction and support to aid users in conducting a more complete IPV risk assessment within SPJ guidelines including risk formulation, scenario planning, and risk management planning. The purpose of this study is to provide evidence and information about the reliability and validity of SARA-V3. Following are observations regarding the analytical findings as well as limitations and future directions for research and practice.

Research Question 1 (What is the distribution of risk ratings on SARA-V3?): The findings related to the distribution of risk ratings were expected. There are a couple of factors that lack variability, but this result can likely be attributed to sampling error. Most of the offenders had significant relationship problems at the time of the assessment, but a follow-up in 2 or 5 years time would likely result in a different finding. Additionally, because I created “ever present” variables this reduced the variability across risk factors. The variability does increase slightly when comparing past versus recent ratings for most of the risk factors. Without the inclusion of the risk factors that had lower variability in this sample, some evaluators may miss important patterns of IPV over time or not consider their significance in a particular case. These risk factors also help evaluators consider problems within the recent past to help determine if violence is desisting or escalating – critical for scenario and management planning. Along that point, I would also argue that inclusion of these risk factors helps evaluators come to a more accurate summary of risk. It is often easy to let the old adage “the best predictor of future behavior is past behavior” dictate our risk decisions. These risk factors help evaluators consider patterns of behavior over time and select the most effect supervision and management strategies. In other words, inclusion of these risk factors amplify an evaluators ability to make meaning of facts and complete a more thorough risk assessment.

Research Question 2 (What is the association among ratings of risk factors on SARA-V3?): In regard to Research Question 2, results were mostly in line with the expectations for the measures of association among risk ratings. Associations among ratings of risk factors revealed all factors have unique variance and do not introduce redundancy within their respective domains. These results are in line with prior research conducted on SARA-V2 (Kropp & Hart, 2000). The findings provided justification to compute numerical total scores for domains and a total combined N and P domain score, which allowed for further and more in depth analysis of SARA-V3.

Research Question 3 (What is the interrater reliability of risk ratings made using SARA-V3?): Findings related to Research Question 3 fell in line with expectations. Across the various measures of interrater reliability, there is evidence that SARA-V3 is reliable. The numerical total scores had the highest ICC values followed by individual risk factors and then summary risk ratings. Achieving high ICCs using numerical scores, rather than qualitative categorical ratings, is more likely statistically due to greater variance across the domains and in numerical scores. This was the case for analyses in this study as total numerical SARA-V3 factors demonstrated both high rater agreement and narrow confidence intervals that generally fell within the substantial to almost perfect ranges of agreement proposed by Landis and Koch (1977). Rater agreement at the individual factor level is typically expected to be more variable (Douglas & Ogloff, 2003) and this was the case for some SARA-V3 risk factors.

There was more variability in rater agreement across summary risk ratings for Case Prioritization and Risk for Imminent Violence – ICCs were lower than expected. Lower ICC values are typically expected with summary risk ratings due to the lack of variability in ordinal data (versus interval or ratio data), however in this study these values fell in the fair range. Of course the obvious point that the raters within this study simply do not agree on what is a *Low*, *Moderate*, or *High* risk case must be considered. Indeed, some in the field are critical of this risk communication method adopted in the SPJ approach (Hilton, Carter, Harris, & Sharpe 2008; Mills & Kroner, 2006).

As this is the first research investigation of SARA-V3, there are no directly comparable studies. However, when compared to prior studies on SARA-V2, rater agreement as measured by numerical scores on SARA-V3 are analogous. For example, Hart and Kropp (2000) found rater agreement pertaining to numerical total scores on SARA-V2 to generally fall in the substantial to almost perfect ranges. Grann and Wedin (2002) also reported similarly high rater agreement for numerical total SARA-V2 scores.

The importance of the relationship between reliability and validity in investigations of psychometric analysis cannot be overstated. Although the work of validating assessments such as SARA-V3 can be tedious and time consuming – it is necessary work. The decisions made with the aid of violence risk assessments affect the lives and well being of many stakeholders – the person being assessed, the person(s) conducting the assessment, victims of violence, the criminal justice system, and society at large (Hart, Douglas, & Guy, 2016). Of utmost importance is that raters are able to agree on the factors that comprise SARA-V3 and overall ratings of risk generated by using the guide (Hart & Kropp, 2000). If this is not the case, the guide and propagated decision making model are essentially useless. The current study satisfactorily provided evidence that SARA-V3 is reliable in terms of rater agreement.

Research Question 4 (What is the association between risk ratings made using SARA-V3 and of those made using other procedures to assess risk for IPV?): The expectations for Research Question 4 were exceedingly met. SARA-V3 demonstrated concurrent validity with other measures of IPV via large, positive correlations. Discriminant validity between SARA-V3 domains, SARA-V2 Parts, and B-SAFER Sections was evidenced by large associations across corresponding assessment areas and small to no associations across differing areas. Previous studies have found similar results in terms of the large associations between SARA-V2 and other IPV risk assessments (Hilton et al., 2004; Kropp & Hart, 2000; William & Houghton, 2004). Additionally, a comparison of SARA-V3 and SARA-V2 Case Prioritization ratings resulted in a majority of cases receiving the same *Moderate* or *High* risk rating, however no low cases were assigned the same rating between the versions of the guide.

Finally, comparisons between SPJ assessments and actuarial assessments in regard to overall risk ratings show that SPJ risk judgments resulted in an overall lower risk categorization of the cases than did the actuarial assessments included in this study. Actuarial assessment instruments within the sample tended to estimate a majority of cases in the highest categories of risk (e.g., severe or extreme danger, risk bins with associated recidivism rates ≥ 59%). Given the various descriptive features of these cases, it is likely that most of these offenders are moderate risk. Many of the offenders were on bail and in the community at the time of the original risk assessment. Predictive validity studies have also demonstrated that, in general, actuarial risk assessments tend to be imprecise in terms of the assigned risk categories or bins potentially leading to an overestimation of risk (Hart, Michie, & Cooke, 2007; Mills, Jones, & Kroner, 2005; Ryan, Gray, Storey, Hart, 2015).

Implications and Future Directions

The current study was limited in several ways. While the files reviewed for this study generally included ample information about the offender and case (handwritten, original interview notes) reviewing a file and scoring an assessment tool is never a completely adequate substitute for following the recommended procedures and interviewing the individual being assessed. Indeed, practice guidelines generally dictate that face-to-face contact with the examinee is an essential part of all psychological assessments (APA, 2013). Additionally, there was generally less information associated with victims and their circumstances at the time of the assessment, which almost entirely eliminated analysis of the Victim Vulnerability domain.

This was a relatively small sample, especially the interrater subsample. It was not particularly demographically diverse and we increased rater variability by including SARA-V2 ratings that were made by a psychologist, rather than researchers in the project. This sample also comprised mostly moderate risk cases in which many of the offenders committed relatively minor crimes. This truncation of the full spectrum of risk does not allow for a full evaluation of the risk discrimination abilities of evaluators using SARA-V3. Future work should focus on more culturally diverse samples. This work is becoming incredibly important in the field. The recent Canadian Federal Court case, *Canada v. Ewert*

(2015), highlights the potential for legal challenges to psychological assessment assessment instruments due to cultural bias. Indeed, some in the field have cited this issue—cross-cultural validation and application of forensic assessment instruments—as perhaps the most challenging future problem facing the field of forensic psychology (Grisso, 2016). It is quite likely that future legal challenges will occur. Some researchers are already addressing the cross-cultural evaluation of some risk assessment instruments (Olver et al., 2016), however more work is needed.

Additionally, research with increased sample sizes and number of raters to better analyze interrater reliability should be conducted. Generalizability theory could help determine the sources of variance within rater agreement. Was the variance between TR and SC in the current study due to the rater, the cases, time when the assessment was completed, or something else? The current design cannot answer these questions and it is important to know which types of error variance are likely to affect SARA-V3 ratings. Also, research conducted with more variant samples in terms of offender severity to better examine if SARA-V3 can discriminate between low, moderate, and high risk cases should be done. In terms of priority, this work is lower in importance.

The present study did not include follow-up data, therefore an examination of predictive validity was not possible. Evaluating predictive validity with SPJ assessments is particularly difficult. The purpose of the risk assessment is to develop management strategies in order to decrease violence in the future, therefore management strategies can be thought of as a suppressor variable. Past research has outlined some of these issues (Belfrage & Strand, 2012). Future research in this area should not only focus on binary re-offense outcomes, but also on recidivism rates given the management strategies proposed as well as the *type* of recidivism when management strategies are implemented. Can we use formulation and scenario planning to forecast the type of violence a person might engage in across certain types of circumstances? Questions such as these are of primary importance for future investigation.

The raters in this study did not complete all SARA-V3 steps and so one of the highest priority areas for future research is a focus on risk formulation, scenario planning, and risk management planning. Is it possible to achieve reliability in these areas?

Considering these processes are so case, examiner, sometimes geographically and culturally specific, it is unclear if the classic standards and conceptualizations of interrater agreement are appropriate for examining these processes. Are we interested in if raters can come up with a similar risk formulation or if raters can derive risk formulations that may differ, but both be of high quality? Few researchers have studied the many issues around violence risk formulation (Minoudis et al, 2013; Sutherland et al., 2012; Wilson, 2013) and this area is a top priority moving forward.

References

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist*, *68*, 7-19.

Au, A., Cheung, G., Kropp, R., Yuk-chung, C., Lam, G. L. T., & Sung, P. (2008). A preliminary validation of the Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER) in Hong Kong. *Journal of Family Violence, 23,* 727-735.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist*, *34*, 341-382.

Babcock, J. C., Green, C. E., & Robie, C. (2004). Does batterers' treatment work? A meta-analytic review of domestic violence treatment. *Clinical Psychology Review*, *23*, 1023-1053.

Belfrage, H., & Strand, S. (2012). Measuring the outcome of structured spousal violence risk assessments using the B-SAFER: Risk in relation to recidivism and intervention. *Behavioral Sciences & the Law*, *30*, 420-430.

Belfrage, H., Strand, S., Storey, J. E., Gibas, A. L., Kropp, P. R., & Hart, S. D. (2012). Assessment and management of risk for intimate partner violence by police officers using the Spousal Assault Risk Assessment Guide. *Law and Human Behavior*, *36*, 60-67.

Bowlus, A., McKenna, K., Day, T., & Wright, D. (2003). *The economic costs and consequences of child abuse.* Report prepared by The Law Commission of Canada, University of Western Ontario, Canada.

Campbell, J. C. (1986). Nursing assessment of risk of homicide for battered women. *Advances in Nursing Science, 8*, 36-51.

Campbell, J. C., O'Sullivan, C., Roehl, J., & Webster, D. W. (2005). *Intimate partner violence risk assessment validation study: The RAVE study.* Final Report to the National Institute of Justice (NCJ 209731–209732). Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/209731.pdf

Campbell, J. C., Webster, D. W., Koziol-McLain, J., Block, C., Campbell, D., Curry, M. A., … Laughon, K. (2003). Risk factors for femicide in abusive relationships: Results from a multisite case control study. *American Journal of Public Health, 93,* 1089-1097.

Campbell, J .C. (1995). *Assessing dangerousness*. Newbury Park, CA: Sage Publishing.

Canadian Centre for Justice Statistics. (2015). *Family violence in Canada: A statistical profile, 2013* (Statistics Canada – Catalogue no. 85-002-X). Ottawa, ON: Minster of Industry.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86,* 127-137.

Cocozza, J. J., & Steadman, H. J. (1978). Prediction in psychiatry: An example of misplaced confidence in experts. *Social Problems, 25,* 265-276.

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

Coker, A. L., Davis, K. E., Arias, I., Desai, S., Sanderson, M., Brandt, H. M., & Smith, P. H. (2002). Physical and mental health effects of intimate partner violence for men and women. *American Journal of Preventive Medicine, 24*, 260-268.

Day, T. (1995). *The health related costs of violence against women: The tip of the iceberg*. Report prepared by the Centre for Research on Violence Against Women and Children, University of Western Ontario, Canada.

Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence risk assessment: Science and practice. *Legal and Criminological Psychology*, *4*, 149-184.

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

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20^{V3}: Assessing risk for violence – User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, Version 3 (HCR-20^{V3}): Development and overview. *International Journal of Forensic Mental Health*, *13*, 93-108.

Douglas, K. S., Hart, S. D., Groscup, J. L., & Litwack, T. R. (2013). Assessing violence risk. In I. B. Weiner & R. K. Otto (Eds.), *The handbook of forensic psychology*, (4th ed.) (pp.385-442). New York, NY: Wiley.

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

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

Dvoskin, J. A., & Heilbrun, K. (2001). Risk assessment and release decision-making: Toward resolving the great debate. *Journal of the American Academy of Psychiatry and the Law, 29*, 6-10.

Ellsberg, M., Jansen, H. A. F. M, Heise, L., Watts, C. H., & Garcia-Moreno, C. (2008). Intimate partner violence and women's physical and mental health in the WHO multi-country study on women's health and domestic violence: An observational study. *Lancet*, *371*, 1165-1172.

Ewert v. Canada, FC 1093 (2015).

Feder, L., & Wilson, D. B. (2005). A meta-analytic review of court-mandated batterer intervention programs: Can courts affect abusers' behavior? *Journal of Experimental Criminology*, *1*, 239-262.

Fleiss, J. L., & Shrout, P. E. (1977). The effects of measurement errors on some multivariate procedures. *American Journal of Public Health*, *67*, 1188-1191.

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

Golding, J. M. (1999). Intimate partner violence as a risk factor for mental disorders: A meta-analysis. *Journal of Family Violence*, *14*, 99-132.

Goodman, L. A., Dutton, M. A., & Bennett, L. (2000). Predicting repeat abuse among arrested batterers use of the Danger Assessment Scale in the criminal justice system. *Journal of Interpersonal Violence*, *15*, 63-74.

Grann, M., & Wedin, I. (2002). Risk factors for recidivism among spousal assault and spousal homicide offenders. *Psychology, Crime and Law*, *8*, 5-23.

Grisso, T. (2016, March). Finding your way without GPS. In S. L. Broksky *The 2-way psych-law time machine: Looking at 1968 from 2016, and 2016 from 2064.* Symposium conducted at the annual meeting of the American Psychology-Law Society, Atlanta, GA.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19-30.

Hanson, R. K., Helmus, L., & Bourgon, G. (2007). The validity of risk assessments for intimate partner violence: A meta-analysis (User Report No. 2007–07). Ottawa, ON: Public Safety Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist – Revised manual.* Toronto, Canada: Multi-Health Systems Inc.

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

Harris, G. T., Rice, M. E., Quinsey, V. L., & Cormier, C. A. (2015). The actuarial prediction of violence. In *Violent offenders: Appraising and managing risk (*3rd ed.*)* (pp. 121-168). Washington, DC: American Psychological Association.

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, *3*, 121-137.

Hart, S. D., Douglas, K. S., & Guy, L. S. (2016). The structured professional judgment approach to violence risk assessment: Origins, nature, and advances. In D. P. Boer (Ed.), *The Wiley Handbook on the Theories, Assessment and Treatment of Sexual Offending* (pp. 643-666). New York, NY: Wiley.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments. *British Journal of Psychiatry*, *190*, 60-65.

Heckert, D. A., & Gondolf, E. W. (2004). Battered women's perceptions of risk versus risk factors and instruments in predicting repeat reassault. *Journal of Interpersonal Violence*, *19*, 778-800.

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: the importance of legal decision-making context. *Law and Human Behavior*, *21*(4), 347-359.

Helmus, L., & Bourgon, G. (2011). Taking stock of 15 years of research on the Spousal Assault Risk Assessment Guide (SARA): A critical review. *International Journal of Forensic Mental Health 10*, 64-75.

Hilton, N. Z., Carter, A. M., Harris, G. T., & Sharpe, A. J. (2008). Does using nonnumerical terms to describe risk aid violence risk communication? Clinician agreement and decision making. *Journal of Interpersonal Violence*, *23*, 171-188.

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario Domestic Assault Risk Assessment. *Psychological Assessment*, *16*, 267-275.

Hilton, N. Z., Harris, G. T., Rice, M. E., Houghton, R. E., & Eke, A. W. (2008). An indepth actuarial assessment for wife assault recidivism: The Domestic Violence Risk Appraisal Guide. *Law and Human Behavior*, *32*, 150-163.

Klein, A. R. (2009). *Practical implications of current domestic violence research: For law enforcement, prosecutors and judges.* (Report No. NCJ 225722). Washington, DC: Department of Justice Office of Justice Programs.

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.

Kropp, P. R., Hart, S. D., & Belfrage, H. (2005). *Brief spousal assault form for the evaluation of risk (B-SAFER): User manual.* Vancouver, Canada: ProActive ReSolutions, Inc.

Kropp, P. R., Hart, S. D., & Belfrage, H. (2010). *Brief spousal assault form for the evaluation of risk (B-SAFER), Version 2: User manual*. Vancouver, Canada: ProActive ReSolutions Inc.

Kropp, P. R., Hart, S. D., Webster, C. W., & Eaves, D. (1994). *Manual for the Spousal Assault Risk Assessment Guide.* Vancouver, Canada: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. W., & Eaves, D. (1995). *Manual for the Spousal Assault Risk Assessment Guide,* 2nd ed*.* Vancouver, Canada: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. W., & Eaves, D. (1999). *Spousal Assault Risk Assessment: User's Guide.* Toronto: Multi-Health Systems Inc.

Kropp, P. R., & Hart, S. D. (2015). The Spousal Assault Risk Assessment Guide Version 3 (SARA-V3). Vancouver, Canada: ProActive ReSolutions Inc.

Kropp, P. R., Hart, S. D., & Lyon, D. R. (2007). Stalking Assessment and Management. Vancouver, Canada: ProActive ReSolutions Inc.

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

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, *7*, 409-443.

Litwack, T. R., Zapf, P. A., Groscup, J. L., & Hart, S. D. (2006). Violence risk assessment: Research, legal, and clinical considerations. In I. B. Weiner and A. K. Hess (Eds.), *The handbook of forensic psychology,* (3rd ed.) (pp. 487-533). Hoboken, NJ: Wiley.

Max, W., Rice, D. P., Finkelstein, E., Bardwell, R. A., & Leadbetter, S. (2004). The economic toll of intimate partner violence against women in the United States. *Violence and Victims*, *19*, 259-272.

Meehl, P. E. (1954). *Clinical vs. statistical predictions: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Messing, J. T., & Thaller, J. (2012). The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence*, *28,* 1537-1558.

Mills, J. F., Jones, M. N., & Kroner, D. G. (2005). An examination of the generalizability of the LSI-R and VRAG probability bins. *Criminal Justice and Behavior*, *32*, 565-585.

Mills, J. F., & Kroner, D. G. (2006). The effect of base-rate information on the perception of risk for reoffense. *American Journal of Forensic Psychology*, *24*, 45-56.

Minoudis, P., Craissati, J., Shaw, J., McMurran, M., Freestone, M., Chuan, S. J., & Leonard, A. (2013). An evaluation of case formulation training and consultation with probation officers. *Criminal Behaviour and Mental Health*, *23*, 252-262.

Olver, M. E., Sowden, J. N., Kingston, D. A., Nicholaichuk, T. P., Gordon, A., Christofferson, S. M. B., & Wong, S. C. P. (2016). Predictive accuracy of Violence Risk Scale–Sexual Offender Version Risk and change scores in treated Canadian Aboriginal and Non-Aboriginal sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*. Advance online publication. doi: 10.1177/1079063216649594

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

Rettenberger, M., & Eher, R. (2013). Actuarial risk assessment in sexually motivated intimate-partner violence. *Law and Human Behavior*, *37*, 75-86.

Ryan T. J., Gray, A. L., Storey, J. E., & Hart, S. D. (2016, March). *Cross-validation of the VRAG: A 10-year prospective study.* Paper presented at the Annual Meeting of the American Psychology-Law Society, San Diego, CA.

Sartin, R. M., Hansen, D. J., & Huss, M. T. (2006). Domestic violence treatment response and recidivism: A review and implications for the study of family violence. *Aggression and Violent Behavior*, *11*, 425-440.

Slaney, K. L., Storey, J. E., & Barnes, J. (2011). Is my test valid? Guidelines for the practicing psychologist for evaluating the psychometric properties of measures. *International Journal of Forensic Mental Health*, *10*, 261-283.

Storey, J. E., Kropp, P. R., Hart, S. D., Belfrage, H., & Strand, S. (2014). Assessment and management of risk for intimate partner violence by police officers using the Brief Spousal Assault Form for the Evaluation of Risk. *Criminal Justice and Behavior*, *41*, 256-271.

Sutherland, A. A., Johnstron, L., Davidson, K. M., Hart, S. D., Cooke, D. J., Kropp, P. R., Logan, C., Michie, C., & Stocks, R. (2012). Sexual violence risk assessment: An investigation of the interrater reliability of professional judgments made using the Risk for Sexual Violence Protocol. *International Journal of Forensic Mental Health, 11*, 119-133.

Truman, J. L., & Langton, L. (2015). *Criminal victimization, 2014.* (Report No. NCJ 248973). Washington DC, Bureau of Justice Statistics.

United Nations Department of Public Information. (2009). *Violence against women.* (Report No. DPI/2546A). New York: Author.

Williams, K. R., & Houghton, A. B. (2004). Assessing the risk of domestic violence reoffending: A validation study. *Law and Human Behavior*, *28*, 437-455.

Wilson, C. M. (2013). Reliability and consistency of risk formulations in assessments of sexual violence risk (Doctoral Thesis). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 38379)

World Health Organization. (2012). *Understanding and addressing violence against women.* (World Health Organization – WHO/RHR/12.36). Retrieved from: http://apps.who.int/iris/bitstream/10665/77432/1/WHO_RHR_12.36_eng.pdf.

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136,* 740-767.

Zhang, T., Hoddenbagh, J., McDonald, S., & Scrim, K. (2012). *An estimation of the economic impact of spousal violence in Canada, 2009.* (Report No. rr12-07-e). Ottawa: Department of Justice Canada.