An Evaluation of the Reliability and Quality of Expert and Novice Forensic Case Formulations

by

Tara J. Ryan

M.A., Simon Fraser University, 2017

B.Sc., Creighton University, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Summer 2020

Approval

Name:	**Tara J. Ryan**
Degree:	**Doctor of Philosophy (Psychology)**
Title:	**An Evaluation of the Reliability and Quality of Expert and Novice Forensic Case Formulations**
Examining Committee:	**Chair:** Ralph Mistlberger Professor
	Stephen Hart Senior Supervisor Professor
	Kevin Douglas Supervisor Professor
	P. Randall Kropp Supervisor Adjunct Professor
	Maaike Helmus Internal Examiner Assistant Professor School of Criminology
	Rajan Darjee External Examiner Senior Lecturer Centre for Forensic Behavioural Science Swinburne University of Technology
Date Defended/Approved:	August 13, 2020

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

Forensic case formulation is an under-studied and growing area within the violence risk assessment literature. The current study aimed to address gaps in the literature by examining the interrater reliability (IRR) and quality of forensic case formulations by comparing Expert and Novice raters. *N* = 50 intimate partner violence offender files were accessed. Four raters (*n* = 2 Experts, *n* = 2 Novices) rated each file using all steps of Spousal Assault Risk Assessment Guide-Version 3 (SARA-V3; Kropp & Hart, 2015). Cases were formulated using a Decision Theory approach in which motivating, disinhibiting, and destabilizing mechanisms were identified. The distribution of ratings for these mechanisms was presented. IRR was examined using a novel coefficient, Gwet's AC. Raters also completed narrative case formulations. Then a *Within Case* and *Across Case* paired case design involving *n* = 143 narrative formulation pairs was conducted with three new raters. The similarity of paired formulations was evaluated. Raters also assessed the quality of formulations using the Case Formulation Quality Checklist-Revised (CFQC-R; McMurran & Bruford, 2016). For most formulation mechanisms, distribution of Presence ratings was skewed. Overall, across Experts and Novices, the IRR of formulation mechanisms ranged from *poor* to *almost perfect* (AC_2 = .10. - .98), with most coefficients falling between the *moderate* and *almost perfect* ranges. The similarity of formulations was established; *Within Case* paired formulations were judged as more similar than *Across Case* paired formulations. Finally, formulations were high in quality; Experts produced higher quality formulations than Novices.

Keywords: forensic case formulation; SARA-V3; Spousal Assault Risk Assessment; violence risk assessment; formulation interrater reliability; Structured Professional Judgment

Acknowledgements

Seven years ago, I picked up my life in Omaha, Nebraska and moved to Canada to attend SFU and work with Dr. Stephen Hart. This is still somewhat surreal to me. Steve's mentorship, kindness, and pragmatism have been invaluable during my graduate training. I strive to think critically about the work I do in our field and to be especially aware of its implications, which is often on individuals who are some of the most disadvantaged in our society. This outlook is not always prevalent among psychologists in the forensic field. I know that my views in this regard have been profoundly impacted by Steve and my training at SFU. I could not be more appreciative to have had your guidance throughout graduate school—thank you, Steve.

I would also like to thank my committee members, Drs. Randy Kropp and Kevin Douglas, for their support as I completed my training, thesis, and dissertation work. I am grateful to have their perspectives as I move forward in my career. I am also thankful to Drs. Rajan Darjee and Maaike Helmus for their willingness to examine my dissertation during this unprecedented time.

My life would look very different had I not enrolled in Dr. Matthew Huss' Introductory Psychology course in 2007 at Creighton University. Dr. Huss has been a stable, reliable, and ever-encouraging mentor since that time. I cannot thank him enough for being a frequent voice of reason and support as I have navigated life before and during grad school. He is one of the hardest working forensic psychologists I know and is unmatched in the ways in which he gives back to this field and his students.

Dr. Karen Whittemore provided me with invaluable clinical (and let's be honest, life!) supervision during my graduate training. Karen is uniquely dedicated to training students. Most of what I know about applied violence risk assessment is because of the work I completed with her—I know that many of her former students share similar sentiments. Thanks for the many hours of supervision, lunchtime jogs, and the occasional beer and hike, Karen! I would also like to thank the Surrey Forensic Clinic team for being great friends of research as well as great colleagues during my time there. Special thanks to Lisa Casagrande for becoming such a wonderful and inspiring friend. Dr. Eugene Wang and Bobbei kept me well fed during my final training years. Thank you for your kindness, friendship, and helpful feedback on an early draft of my dissertation.

Ellen Kurz (aka my Vancouver mom) and her family have been so kind as I worked my way through grad school. Thank you for the laughter, many chats, meals, and Karl's whiskey! :P

Life at SFU and in Vancouver would have been much less fulfilled without the friendship of Dylan and Cady Gatner. Dylan, your peer mentorship and camaraderie as we tackled courses, clinical work, and research were invaluable to me. The Gatners really made me part of their lives in Vancouver, so much so that I had a sense of “home” in this once very foreign city. Thanks to you both. I cannot wait to visit you soon!

I am not sure how I would have managed to get through my grad training without Sarah Coupland. From reminding me when and where I had to be during our first years at SFU to many validating chats in between, Sarah has become one of my closest friends. She has read cover letters, applications, and offered encouragement over the past seven years. She also made both my MA and PhD research possible by collecting data despite having a lot on her own plate. I look forward to a lot more research, conferences, and getaways in the future. And I promise to always fill up the gas tank before we leave Hope. Thank you, Sarah!

Erin Fuller has become a great friend during my time at SFU. It’s rare to find someone with her candour and authenticity. I laugh often when I am with Erin, which was crucial to getting through grad school. Kate Hanniball, I strive to achieve the balance and passion of both work and sport that you have created in your life. I am so glad our paths crossed at SFU and I look forward to some outdoor adventures with you. I can always count on Leila Wallach to be my gastronome co-pilot; let’s work on that cloning plan for Chispa. Dave Schuberth, I can’t believe I’m moving to your beloved Woo! Looking forward to your future beer projects (and being a taste tester). Dylan Wiwad, thanks for all the bike advice! Here’s to hoping for some RIP Shift Drinks in 2021, Lee Vargen. Nicole Muir, Jenny Pink, Kat O’Donnell, Yan Lim, and Tiff O’Connor–thanks for being such wonderful supports throughout these last seven years. I am so excited to see where life takes everyone.

Research is hard work and this study felt especially arduous at times. I want to sincerely thank the research assistants who worked on this project. I would also like to thank my funders, Protect International and Mitacs Accelerate. This monetary support made the project possible.

To my close friends back home–Lee, Kelly, Erin, Corrie, Dani, Casey, Emily, and Alycia–I have missed many of your life events, big and small, during my grad training. I am excited to be more present and for many more trips home and away with all of you.

Finally, my perseverance and resilience are both attributable to my mom. She offers the definition of unconditional love and support and this accomplishment would never have been possible without her. Thank you, mom. I love you.

Table of Contents

List of Tables

List of Figure

Chapter 1. Introduction

Rates of general violence continue to fall in North America (Federal Bureau of Investigation, 2017; Statistics Canada, 2017). Similarly, the rate of intimate partner violence (IPV) has been declining since the 1990s in both the United States and Canada (Catalano, 2015; Sinha, 2013). This decrease has been attributed to a number of demographic and dynamic factors as well as changes in societal perspectives. For example, Farmer and Tiefenthaler (2003) found the greater independence of women due to educational attainment and resulting financial security, and the increased access to legal services specific to victims of IPV, have impacted rates. Others have also attributed the decline in part to modernized policing practices such as pro-arrest policies (Maxwell et al., 2002; Sherman & Berk, 1984; Sugarman & Boney-McCoy, 2000) and shifting attitudes and expectations around divorce, marriage, and the availability of women's services (Dugan et al., 1999).

Despite these optimistic trends over the last three decades, IPV continues to pose a serious societal problem. In Canada between 2005 and 2011, nearly 60% of completed adult court cases were categorized as involving IPV (Beaupré, 2015). Further, of those cases, nearly 60% resulted in a guilty verdict on at least one charge. A majority of those who pled or were found guilty received a sentence requiring supervision. These sentences resulted in obvious monetary and resource burdens on the criminal justice system and ultimately taxpayers. Further, around the world, the frequency of IPV is variable, with women in some countries reporting very high lifetime prevalence (Alhabib et al., 2010; Garcia-Moreno et al., 2006). Notably, since the onset of the COVID-19 pandemic in January 2020, a number of organizations and researchers have raised alarms about the potential increased risk of IPV given financial strain within family units and social distancing requiring sheltering in our homes (Campbell, 2020; van Gelder et al., 2020). Bottom line: IPV is the most common form of violence experienced by women worldwide and efforts to reduce its occurrence continue to be imperative.

Changing social factors and policing practices have certainly impacted rates of IPV. Another important component to the reduction of IPV is empirically informed violence risk assessment decision support tools. Empirical research has demonstrated that advances in risk assessment practices have benefitted the assessment and management

of IPV offenders compared to the use of unstructured clinical judgment (Singh & Fazel, 2010; Singh, Grann, & Fazel, 2011; Yang et al., 2010; see also Fazel et al., 2012, with respect to actuarial instruments focused on historical factors and unstructured clinical judgment). IPV, also known as spousal assault or domestic violence, is "the actual, attempted, or threatened physical harm of a current or former intimate partner" (Kropp & Hart, 2015, p. 1). Since the mid-1990s, the research and development of two dominant violence risk assessment approaches have transformed how many psychologists, psychiatrists, law enforcement officers, probation officers, and threat assessment teams think about and conduct assessments of violence risk. The emergence of the actuarial and Structured Professional Judgment (SPJ) approaches to risk assessment have led to considerable debate about the comparative clinical utility, validity, IRR, and predictive accuracy of the two approaches.

The actuarial method for assessing violence risk differs from the SPJ approach in three predominant ways: development, scoring procedures, and risk conclusions. In regard to development, the item composition of actuarial risk assessment instruments is typically derived from a sample of offenders in which regression analyses determine the factors most predictive of recidivism (Harris et al., 2015). Items are rated and then combined according to an algorithm to create a total score that corresponds to a specific risk probability associated with recidivism in the development sample. In contrast, the risk factors that comprise SPJ guidelines are derived from systematic review of the literature and consultation with subject matter experts. Administration procedures also differ. Typically, after gathering information and rating the presence of risk factors, users determine the relevance factors using two types of case formulation. The first type of formulation focuses on the development of an explanation of past violence. The second type of formulation focuses on the development of plausible future scenarios and management strategies for those scenarios. Finally, raters consider all information from the assessment process, including how reasonably the risk management strategies can be implemented, and make conclusory opinions using a 3-point ordinal scale [*low*, *moderate*, *high]* in relation to different dimensions of risk – likelihood, serious harm, and imminence. A myriad of book chapters, journal articles, and conferences presentations have been published and presented detailing the minutiae of the ongoing debate between the utility of the two approaches (see Dvoskin & Heilbrun, 2001; Hart, 1998; Hart et al., 2016; Hart et al., 2007; Heilbrun, 1997; Quinsey et al., 1998; Yang et al., 2010). Notably,

one of the common findings in various meta-analyses comparing the predictive validity of actuarial risk assessment instruments and SPJ guidelines is that there is no substantial difference between them (Williams et al., 2017). Overall, it seems all of these tools are, at best, moderately predictive of future violent behaviour. This, of course, leads some to question why there is such disagreement in the field regarding which of the two approaches is “best” and why one might choose to use a particular tool over another.

One reason some clinicians prefer one approach over the other centers around the potential problem of using predictive validity as the primary psychometric criterion for selecting a tool. Given the difficulty in predicting future human behaviour, especially low base rate behaviours such as violence, it is possible that we, as a field, will not drastically improve our predictive accuracy. Thus, contemporary violence risk assessment tools are more focused on the identification of dynamic risk factors and reducing risk through addressing those risk factors via rehabilitation and management strategies (Williams et al., 2017). Inherent to the process of conducting an SPJ risk assessment is the formulation of violence risk and consideration of management planning (i.e., monitoring, treatment, supervision, victim safety planning strategies) including attention to resources available, the offender’s response to past treatment and monitoring, the offender’s current interest and insight regarding need for treatment, and attitudes toward supervision conditions. The presence and relevance of risk factors *as well as* consideration of appropriate management strategies help inform conclusory opinions around risk. Ultimately, the goal of an SPJ violence risk assessment is not to predict risk—the goal is to develop a risk management plan. This plan should inform the judge or decisionmaker about the reasonableness that the individual can be managed in a manner that least impacts personal liberties and freedom while protecting both the individual and the public from violence. Singh et al. (2014) conducted a survey of professionals who use violence risk assessment tools and identified the frequency of use and perceived utility of various actuarial and SPJ violence risk assessment instruments and guidelines. Results indicated that those who used SPJ guidelines rated them as very useful with regard to developing risk management plans. Given that management strategy planning is not inherent to the process when conducting an actuarial risk assessment, professionals who are interested in a more comprehensive process of assessing violence risk may opt to use SPJ guidelines.

IPV Risk Assessment within the SPJ Approach

The SARA, the first SPJ guideline, was developed for the assessment of risk for IPV. It is one of the most widely used IPV risk assessment tools in the field (Hanson et al., 2007). Version one of the guide was developed and published in 1994 and version two was released in 1995. Minor updates to Version 2 were completed in 1999 and 2008 (SARA-V2; Kropp et al., 2008). Following the standard SPJ development procedures, SARA risk factors were derived through an extensive literature review. SARA-V2 comprises 20 risk factors thematically organized into Part 1 (risk factors associated with criminal history and psychosocial adjustment) and Part 2 (risk factors associated with spousal assault history and the index offence). Administration of SARA-V2 involves the typical process of information gathering and review. Users then rate the presence of risk factors. Next, raters determine the relevance of "critical factors" that, on their own, indicate the examinee poses an imminent risk of harm. Finally, raters make two Conclusory Opinions, imminent risk of harm to spouse and imminent risk of harm to another person. Research has indicated that risk ratings made using the SARA-V2 have good to excellent IRR, good to excellent concurrent validity with other IPV risk assessment tools, and moderate to good predictive validity (Belfrage et al., 2012; Graham et al., 2019; Grann & Wedin, 2002; Hanson et al., 2007; Heckert & Gondolf, 2004; Helmus & Bourgon, 2011; Hilton et al., 2004; Hilton et al., 2020; Kropp & Hart, 2000; Messing & Thaller, 2013; Nicholls et al., 2013; van der Put et al., 2019; Williams & Houghton, 2004).

The SARA-V2 was recently revised to better reflect advances in empirical research and SPJ assessment procedures. The SARA-Version 3 (SARA-V3; Kropp & Hart, 2015) comprises three domains: Nature of IPV Factors (N), 8 factors; Perpetrator Risk Factors (P), 10 risk factors; and Victim Vulnerability Factors (V), 6 factors. The domains and factors are presented in Table 1. The administration procedure includes the standard 6 steps: (1) Gathering case information, (2) Rating the presence of risk factors, (3) Rating the relevance of risk factors based on a formulation of violence risk, (4) Scenario planning, (5) Developing risk management plans, (6) Providing Conclusory Opinions.

Given the recent publication of SARA-V3, there is currently a lack of research on the guidelines. Only two unpublished dissertations have been conducted on SARA-V3. In the first study, Ryan (2016) reported fair to good IRR for presence and relevance ratings of individual risk factors as well as Conclusory Opinions. She also reported moderate to

high concurrent validity with respect to ratings made using other actuarial instruments and SPJ guidelines. However, Ryan did not evaluate the formulations of violence risk or scenario-based management plans in more detail. One other recent unpublished dissertation, (Schafers, 2019), reported there were significant decreases in SARA-V3 risk ratings made pre- versus post- treatment, and also found that those decreases were associated with lower post-release recidivism.

Table 1 SARA-V3 Factors

Nature of IPV: History includes…
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: Problems with…
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. Violence/Suicidal Ideation
P10. Distorted Thinking About IPV
Victim Vulnerability Factors: Problems with…
V1. Barriers to Security
V2. Barriers to Independence
V3. Interpersonal Resources
V4. Community Resources
V5. Attitudes or Behaviour
V6. Mental Health

One problem with empirical investigations of SPJ guidelines is that most of them did not study ratings of the presence and relevance of individual risk factors, but rather recoded them numerically and summed them to yield composite scores. This is antithetical to the administration procedures within the SPJ framework and thus ultimately an

inaccurate way of testing the predictive validity of SPJ guidelines (Hart & Boer, 2010). In addition, given that inherent to the process of conducting an SPJ violence risk assessment is the selection and recommended implementation of management strategies, if actually implemented, the recidivism rate *should* be lower (Hart et al., 2016). Thus, research more analogous to SPJ guidelines would include an examination of the entire SPJ risk assessment process, including case formulation, scenario planning, risk management planning, the predictive validity of the case formulation and scenario planning, as well as evaluation of which management strategies were implemented and if and how those strategies impacted recidivism. This type of research is incredibly difficult and resource intensive.

Case Formulation – Theory and Process

Case formulation is the process of integrating various and often extensive information about a client into a coherent and explanatory narrative. As typically described in clinical psychology and psychiatry, clinical case formulation focuses on the diagnosis or treatment of an individual. When formulating, a clinician takes into account all sources of information about the client including, but not limited to: presenting problem(s), psychosocial history, past and current psychosocial functioning, interpersonal relationships, culture, childhood trauma, and genetic and medical history (Eells, 2014). Eells notes that case formulation changes over time and aids clinically in assigning diagnoses, understanding conditioned behavioural responses, treatment planning, and making sense of client information, especially when there are contradictions in a client's behaviours, emotions, and thoughts. Importantly, formulations are often shared with clients to increase their understanding of their problems as well as how treatment may be structured and of benefit. Because of the integral role case formulation plays in clinical practice, psychologists and psychiatrists view formulation skills as necessary for competent practice. To that end, various licensing organizations have specified practice standards regarding formulation skills including the American Psychological Association (American Psychological Association, 2006), American Board of Professional Psychology (American Board of Professional Psychology, n.d.), the Canadian Psychological Association (Canadian Psychological Association, 2012), and the British Psychological Society (British Psychological Society, 2017).

According to Persons and Tompkins (1997), the quality of clinical case formulations should be evaluated across a number of areas—a good case formulation has treatment utility, is parsimonious in that it offers the minimum information necessary to guide treatment, and is evidence based. They also suggest clinical case formulations should use nomothetic formulations (general laws of behaviour) as a foundation for idiographic explanations about a particular case. Eells and Lombart (2011) suggest that formulations should also be testable and refined if hypotheses are not supported in the client's behavioural or emotive experience and that formulations should be accurate with respect to the client factors and balanced between description and explanation. Eells and Lombart propose formulations serve a number of purposes including aiding in the organization of information about a person, providing a blueprint for treatment, acting as a means to measure change, and helping the therapist better understand the patient. Persons and Tompkins (1997) outline a 7-step process for clinical case formulations that involves identifying presenting problems, providing a DSM-5 (American Psychiatric Association, 2013) diagnosis, selecting an 'anchoring diagnosis,' selecting a nomothetic formulation if one exists, individualizing the template by collecting idiographic data, proposing hypotheses about the origins of the mechanisms, and describing the precipitants of the current episode of illness or symptom exacerbation.

Cases can be formulated in any number of ways and using various psychological theoretical frameworks posing challenges for empirical evaluation. Ridley et al. (2017) discuss a number of challenges facing clinicians and other providers using case formulation including the lack of a consensus definition, variety of methods and orientations from which to formulate, and judgment or inferential errors that clinicians fall into when formulating—all of which contribute to case mis-conceptualization. Despite this, one recent systematic review examining the IRR of clinical case formulations highlighted some promising results (Flinn et al., 2015). The review included a total of N = 18 articles in the analyses across different therapeutic orientations, including cognitive (n = 6), behavioural (n = 1), psychodynamic (n = 6), and integrative (n = 5). The IRR of formulations across the modalities was mixed and ranged from slight to substantial agreement. Formulations grounded in the psychodynamic framework produced the highest levels of IRR. Flinn and colleagues concluded that IRR can be achieved across all orientations, but indicated that training on the skills required for case formulation led to higher rates of agreement in the studies included in the review.

Still other issues arise when considering the empirical evaluation of case formulations. For example, if a case is formulated from a psychodynamic perspective versus a humanistic perspective—which is correct? And what is *correct*, anyway? Butler (1998) postulates that there is no correct formulation as formulations are merely hypotheses and not facts. So then how do we, as a field, evaluate case formulations? Butler suggests one criterion to consider is usefulness to clinical practice (e.g., Does the case formulation lead to therapeutic change and ultimately patient improvement? Is it useful to the client in understanding their problems? Does it help the therapist understand the client's problems?). This raises an important issue with regard to the evaluation of IRR of case formulations as it is entirely possible that two clinicians may arrive at different formulations about the same client and although IRR may be lacking, both can still be putting forth plausible and high-quality hypotheses about the client. So, given the inherently dynamic nature of the task that is case formulation, it is necessary to advance research that taps into and simultaneously tests various aspects of case formulations including IRR, validity (criterion and predictive), quality, and similarity. Indeed, Sturmey and McMurran (2011) underscore that no studies have been conducted that test if, within one research design, clinicians can agree on a clients' target behaviours and the mechanisms that influence those behaviours, and then match treatment plans for those behaviours.

Case Formulation in Forensic Psychological Practice

Case formulation skills are similarly integral to forensic psychological and psychiatric practice (Sturmey & McMurran, 2011). Particularly within the area of violence risk assessment, forensic practitioners have made use of case formulation skills to understand and manage the perpetration of violence. As the focus on the importance of formulation in psychotherapy has increased in recent years (Eells, 2007), so has the emphasis of formulation within forensic practice and forensic case formulation is indeed considered by some to be the next advancement in violence risk assessment research and practice (Hart & Logan, 2011). Just as within the general clinical context, forensic case formulations can be conceptualized within the framework of any psychological theory (e.g. psychodynamic, cognitive-behavioural, humanistic, integrative, etc.). An interesting complication of forensic case formulation in the context of violence risk assessment is that the goals may be twofold: to understand violence risk *and* to develop risk management plans. But most people focus on etiology. For example, Delle-Vergini and Day (2016),

"The formulation provides a structure to develop answers to questions such as why the person offends and which are the most relevant factors that maintain offending" (p. 241). Although understanding the etiology of violence is integral to the formulation, it should be for the purpose of implementing strategies to manage violence. Douglas et al. (2013) note, "Ideally, we need to tell a story about an individual that integrates the many pieces of information available to us. It is necessary to derive an individual theory of risk, to help us make sense of risk, and therefore how to best intervene and manage risk" (p. 54) (see also Banasik et al. 2018). A handful of formulation approaches specific to offending behaviour have been developed and are used among practicing clinicians, however the SPJ method of formulating was developed specifically to understand violence perpetration. These approaches provide clinicians with a process to organize the often immense data points that are collected on clients in a more systematic and theory-driven method.

Approaches to Case Formulation: Etiology of Violence Risk

Risk-Needs-Responsivity (RNR) (Andrews et al., 1990). The RNR framework has been transformative in the way that clinicians, researchers, and some correctional systems approach their work with offenders. Indeed, the RNR model has, to varying degrees, impacted the development of the remaining case formulation approaches outlined below. Culminating in a decade of writing and research in the field related to classifying offenders based on risk and needs, Andrews et al. proposed a model of reducing recidivism via matching (the responsivity principle) individual offenders to a level of service analogous to their unique risk to recidivate and to their criminogenic needs, incorporating both general and specific approaches to treatment and management. Development of the RNR set of principles has resulted in further application to focus on eight risk/need factors associated with criminal conduct including antisocial personality, antisocial attitudes, antisocial associates, substance abuse, family/marital relationships, school/work, and prosocial recreational activities (Bonta & Andrews, 2007). Other non-criminogenic (minor) needs include self-esteem, vague feelings of personal distress, major mental disorder, and physical health. Formulation using the RNR framework first involves assessing the risk of recidivism and then matching the services provided to the risk. Given that research and theory have established that higher risk offenders recidivate at a higher rate than lower risk offenders, the level of interventions (e.g., dosage, setting, etc.) should be more intensive with higher risk offenders. In order to properly provide interventions, the

offender's criminogenic needs must be identified and targeted in treatment (Bonta & Andrews). Bonta and Andrews provide many recommendations regarding the responsivity principle and the delivery of services in the spirit of the RNR framework including the use of cognitive social learning methods (cognitive-behavioural interventions) and tailored application of such interventions considering individual differences (e.g., intellectual disability). After development of the model, Andrews and Bonta created the LSI and later revisions to aid practitioners in the application of the model.

One limitation of the RNR framework is that it was not developed specifically for forensic case formulation, but rather as a comprehensive model of offender rehabilitation. Critics of the model suggest it is not theoretically robust enough to provide therapists with adequate tools to help offenders make substantive changes that will impact offending (Ward & Brown, 2004; Ward & Stewart, 2003). The other critiques offered by Ward and colleagues tend to focus on what could be considered factors that affect the quality and necessary components of a violence risk formulation (e.g., lacks individualization, focus on criminogenic needs is too narrow). Polaschek (2012) notes several strengths of the RNR model including is explanatory depth and its practical utility; several weaknesses are also addressed such as the lack of simplicity and parsimony and the underdevelopment of the responsivity prong, among others. Some research indicates that adherence to the RNR framework results in improved treatment outcomes (Hanson et al., 2009).

The Four Ps (Weerasekera, 1993, 1996). Weerasekera identified a number of issues around the education of case formulation for medical school students and developed the Four Ps model of formulation to address issues in the field of psychiatry and mental health. Notably, Weerasekera identified the lack of treatment recommendations in many formulations as problematic and proposed a multi-perspective model in which clinicians consider the Four Ps (predisposing, precipitating, perpetuating, and protective factors) to explain a number of individual client and systemic factors. Individual factors include biological, behavioural, cognitive, and dynamic and systemic factors include relevant systems in an individual's life—the couple, family, occupation/school, and social system. Two other factors, coping-response style and treatment, were also included to capture treatment considerations. Weerasekera advocated for an iterative process in which, during the course of a client's treatment, formulations would be tested and revised based on new information and progress, or lack thereof, in treatment. The Four Ps model was developed for general clinical case

formulation, but is also often applied to forensic case formulation to explain violence perpetration and offending behaviours (Mindoudis et al., 2013). Now commonly referred to as the 5 Ps Model, clinicians also consider the clients presenting problem in their formulation.

The Four Ps approach to case formulation is potentially limited in the context of violence risk formulation as it was conceived as a formulation method for general clinical work. There have been no specific empirical investigations regarding the validity and IRR or utility of the Four Ps formulation method.

Offence Paralleling Behaviour (Daffern et al., 2007; Jones, 1997, 2004). The Offence Paralleling Behaviour (OPB) framework views offending behaviours in terms of patterns, and importantly, the focus is on the sequential development of such behaviours. Jones proposes a functional analysis of behaviour when examining offences and conceptualizes OPBs as the pattern of behaviours that occur prior to and just after an offence or fantasized offence. Jones indicated that the pattern of behaviours does not necessarily have to result in an offence but can merely resemble the behaviours that led to a prior offence. Important to the analysis is the exploration of the behaviours, thoughts, and emotions the person engaged in and experienced within the pattern. Jones has indicated that this identified offence cycle is a testable hypothesis about the function of the behaviours the offender has engaged in and may repeat in the future. Of course, it is relatively easy to identify the usefulness of this approach to formulating violence risk as it helps practitioners as well as offenders to develop strategies to recognize their pattern of violence and interventions to prevent violence before it occurs in the future.

Much of the peer-reviewed writing about the OPB paradigm is theoretical in nature. However, there are a limited number of empirical studies focused on comparing inpatient behaviours with offenders' index offences, with one study also focusing on recidivism (see: Daffern et al., 2007; Daffern et al., 2008; Daffern et al., 2009). The findings were mixed. Generally, these studies found that, in patients with personality disorders, there is some similarity between inpatient aggression and the index offence or recidivistic offending behaviours, but not for sexual violence. In addition, no relationship between pre- and post-aggressive behaviour and inpatient aggression was found. Much like the other formulation models, research regarding OPB is lacking.

Good Lives Model (Laws & Ward, 2011; Ward, 2002; Ward & Brown, 2004; Whitehead et al., 2007). The Good Lives Model (GLM) was developed in response to nearly two decades of research related to offender rehabilitation and the RNR model. Ward and colleagues recognized the importance of risk management, an integral prong of the RNR model, but did not view it as sufficient when working to rehabilitate offenders. Instead, they argue that helping offenders live more fulfilling lives is the best way to reduce recidivism. Ward and colleagues promote a more positive, strengths-based intervention model. The focus of the model is not just on risk factors, but on 11 primary goods that offenders first learn to prioritize based on their own personal preferences and motivations and then engage in behaviours that align with the promotion of these goods. Some of these primary goods include life (healthy living and functioning), excellence in play (hobbies and recreational pursuits), relatedness (intimate, romantic, familial relationships), spirituality (finding meaning and purpose), pleasure, and creativity. Within the GLM framework, criminogenic needs and antisocial behaviour arise from these basic needs and goods not being met. Initially developed and most often used in sexual offender treatment settings, the model has also been generalized to treat other types of violence including IPV. Glaser (2011) notes a limitation of the GLM is its paternalistic approach to working with offenders. Others have noted the model's focus on non-criminogenic needs as potentially problematic (Ogloff & Davis, 2004). Empirical work examining treatment outcomes are relatively positive (Willis & Ward, 2013), but much of this work has been conducted by Ward and colleagues.

Given the scope of this study, the focus will be on the an approach to violence risk formulation often recommended for use with SPJ guidelines, which has two general steps. First, past violence is conceptualized using a Decision Theory, or Action Theory, framework. Then a second formulation for future possible scenarios of violence is considered (Hart & Logan, 2011). Using Decision Theory in an SPJ framework encourages the examiner to consider motivators, disinhibitors, and destabilizers that impact a person's decision to engage in violence. These motivators, disinhibitors, and destablizers are linked to present and relevant risk factors to help formulate an explanatory narrative of a person's decision to engage in violence (see below). This formulation serves as the basis for scenario planning, devising risk management strategies, and arriving at a conclusory opinion of risk for a particular case.

Decision (Action) Theory

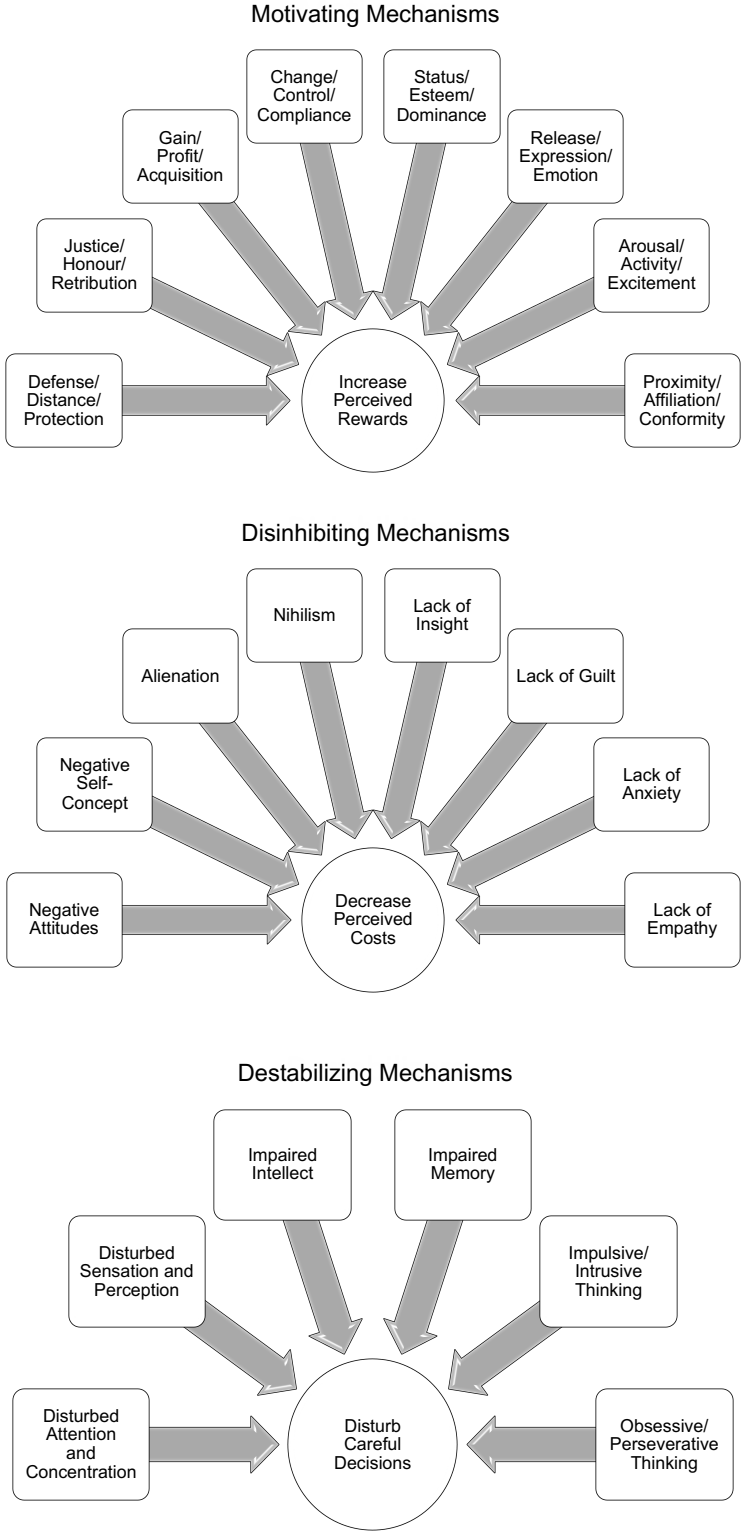
The basis for the SPJ method of violence risk formulation lies within the Decision, or Action, Theory framework. According to the Decision Theory framework, violence is conceptualized as a choice—an individual who commits an act of violence is acting with intention, making the decision, or choosing to engage in violence (Hart & Logan, 2011). The decision could be impulsive and reactive, or it could be carefully planned. Within this theory, decisions to engage in violence are goal-directed, planful and thoughtful, but are not necessarily conscious or rational. When conceptualizing violence as a decision, there must be a goal, motivation to pursue this goal, intent with volition, and action. An individual must first view violence as a viable response option and then perceive potential rewards or benefits, the cost of the violence must be acceptable to the person, and the plan to commit violence must be feasible (Hart & Logan).

The Decision Theory framework of explaining violence perpetration is informed by Felson (2009) and Wikström and Treiber (2009). Felson (2009) argues that violent crime, different from violence and crime, can only be conceptually explained by considering both theories of aggression and deviance. Felson demonstrates that attempting to explain violent crime with either only aggression or only deviance will result in an insufficient explanation. He further conceptualizes violent crime as instrumental behaviour that can be traced back to our most basic human desires. In addition, violent crime is ultimately motivated by reward-seeking. Felson contends that the constructs of violence and crime are distinct but overlap. This theory incorporates rational choice theory, which explains the 'why' when they overlap—when violence and crime become violent crime. Conceptualizing violence as a rational choice can be controversial. Often those who act violently are acting on impulse, out of anger, while drunk or high, and so forth. So how is it possible that violence is a rational choice? Here, Felson offers a convincing argument to conceptualize violence as a *bounded* rational choice: the rationality of the violence (reward-seeking behaviour) should be considered within the individual's specific context—their situation and judgments about the perceived rewards and costs. To an outsider looking in, the decision to engage in violence may not seem rational. However, within this framework, the individual deciding to engage in violent behaviour and their cost-benefit analysis is impacted by many idiographic factors. In many ways, because of this focus on idiographic factors this theory is well matched to the SPJ case formulation model. Wikström and Treiber (2009) offer a general theory of violence via the Situational Action

Theory, which complements Felson (2009). Wikström and Treiber (2009) explain violence as moral action that integrates various influences in the decision to engage in violence including particular circumstances, emotion, perceived lack of other options, and habituation among others. Together, these two broad theories of violence inform the SPJ method of case formulation.

A conceptual framework to help users of SPJ guidelines organize their thinking when formulating an individual's violence risk has been developed. This framework is not a measurement model, but rather a conceptual framework. When formulating violence within the SPJ framework, users engage in an analysis of how risk factors work through mechanisms which motivate, disinhibit, and destabilize the offender that ultimately influence the decision to engage in violence (see Figure 1). Motivators increase the perceived rewards of engaging in violence, disinhibitors decrease perceived costs or negative consequences of violence, and destabilizers disturb careful decision making (Hart & Logan, 2011). Motivating mechanisms include attempts at self-defence or protection; seeking justice, honour, or vengeance; seeking control, change, or compliance in a situation of conflict; asserting one's status, esteem, or dominance over another; release or expression of emotion; seeking arousal, activity, excitement, or stimulation; and establishing or strengthening proximity, affiliation, or conformity. Disinhibiting mechanisms include holding negative attitudes; negative self-concept; alienation from others; nihilistic attitudes and beliefs; and lack of insight, guilt, anxiety, and empathy. Finally, destabilizers include disturbed attention and concentration; disturbed sensation and perception; impaired memory; impaired reasoning; inflexible, obsessive, and perseverative thinking; and impulsive and intrusive thinking. Analysis of how relevant risk factors impact the decision to engage in violence results in an explanatory narrative that is unique to the individual and simple to understand. After determining the motivators, disinhibitors, and destabilizers (an explanation of past violent behaviour), the user then engages in another form of violence risk case formulation known as scenario planning. Scenario planning is an exercise in which possible futures are considered, keeping in mind the relevant risk factors and violence risk case formulation. Repeat, twist, escalation, and desistance scenarios are considered. These two case formulations then inform management strategy planning.

Figure 1 **Motivating, Disinhibiting, and Destabilizing Formulation Mechanisms (after Hart, 2015)**



The Decision Theory method is inherently abductive and philosophically pragmatic – with emphasis on reaching the best and simplest explanation of violent behaviour given the information available (Hart et al., 2011; Ryan et al., 2019). The objective is to explain mechanisms of violence and this is contrasted with the actuarial approach which uses a form of logic known as inductive projection, where the goal is to categorize individuals in order to predict violence. Abductive reasoning is well matched to the adversarial legal system (Pardo & Allen, 2008) in which transparent, easily explained and understood reasoning is paramount to decision makers (i.e., judges, juries).

Forensic Case Formulation Literature

Currently, there are a limited number of empirical investigations examining forensic case formulation. Also known as violence formulation and forensic case conceptualization, a number of aspects of forensic case formulation are crucial to evaluate. Informed by the general clinical case formulation literature, important areas for empirical study for forensic case formulations include most aspects of measurement validity, particularly IRR and predictive validity, as well as the quality of case formulations, treatment utility, and issues of training and competency (Mumma, 2011). The small number of existing articles dedicated to forensic case formulation fall into two broad areas: theoretical or conceptual analyses and empirical investigations.

The theoretical analyses of forensic case formulation have begun to lay the groundwork for important empirical investigations (Hart et al., 2011; Lewis & Doyle, 2009; Sturmey & McMurran, 2011). In addition, Davies et al. (2013) present issues around the pragmatics of using case formulation in forensic settings through two case studies. Problems such as client understanding and buy-in of case formulations as a guide for treatment and problem-targeting and the development of competent case formulation skills in trainees and professionals were addressed. Ultimately, Davies et al. suggest a useful future area of forensic case formulation research may be evaluating the utility of case formulations and focusing on the development of quality frameworks for use with forensic clients. Highlighting the role of forensic case formulation in the courtroom, Kapoor and Williams (2012) briefly critique psychiatry's movement away from psychodynamic conceptualizations of behaviour and use case studies to illustrate the ostensible lack of congruence of psychodynamic formulations in the courtroom. Kapoor and Williams caution forensic professionals from formulating cases too narrowly, in fear of an inadmissibility

ruling, as this approach may inhibit a full and complex understanding of a person's behaviour and psychological processes. Although focused specifically on the psychodynamic framework, the concerns raised regarding the congruence of case formulations and the law are important to consider for all frameworks. Introducing further conceptual work, Logan and Johnstone (2010), Gatner et al. (2017), and Gatner (2019) considered the role of forensic case formulations specific to offenders suffering from personality disorder, both generally and also specific disorders including Psychopathic Personality Disorder. Palmer (2016) attempted to outline a number of issues with forensic case formulation, including the identification and use of dynamic risk factors and lack of clarity around the process of formulation, and introduced the Risk Etiology Case Formulation Model as method for clinicians to formulate violence risk.

The tradition of establishing theory or concepts in their infancy through the use of case studies is conventional in the field of psychology and no different in the area of forensic case formulation. In addition to Davies et al. (2013) and Kampoor and Williams (2012), several other case study papers have been published and have helped develop a theoretical basis for forensic case formulation. Across a number of presenting client problems–schizophrenia, sexual offending, personality disorder, and IPV perpetrated by women—various researchers have focused on presenting the utility and therefore conceptual basis for forensic case formulation in practice (Bjørkly et al., 2014; Connell, 2015; Mappin et al., 2013; Vess et al., 2008).

Although the number of studies empirically examining various components of forensic case formulation is limited, the literature has started to accumulate. The small number of studies are focused on issues around knowledge of case formulation, training, and competency. Minoudis et al. (2013) conducted a pre-post study with probation officers (POs) in which POs provided a formulation based on case vignettes pre-training. Both vignettes were fictional, but written in a way to resemble clients typically seen in the service where the study was conducted. In one vignette, antisocial personality traits were more prominent and in the other borderline personality traits. POs then attended a formulation training program based on the '5 Ps' framework (Weerasekera, 1996). The quality of case formulations was evaluated with the Case Formulation Quality Checklist (CFQC; McMurran et al., 2012). This also provided Minoudis and colleagues the opportunity to examine psychometric properties of the checklist. The researchers found the CFQC to have good psychometric properties. Intraclass correlation coefficients (ICCs) were used

to examine IRR and resulted in moderate to good agreement. In the study, ratings made using the CFQC had good internal consistency, α = .92 and good test-retest reliability, $ICC_{(2,1)}$ = .85 - .99. In regard to training effects, there was no pre-post difference in quality scores for Vignette 1. However, there was a training effect for Vignette 2, $t(12) = 2.28$, $p < .05$. Overall, the authors concluded that the CFQC demonstrated good psychometric properties and may be appropriate to evaluate formulation. In regard to training and competency around the task of forensic case formulation, the POs struggled with the formulation skills. Minoudis and colleagues hypothesized this could be an indication of potential problems with their training program or perhaps a fundamental misunderstanding on the part of POs. Hopton et al. (2018) conducted another study evaluating the quality of case formulations. Examining case formulations developed using HCR-20^{V3}, Hopton et al. compared the quality of HCR-20^{V3} case formulations between HCR-20 Version 2 and 3 also using the CFQC-Revised (CFQC-R; McMurran & Bruford, 2016) across N = 121 forensic psychiatric hospital patients. The overall median score at the item level on the CFQC-R across cases was 6 with a mean total score of 53, indicating poor to intermediate quality. Formulations made using Version 3 were rated as higher quality overall than those made using Version 2.

In addition to Hopton et al. (2018), a small number of additional studies have focused on case formulation within the SPJ framework. Sutherland et al. (2012) conducted a study examining the IRR of the Risk for Sexual Violence Protocol (RSVP; Hart et al., 2003) using a sample of N = 28 professionals from varied professional backgrounds including nurses, clinical psychologists and psychiatrists who worked in forensic mental health or intellectual disability settings. These professionals completed a training workshop on use of the RSVP. In addition to examining rater agreement on the presence and relevance of risk factors and Conclusory Opinions, Sutherland and colleagues also examined agreement for RSVP Steps 4 (risk scenario planning) and 5 (risk management strategies) – both steps integral to formulation. Sutherland et al. found that agreement between raters varied based on training and expertise with greater training and expertise associated with greater agreement. In regard to Steps 4 and 5, rater agreement was mixed. Items related to scenario planning generally had fair agreement, although those related to recommended supervision and monitoring had excellent agreement.

Wilson (2013) also specifically focused on an evaluation of the IRR and similarity of RSVP case formulations produced by participants who completed an online training

program. Using a within-case and across-case design, Wilson found some aspects of case formulation (developing risk scenarios, risk management planning) were potentially easier skills to integrate and understand than others (understanding the motivators, destabilizers, and disinhibitors in a particular case). Sher and Gralton (2014) also focused on training, but specific to the administration of the Short-Term Assessment of Risk and Treatability (START; Webster et al., 2009). After *N* = 91 staff at a UK medium secure adolescent service completed START training, attendees were asked to complete a survey (30% of attendees provided completed surveys). In regard to questions around case formulations, formulation skills were most frequently identified as an area in need of training at a rate of 50%.

In the United Kingdom (UK), a new high-risk offender management program, the Offender Personality Disorder Program, is being implemented. A focal point of this program is that interventions for these offenders are psychologically informed and that the interventions include a formulation-based approach to managing high-risk offenders. In anticipation of this program's implementation, Brown and Völlm (2013) have conducted two studies focused on if and how non-clinicians and non-specialists can formulate forensic cases. Using a qualitative design, Brown and Völlm focused on identifying the level of knowledge of case formulations with *N* = 19 probation officers as part of a broader project to develop a case formulation training package for offender managers (probation officers) in the UK. A series of focus groups were used to identify the current, transferrable skills that could serve as a basis for case formulation training. Probation officers tended to ask about many of the psychosocial domains necessary to formulate. Brown and Völlm noted there was a lack of training and knowledge of the case formulation skillset among probation officers, however they were willing to engage in training. In an additional qualitative, focus group study examining the implementation of case formulation skills among probation officers, Brown and Völlm (2016) interviewed probation officers, offenders, and close family or friends of the offenders regarding the implementation of case formulations into the offenders' "pathway plan" (a plan that directs offenders to appropriate interventions in custody or the community). Themes of role conflict were expressed by offenders who were concerned about sharing the information necessary for offender managers (OMs) to develop a quality formulation given the relationship between offenders and OMs (one in which the probation officer is viewed in a role of authority, implementing supervision and surveillance strategies, requiring accountability). Offenders

also expressed concerns regarding power and trust. Some of the concerns expressed by OMs included information gathering, such as mental health, that OMs may not have the skills properly assess or deescalate if the client becomes dysregulated when discussing symptoms. Ultimately, Brown and Völlm note that because of some of these concerns, implementation of the formulation-based approach may be challenging.

Evaluating Forensic Case Formulations

Many of the same difficulties discussed previously regarding the evaluation of clinical case formulations pertain to forensic case formulations. Delle-Vergini and Day (2016) highlight a number of the issues with forensic case formulation including a lack of empirical literature specific to forensic practice, the adversarial circumstances under which forensic clinicians work, and the need for formulations to be understandable to multiple stakeholders (e.g., the evaluee, the judge, attorneys, treatment providers, probation officers). Arguably, two suitable starting points to evaluate case formulations are considerations around quality and IRR. Bucci et al. (2016) conducted a systematic review of measures used to assess the quality of case formulations and found eight measures in the literature, with the CFQC being one of three measures with the most empirical evidence. The authors observed, however, that research around measures of quality in the literature generally lacks empirical evidence.

Current Study

The scope of the current project was primarily focused on the evaluation of forensic case formulations within the SPJ framework using a Decision Theory approach. This was the first study evaluating SARA-V3 that incorporated all SPJ steps—it was also one of few studies conducted on SPJ guidelines to do so. This study had two parts. Part 1 focused on the IRR of forensic case formulations. Part 2 focused on the similarity and quality of forensic case formulations. I used a sample of adult male IPV offenders to examine the following research questions.

Research Questions:

1. What is the distribution of closed-ended forensic case formulation mechanism ratings?

2. What is the agreement between raters (i.e., IRR) of closed-ended forensic case formulation mechanism ratings?

3. What is the similarity of narrative forensic case formulations?

4. What is the quality of narrative forensic case formulations?

5. Does the expertise of raters influence the patterns of findings for questions 2, 3, and 4? (Note: Research Question 5 was addressed within Questions 2, 3, and 4 in the Results section.)

Chapter 2. Method

Overview

The current study was based on archival data. The study consisted of two parts using the same participant sample. An existing dataset comprising *N* = 100 closed IPV offender files that were initially referred to an outpatient forensic clinic for assessment in British Columbia, Canada was utilized for the current study. Ethical approval to conduct this study was granted by Simon Fraser University and British Columbia Mental Health and Substance Use Services. All offenders included in the sample committed at least one violent offence within the context of an intimate partner relationship (or perceived relationship). Referral dates to the outpatient forensic clinic, signifying the index offence in this study, ranged from 2000 to 2009. Given the types of offences committed and that a vast majority of the offenders included in this sample were released on bail and living in the community or in custody at the time of the offence, this sample can be considered one of *moderate to high* risk to manage in the community. Upon referral to the forensic outpatient clinic, all offenders included in the sample participated in a psychological evaluation in which a registered clinical psychologist used SARA-V2 to guide decisions around level of risk management. The psychologist then prepared a forensic psychological pre-sentence report and submitted the findings to the judge ruling in the case. Raters in the current study (TR, YL, LV, WF, KG, JB) made SARA-V3 risk ratings independently (i.e., blind to ratings made by other raters). Raters TR and YL were considered Expert raters due to their status as Ph.D. students at the time the study was conducted, which included extensive prior training in violence risk assessment and the use of violence risk assessment tools in clinical settings. Raters JB, WF, KG, and LV were considered Novice raters as they had little or no prior training or experience using violence risk assessment tools in clinical settings. All raters completed the first *n* = 5 cases. For the remaining cases, each case was coded by *n* = 2 Experts and *n* = 2 Novices. Given that raters completed an unequal number of cases, this design is nested. A visual illustration of the nested study design is located in Table 2.

Table 2 Nested Study Design Visualization

n (cases)	Expert TR	Expert YL	Novice LV	Novice WF	Novice KG	Novice JB
5	X	X	X	X	X	X
11	X	X	X	X		
6	X	X	X		X	
6	X	X	X			X
6	X	X		X	X	
6	X	X		X		X
10	X	X			X	X
N = 50	*n* = 50	*n* = 50	*n* = 28	*n* = 28	*n* = 27	*n* = 27

Next the narrative case formulations written by the *n* = 6 raters were analyzed. The narrative case formulations were evaluated for similarity and quality by three additional research assistants (RAs) (SC, LW, EF). These RAs did not code any SARA-V3 risk assessments and thus were blind to the cases and raters. SC had a Ph.D. in clinical-forensic psychology and had recently completed a postdoctoral fellowship in forensic psychology; she was a registered psychologist. LW was a Master's level graduate student in a clinical psychology program who had previously completed violence risk assessment training on the HCR-20^{V3}. The final RA, EF, was enrolled as a Ph.D. student in the clinical-forensic psychology program who had training in violence risk assessment and practicum experience using SARA-V3.

Two other studies have been conducted using cases from the dataset included in the analysis of the current study. Storey and Hart (2014) published a study evaluating the Danger Assessment (DA; Campbell, 1986; Campbell et al., 2003). The DA is an IPV risk assessment that was developed to predict the risk of lethal IPV. They found that the DA tended to overestimate risk of the offenders included in the sample – more than half of the sample fell in the highest risk category. None of the offenders in the sample committed a lethal act of IPV in the follow-up period (*M* = 5.19 years). Additionally, Ryan (2016) conducted a pilot study to the current investigation. This unpublished thesis examined the IRR and concurrent validity of SARA-V3. One rater (TR) coded *N* = 97 files using SARA-V3 and a second rater coded a subset of *n* = 30 files to evaluate IRR. Only the presence of risk factors, relevance ratings, and Conclusory Opinions were coded. In addition,

concurrent validity was evaluated by comparing SARA-V3 ratings to other previously coded measures of IPV risk. Overall, Ryan found intraclass correlation coefficients (ICCs) fell mostly in the fair to good range suggesting adequate IRR. Medium to large correlations between SARA-V3 and other measures of IPV risk substantiated adequate concurrent validity of SARA-V3.

Cases

A sample of *N* = 50 closed-case files of male offenders referred to an outpatient forensic psychiatric clinic for a pre-sentence psychological evaluation were included in the analysis. Initial assessment dates ranged from 2000 to 2009. The mean age of participants at the time of the initial assessment was 37.70 (*SD* = 8.68) with a range of approximately 19 to 60 years old. Participants included in the sample were legally involved after committing at least one act of violence in the context of an intimate partner relationship. Ethnic composition was predominantly people of European (64%) and South Asian (22%) decent. East Asian (4%), African (6%), and Indigenous (4%) decent comprised the remainder of the sample. With regard to the relationship between the offender and victim at the time of the index offence, a majority of the sample was either divorced or separated (46%) or were formerly in a non-marital relationship (40%). Eight percent were married and 4% were dating. The most frequent index offence charge was assault (*n* = 31), followed by breach of conditions (*n* = 23), threats (*n* = 22), and criminal harassment (stalking) (*n* = 18). Other charges ranged in type and severity but included alleged offences such as assault with a weapon, attempted murder, sexual assault, possession of a weapon, and attempted kidnapping.

Procedure

This was an archival study in which data were coded from file information in order to complete the SARA-V3. The composition of information varied from file to file, but always included police reports, criminal record information, charges and convictions, and a psychological pre-sentence report. Depending on the file, information may have also included victim and witness police statements, victim interviews, probation pre-sentence reports, relevant hospital records such as psychiatric hospital admissions, in-custody

institutional infractions, and participation in mental health treatment or substance use treatment programs.

Case selection. The *N* = 50 cases included in this study were selected from the *N* = 97 used by Ryan (2016). A project investigator who was not involved with coding data selected the cases using stratified random sampling. First, the cases in Ryan (2016) were stratified on the basis of Case Prioritization ratings; second they were further stratified on the basis of total number of P factors rated *yes* for Presence-Past. As only *n* = 4 cases in Ryan (2016) were rated *low* on Case Prioritization, all 4 were selected for inclusion in the present study. The project investigator then selected 23 cases rated *moderate* Case Prioritization and *high* Case Prioritization. Within each of the *moderate* and *high* Case Prioritization groups, the project investigator randomly selected *n* = 7 cases from the bottom tercile of the distribution of total number of P factors rated *yes*, *n* = 9 cases from the middle tercile, *n* = 7 from the top tercile.

After the *N* = 50 cases were selected, the project investigator determined the order in which the cases would be completed. This was done to ensure that each rater coded approximately the same number of cases rated *low, moderate, or high* Case Prioritization and that the raters did not encounter a long run of cases with the same Case Prioritization rating. To this end, the project investigator divided the *N* = 50 cases into 5 groups of 10 that were approximately equivalent in terms of Case Prioritization ratings and also were approximately equivalent in terms of risk. Finally, the project investigator randomized the order of presentation of cases within each group as well as the order in which groups was presented.

Narrative case formulations. There was a total of *n* = 57 Expert narrative formulations and *n* = 70 Novice narrative formulations available for coding and analysis. SC was the primary rater for quality ratings while LW and EF provided primary ratings for similarity ratings. SC provided quality ratings for all *n* = 127 formulations. LW and EF provided IRR ratings (LW rated *n* = 100; EF rated *n* = 27 cases). With regard to similarity ratings, all *n* =127 cases were included in the sample of case pairings. There was a total of *n* = 143 pairings created using a stratified random procedure. There were two higher-order categories of case pairings: *Within Case* pairings included two formulations from the same case written by different raters, and *Across Case* included two formulations from different cases written by different raters. Among both the *Within Case* and *Across Case*

categories, the following pairings were created: Expert-Expert (*n* = 46 pairs: 20 within, 26 across), Expert-Novice (*n* = 50 pairs: 25 within, 25 across), Novice-Novice (*n* = 47 pairs: 24 within, 23 across). Thus, there were a total of *n* = 69 *Within Case* pairings and a total of *n* = 74 *Across Case* pairings.

Rater training. Four Novice raters (LV, WF, KG, JB) enrolled in graduate programs at Simon Fraser University (one in clinical-forensic psychology, one in experimental-forensic psychology, two in criminology) were trained in the use of SARA-V3 via an online, 20-hour training accessed through the Concept Professional Training website (www.concept-ce.com). Dr. Stephen Hart led the pre-recorded online training. The training has been approved by the American, British, and Canadian Psychological Associations, among other regulating bodies and states, as meeting Continuing Education (CE) standards and is worth 20 CE credits. This particular training program was specific to competent use of SARA-V3. The first part of the training focused on foundational risk assessment concepts, then provided more specific information about the development and use of SARA-V3. The training included an overview of each step involved in using SARA-V3 to assess violence risk. In addition, quizzes that required a minimum grade of 70% were completed by trainees to proceed to the next stage of training. The second half of the training involved participants reviewing an actual redacted case and using the SARA-V3 to complete a risk assessment following a discussion of ratings made and completion of a quiz to check for understanding. Novice raters also took part in an hour-long online case formulation training, presented by Dr. Kelly Watt. Raters were provided with a formulation coding book which included definitions of the various motivators, disinhibitors, and destabilizers as well as brief comments on the procedures for forensic case formulation within the SPJ framework. In addition to the practice case completed during the online SARA-V3 training, Novice raters each completed 2 additional practice cases and met with Expert rater TR and one other fellow Novice rater, both of whom completed the same case to discuss ratings, clarify any coding errors, and reach consensus ratings. Novice raters were also given written feedback on their narrative formulations. These cases were coded for practice purposes only and not included in the data analysis. Expert raters YL and TR did not participate in the online trainings because both had already completed extensive didactic risk assessment training (each exceeding 100 classroom hours) and completed clinical practica placements each exceeding 450 hours where SARA-V3 was used on actual cases.

SARA-V3 Coding Procedures. After a comprehensive review of file data (Step 1), raters completed all steps of SARA-V3. In Step 2, raters considered the presence of past and recent Nature of IPV Factors (N factors) and Perpetrator Risk Factors (P factors). Presence ratings were made using a 3-point ordinal scale that were then recoded numerically: [*yes* = 2*, possibly or partially present* = 1*, no* = 0]. Next in Step 3, raters made Relevance ratings for P factors using a 3-point ordinal scale that were then recoded numerically [*yes* = 2*, possibly or partially relevant* = 1*, no* = 0]. Raters then formulated the cases. Data collection of case formulations occurred in both a freeform narrative format and a closed-ended format in which raters identified the particular motivators, disinhibitors, and destabilizers on a 3-point ordinal scale that were then recoded numerically: [*yes* = 2*, possibly or partially present* =1*, no* = 0]. Raters also identified P factors linked to any motivator, disinhibitor, and destabilizer coded as *yes* or *possibly or partially present.* These linked P factors were coded dichotomously (nominally): [*yes* = 1, *no* = 0]. In Step 4, raters completed scenario planning and considered four potential future scenarios: repeat, twist, escalation, de-escalation. For each future scenario, raters also considered a number of additional ratings. These variables included the risk for physical and psychological harm [*low, moderate, high*], the most likely victim(s) within the scenario, the likelihood of the scenario occurring [*low, moderate, high*], and the imminence of the possible scenario [*immediately, within 24 hours to one week; in the coming two to six weeks; in the coming 6 to 12 months; in the coming several years*]. Raters then indicated their judgment of the most likely future scenario. In Step 5, raters developed management plans across four areas: monitoring/surveillance, treatment/assessment, supervision/control, victim safety planning. Raters were instructed to provide at least one recommendation for each area. Finally, in Step 6, raters indicated summary judgments [*low, moderate, high]* for each of the Conclusory Opinions including Case Prioritization, Risk for Serious Physical Harm, Imminent Violence, and Other Risks Indicated.

Each rater coded the first five cases. After each of the first five cases, raters met and discussed their ratings to derive consensus ratings—ratings that all raters agreed were appropriate for N and P factor Presence and Relevance ratings, formulation mechanisms, most likely scenario, and Conclusory Opinions. In this process, raters would discuss their ratings and come to an agreement on any ratings in which they differed. Of note, most victim vulnerability factors were *omitted* as a result of a lack of victim information on file, thus consensus ratings were not completed for V factors. After the first

five cases were completed, raters met to establish consensus ratings after approximately every fifth case. TR led all consensus meetings. After the first five cases, each case was coded by a total of *n* = 4 raters (*n* = 2 Expert, *n* = 2 Novice). Individual raw Expert and Novice ratings were used in analyses related to IRR; consensus ratings were used for distribution analyses.

Similarity and Quality Coding Procedures. The similarity of two paired formulations was examined in Part 2. Pairs were organized *Within Cases* and *Across Cases* and within these two larger categories, paired in the following manner: Expert-Expert, Novice-Novice, Expert-Novice. Three additional RAs (SC, EF, LW) were recruited to evaluate the similarity and quality of narrative case formulations written by the *n* = 6 raters in Part 1. All RAs were oriented to the use of the CFQC-R and similarity coding procedures. Initially, SC worked with another research assistant to code five training cases using both measures. Both RAs completed five cases independently and then met to discuss differences in coding and establish consensus coding for both quality and similarity. The second rater was unable to continue with the project. Thus, EF and LW were brought on to complete the coding. Both EF and LW completed each of the five training cases and their ratings were compared to SC's. EF and LW were provided feedback regarding their coding for each of the training cases. Prior to beginning coding, SC, EF, and LW met to review coding procedures and clarify any problems with coding. The order in which RAs completed similarity and quality ratings was randomized.

Measures and Materials

SARA-V3

SARA-V3 was administered according to typical procedures. The distribution of Presence and Relevance ratings was reported by Ryan (2016) in the sample from which this current sample was drawn. In that study, the distribution was skewed for some Presence and Relevance ratings. Generally, raters in Ryan (2016) achieved a *moderate* range of agreement across Presence and Relevance ratings for N, P, and Conclusory Opinions. The distribution of ratings in the current sample were similar to those reported in Ryan (2016). IRR for Presence and Relevance ratings of individual N and P factors, Scenario Planning, and Conclusory Opinions made using SARA-V3 in the current sample were generally in the *moderate* to *almost perfect* ranges. With regard to the distribution of

the most likely future scenario, these ratings were skewed: repeat (94%), escalation (4%), de-escalation (2%), twist (0%). Interrater reliability for the most likely future scenario was substantial, AC_2 = .84, $p \leq .001$, 95%CI = [.66, 1.00]. Raters designated the most likely future scenario and Experts were more reliable than Novices, AC_2 = .93, $p \leq .001$, 95%CI = [.71, 1.00] and AC_2 = .74, $p \leq .001$, 95%CI = [.44, 1.00], respectively. The distributions and IRR of Presence and Relevance ratings for N and P factors, Scenario Planning, and Conclusory Opinions are presented in the Appendices C through F.

Similarity

Similarity ratings for paired narrative forensic case formulations were made using a form adapted from Wilson (2013), located in Appendix A.1. The similarity form was divided into four sections: 1. Overall similarity, 2. Similarity of the motivators, disinihibitors, and destabilizers, 3. Similarity of scenarios, 4. Similarity of risk management strategies. Most ratings were made using percentages (0% to 100%) in 10% increments using the following anchors: 0% = *not similar at all*; 40% = *different in many important respects, but similar in a few*; 70% = *similar in many important respects, but different in a few*; 100% = *very similar*. On an item in which RAs judged the overall similarity of case formulations, agreement was *fair,* AC_2 = .46, $p \leq .001$, 95%CI = [.32, .59] and percent agreement was 82%.

CFQC-R

The CFQC-R is a 10-item checklist that was used to evaluate case formulations on a 10-point scale that included three anchor points: 0 = *Does not meet this criterion at all,* 5 = *Meets this criterion somewhat,* 10 = *Meets this criterion exceptionally well*. Total scores on the CFQC-R can range from 0-100. A limited number of studies have shown the CFQC-R to have good reliability and internal consistency (Hopton, et al., 2018; McMurran & Bruford, 2016). In the current study, the CFQC-R was adapted slightly to meet the needs of coding in this particular project (see Appendix A.2). In Part 2, RAs used the CFQC-R to assess formulation quality and achieved *substantial* agreement for total scores, AC_2 = .82, $p \leq .001$, 95%CI = [.76, .89].

Data Analytic Strategy

Interrater Reliability

Given the focus on IRR in the current study, a discussion about what I considered when strategizing IRR analyses is necessary. There are over 40 IRR coefficients currently available to researchers, so there was much to consider. There is rigorous debate among statisticians regarding the calculation of IRR (see Feinstein & Cicchetti, 1990; Feng, 2013a; Feng, 2015b; Krippendorff, 2004; Lombard et al., 2002; Zhou, Feng, Liu, & Deng, 2018). Much of the discussion centers around how the coefficients treat chance agreement (Cicchetti et al., 2017). For a multitude of reasons, statisticians tend to agree that all IRR coefficients have problems and limitations. In a review of various IRR coefficients, their assumptions, and paradoxes, Zhao et al. (2013) suggested an entirely new IRR coefficient is necessary given the limitations of the coefficients currently available to researchers. In addition to these problems with estimating IRR, the design of the current study (more than two raters and a nested design) introduced even more limitations in estimating IRR. However, the design of the current study was more true to psychological research–limited time and resources often require nested designs with restricted sample sizes. Some of the factors I deliberated over are outlined below.

One of the most frequently used IRR statistics, Cohen's kappa (κ) (Cohen, 1960), has been criticized in the field over the last few decades (Aickin, 1990; Brennan & Prediger, 1981; Feinstein & Cicchetti, 1990; Gwet 2014). The primary problem identified with Cohen's κ is that if the sample comprises cases that are unequally distributed (i.e., lacking in variance), the IRR coefficient will be artificially reduced–even if raters are in high agreement. This is known as the kappa paradox (Feinstein & Cicchetti, 1990). There are also limits with regard to study design and computation. For example, Cohen's κ is only appropriate for studies with two raters, only suited for nominal data, and does not take into account partial agreement, which is problematic with ordinal data. The alternative weighted κ (Cohen, 1968), which accounts for partial agreement, can only be calculated for two raters. Another IRR coefficient, Krippendorff's alpha (α) (Krippendorff, 1970), can handle any number of raters and missing data as well as any type of variable, including ordinal. Researchers have found that, like Cohen's κ, Cohen's weighted κ, and Krippendorff's α also suffer from kappa paradox-like results when data are invariant (Feng, 2013a; Gwet, 2002; Gwet 2008). Krippendorff (2013; 2016) has enthusiastically rebutted

the arguments against the kappa paradox and questions the definition of IRR used by Feng, Gwet, and others in the field. Krippendorff has argued that a test or tool used to measure a phenomenon, but which results in systematically lopsided variance, is inherently unreliable (but *cf.* Zhao et al., 2018).

As debate around agreement coefficients has circulated, some have hypothesized about the factors that impact the probability of chance agreement. Feng (2013) examined the various factors that impact IRR coefficients via a Monte Carlo analysis. Feng suggested IRR coefficients are impacted primarily by the level of difficulty of items being rated, the marginal distribution of ratings, and the interaction of these two factors. According to Feng, κ, Scott's Pi (π) (Scott, 1955), and α were especially impacted by marginal distributions. In other words, these statistics calculate chance agreement with the assumption that the coding work is always challenging and thus skewed marginal distributions result in high chance agreement. Feng argued that skewed marginal distributions are more likely the result of an easy coding task, which is why these three IRR coefficients are often impacted by the kappa paradox (i.e., high agreement, low reliability). Within the current sample, the distribution of ratings on formulation various items was skewed for a number of reasons. One possible explanation is the sample comprised a homogenous type of violence—IPV. When considering violence pathways, it makes sense that certain formulation mechanisms would be inherently more relevant to formulating IPV and there may be a systematic lack of variability across cases as the sample was predominantly *moderate-high* risk overall.

Another consideration for the current study was the number of raters per case. Some IRR coefficients only calculate agreement statistics for two raters, which automatically negates them from use in the current study. There are a number of agreement coefficients for more than two raters (Bennett et al., 1954); Conger, 1980; Guttmann, 1945; Light, 1971; Maxwell, 1977). However, these coefficients require a fully crossed design and are inappropriate for the current analysis given the data are nested. It is possible to calculate Intraclass Correlation Coefficients with missing data, but statistical programs typically use casewise deletion to handle missing data. One might inquire why missing data in the current study could not simply be imputed, but imputing missing data, (either systematic or random) does not make much sense in an SPJ framework given algorithms and total scores are not employed. This reasoning extended to the formulation mechanisms. Thus, imputing data was not an appropriate solution within

the current study. Finally, Fleiss' κ, a generalization of Scott's π, can calculate coefficients for more than two raters and handles missing data; however, the rating categories are assumed to be unordered (nominal) and thus ratings are not weighted. Much of the data in the current study was ordered and so it was not appropriate for most of the variables.

In sum, all IRR coefficients have various limitations (Feng, 2013b). Given the limitations of the coefficients previously discussed and the design of the current study, the selection of a coefficient that best estimated IRR was a relatively difficult task with no perfect solution. Of the various statistics that were appropriate for the data in the current sample, one that was well-suited for the analysis of IRR was Gwet's AC. Gwet's AC_1 was used for nominal ratings and AC_2 was used for ordinal ratings. I selected this statistic for the current analysis for several reasons. First, given the uneven distributions of the formulation mechanisms, the kappa paradox was likely to occur. Indeed, Wongpakaran et al. (2013) found that Gwet's AC_1 was more stable than both Cohen's κ and Scott's π. Quarfoot and Levine (2016) compared multi-rater IRR coefficients and found Gwet's AC_2 and Brennan-Prediger were more stable with varying marginal distributions. Of course, Gwet's AC_1 and AC_2 are not without problems. Cicchetti et al. (2017) disagreed with the manner in which Gwet's AC_1 chance agreement is calculated. Zhao et al. (2013) examined many of the available IRR coefficients used with nominal data and concluded that, comparatively, Gwet's AC_1 tended to produce unfairly high IRR coefficients when used with data that have highly uneven individual and average distributions and when the number of nominal coding categories was equal to or greater than three.

kappaetc Command

Klein (2018) wrote a new Stata 16.0 command, kappaetc, that generates five IRR coefficients, including Gwet's AC, percent agreement, Cohen/Conger's κ, Scott/Fleiss' κ, and Krippendorff's α. Klein provided an overview of the coefficients available through Stata and community generated coefficient commands. Depending on the type of variable, corresponding bipolar, ordinal, and ratio weights were applied. When more than two raters' data were included in the analysis, unconditional standard errors were applied. Because of the study's nested design, a matrix vector for missing values is added (see Klein). Zapf et al. (2016) examined the impact that randomly missing values in nominal data had on Krippendorff's α and Fleiss' κ. The researchers found that Krippendorff's α produced slightly lower IRR coefficients than Fleiss' κ. However, these results were not tested in

systematically missing data or in non-nominal data and thus the effects of such missing data are unknown. Of note, because of the nested data in the current design, most of the missing data could be considered systematically missing.

The results for other IRR coefficients including Brennan-Prediger coefficient, Cohen/Conger's κ, Scott/Fleiss' κ, and Krippendorff's α for the various formulation mechanisms are presented in the Appendix. These coefficients have a large range that span the Landis and Koch (1977) classifications, from *poor* to *almost perfect*. Generally, the coefficients calculated with Brennan-Prediger tended to show better IRR among raters for the formulation mechanisms than the other coefficients. Cohen/Conger's κ, Scott/Fleiss' κ, and Krippendorff's α tended to produce lower coefficients. Comparative coefficients presented by Expert raters and Novice raters are also presented in various Appendices.

Benchmarking

It is important to consider how the results of an analysis of IRR will be interpreted and communicated. Unsurprisingly, there is also much discussion among academics with regard to the qualitative labeling, or benchmarking, of IRR coefficients. Several authors have put forth suggestive interpretive guidelines regarding how to classify a specific IRR coefficient. Most of these scales have been suggested for the Cohen's κ, but are also applied to other agreement statistics (Gwet, 2014). Landis and Koch (1977) offer a scale with six categories (see Table 3). A revision of the Landis and Koch scale was suggested by Fleiss (1981), which provides three categories of benchmarking: <.40 = *poor*, .40 – .75 = *intermediate to good,* > .75 = *excellent.* Finally, Altman (1991) offered a 5-point scale that is mostly consistent with Landis and Koch (1977): <.20 = *poor,* .21 – .40 = *fair,* .41 – .60 = *moderate,* .61 – .80 = *good,* .81 – 1.00 = *very good.* Finally, another commonly used benchmarking scale was put forward by Cicchetti and Sparrow (1990): < .40 = *poor,* .41 – .59 = *fair,* .60 – .74 = *good,* .75 – 1.00 = *excellent.*

There are a number of considerations when selecting a benchmarking scheme for a particular study including the goal of the analysis (e.g., to establish dichotomous *poor* versus *good* agreement, finer categorization). Among other considerations is the design of the study. Gwet (2014) provides a Monte Carlo simulation using Cohen's κ to demonstrate how the number of categories, raters, and subjects impact precision of IRR coefficients. Ultimately, Cohen's κ is more accurate when the number of subjects or

categories increase. Given that these benchmarking schemes are deterministic and the true coefficient could change depending on the design of the study, Gwet suggests a probabilistic approach to benchmarking given that coefficients have inherent error associated with the number of raters, subjects, and categories. According to Gwet (2014) and Klein (2018), the proposed probabilistic approach to benchmarking coefficients is achieved by selecting a benchmarking scale (e.g., Landis and Koch, Fleiss, Altman), computing the standard error of the IRR coefficient, then for each benchmark interval computing the probability that a coefficient falls within it. Finally, examination of the cumulative probability helps determine the proper interval. In other words, starting at the highest interval, determine the cumulative probability moving down until the probability exceeds a pre-selected threshold (e.g., 95%).

Table 3 Landis and Koch (1977) Kappa Benchmarking

***Kappa* Statistic**	**Strength of Agreement**
< .00	Poor
.00 – .20	Slight
.21 – .40	Fair
.41 – .60	Moderate
.61 – .80	Substantial
.81 – 1.00	Almost Perfect

In the current study, I have selected the probabilistic benchmarking approach suggested by Gwet (2014) using Landis and Koch (1977) qualitative guidelines. Gwet (2014) argues that often confidence intervals span multiple benchmarking intervals. For example, if a particular set of ratings result in an IRR coefficient of .78 with a 95%CI [.35, .93], a deterministic approach would involve considering only the calculated IRR coefficient and designating a qualitative interval, in this case *Substantial*. However, the confidence interval, which is impacted by the number of raters, categories, and observations, spans from the *Fair* to *Almost Perfect* ranges. Thus, it is unclear where the strength of the agreement actually lies. Gwet argues that rather than using a deterministic approach, a probabilistic approach is superior given the impact that various sample characteristics (e.g., number of raters, categories, distributions) have on a calculated IRR coefficient. I find this argument compelling and thus have chosen to implement a probabilistic benchmarking method in my study.

Chapter 3. Results

Research Question 1. What is the distribution of closed-ended forensic case formulation mechanism ratings?

As a reminder, the individual cases were rated by at least *n* = 4 different raters (*n* = 2 Experts, *n* = 2 Novices). Given that the sample comprises IPV offenders, it was expected that certain mechanisms would be rated as present with more frequency than others. In other words, if the sample comprised more heterogeneous types of violence (e.g., general, IPV, sexual) ratings would be expected to be more evenly distributed across response options [*yes, possibly or partially present, no*]. Theoretically, certain motivating mechanisms could be associated with acts of IPV more frequently than others.

To test skewness, a large number of statistical calculations would be required. As such, ratings for mechanisms were considered skewed when >80% fell in one particular response option. The distributions of mechanism ratings for each rater were inspected and a number of them were skewed in similar patterns to the consensus ratings. To present rating distributions at the individual rater level would be quite detailed and cumbersome, thus only the distributions of consensus ratings are presented here for efficiency. Further, the IRR analyses that follow provided evidence that raters were in agreement with regard to the Presence of formulation mechanism ratings

To examine the distribution of formulation ratings, first an inspection of the frequency of Presence ratings for response options [*yes; possibly or partially present; no]* across the formulation mechanisms was conducted (see Table 4). As expected, the ratings for 11 mechanisms were skewed either *yes* or *no*. With regard to Presence ratings for motivating mechanism M1 (Defence/Distance/Protection), M3 (Gain/Profit/Acquisition), and M7 (Arousal/Activity/Excitement), they were skewed *no.* Presence ratings for M4 (Change/Control/Compliance) and M6 (Release/Expression/Emotion) held a similar pattern, but opposite in direction: they were mostly skewed *yes* and *possibly or partially present.* M5 (Status/Esteem/Dominance) had less extreme skewness, but trended *no.* Interestingly, ratings for two motivators, M2 (Justice/Honour/Retribution) and M8 (Proximity/Affiliation/ Conformity), were more evenly distributed across response options. Overall, based on the frequency distributions, the raters considered the primary motivating mechanisms in the current study to be M2 (Justice/Honour/ Retribution), M4

(Change/Control/Compliance), M6 (Release/Expression/Emotion), and M8 (Proximity/ Affiliation/Conformity). Raters considered offenders in the current sample to be less motivated to engage in violence through mechanisms M1 (Defence/Distance/ Protection), M3 (Gain/Profit/Acquisition), M5 (Status/Esteem/Dominance), and M7 (Arousal/Activity/ Excitement).

Table 4 Presence of Formulation Mechanisms

Motivators	Distribution, Presence %		
	Y	P	N
M1. Defence/Distance/Protection	0	4	96
M2. Justice/Honour/Retribution	22	28	50
M3. Gain/Profit/Acquisition	4	12	84
M4. Change/Control/Compliance	76	22	2
M5. Status/Esteem/Dominance	2	22	76
M6. Release/Expression/Emotion	66	26	8
M7. Arousal/Activity/Excitement	0	4	96
M8. Proximity/Affiliation/Conformity	42	30	28
Disinhibitors			
D1. Negative Attitudes	66	28	6
D2. Negative Self-Concept	0	6	94
D3. Alienation	2	24	74
D4. Nihilism	4	20	76
D5. Lack of Insight	94	4	2
D6. Lack of Guilt	90	8	2
D7. Lack of Anxiety	92	8	0
D8. Lack of Empathy	88	10	2
Destabilizers			
De1. Disturbed Attn. and Concent.	0	2	98
De2. Disturbed Sens. and Perc.	4	2	94
De3. Impaired Memory	0	0	100
De4. Impaired Reasoning	36	36	28
De5. Obsessive, Persev. Thoughts	62	28	10
De6. Impulsive, Intrusive Thoughts	76	18	6

Notes. Y = ye*s, present*; P = *possibly or partially present*, N = *no, not present*.

The distributions of Presence ratings for disinhibiting mechanisms were also skewed. Presence ratings for D2 (Negative Self-Concept) were skewed *no* and D5 through D8 (Lack of Insight, Guilt, Anxiety, Empathy) were skewed *yes.* Presence ratings for D3 (Alienation) and D4 (Nihilism) trended *no,* but also *possibly or partially present.* D1 (Negative Attitudes) ratings trended *yes,* but 28% of Presence ratings were *possibly or partially present*. Presence ratings indicated that within the current sample, D1 (Negative Attitudes) and D5 through D8 (Lack of Insight, Guilt, Anxiety, and Empathy) were the primary disinhibitors and that raters did not consider D2 (Negative Self-Concept), D3 (Alienation), and D4 (Nihilism) as primary disinhibitors in the decision to engage in violence among offenders in the sample.

Similarly, ratings for destabilizing mechanisms, or mechanisms that disturb careful decision making, were skewed. Presence ratings for De1 (Disturbed Attention and Concentration), De2 (Disturbed Sensation and Perception), and De3 (Impaired Memory) were grossly skewed *no.* Conversely, ratings for De5 (Obsessive, Perseverative Thinking) and De6 (Impulsive, Intrusive Thinking) trended *yes* and *possibly or partially present*. De4 (Impaired Reasoning) was more evenly distributed across response options. Raters did not identify De1 (Disturbed Attention and Concentration), De2 (Disturbed Sensation and Perception), or De3 (Impaired Memory) to be frequently destabilizing the decision to engage in violence among offenders in the current sample.

Distribution of Linked P Factors

Consistent with the process of formulation, raters also selected P factors that were linked to the individual motivators, disinihibitors, and destabilizers. Understanding the distribution of these linked P factors within each mechanism could be helpful in understanding how raters formulated cases and the related risk factors associated with their formulations in the sample generally. Given the limited sample of *moderate* to *high* risk offenders, it was predicted that some P factors would be linked to formulation mechanisms more frequently than others. In line with this hypothesis, three P factors were frequently linked to motivating mechanisms including P1 (Problems in Intimate Relationships), P7 (Personality Disorder), and P10 (Distorted Thinking about IPV) (see Table 5).

P factors linked to disinhibiting mechanisms are located in Table 6. Overall, P1 (Problems with Intimate Partner Relationships), P7 (Personality Disorder), and P10

(Distorted Thinking about IPV) were most frequently linked to disinhibitors, and often D5 through D8 (Lack of Insight, Guilt, Anxiety, and Empathy). Indeed, when P10 was rated as present, it was linked to D5 through D8 nearly or at 100% of the time. P7 (Personality Disorder) was frequently linked to other disinhibitors including D1 (Negative Attitudes) and D4 (Nihilism). Interestingly, P1 (Problems in Intimate Relationships) was the only P factor linked to all disinhibitors, but at varied frequencies. It was linked to D5 through D8 (Lack of Insight, Guilt, Anxiety, and Empathy) in approximately 25% of cases.

Table 5 Frequency of P Factors Linked to Motivators

	Distribution, Presence – Yes %							
SARA-V3 Factors	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
P1. Intimate Relatio…	4	50	12	92	20	78	0	70
P2. Non-Intimate…	1	0	0	6	2	6	0	2
P3. Employment…	0	0	16	4	0	8	0	2
P4. Trauma/Victimi…	0	2	0	0	0	4	0	4
P5. General Antisoc…	0	2	2	4	2	6	2	0
P6. Major Mental…	0	4	0	6	0	8	0	4
P7. Personality Dis…	0	26	6	52	0	52	2	40
P8. Substance Use	0	0	6	8	0	12	2	0
P9. Violent/Suici…	0	4	0	2	0	2	0	0
P10. Distorted Think…	4	50	14	98	24	92	4	72
Other	0	0	0	0	0	0	0	0

Notes. M1 = Defence/Distance/Protection, M2 = Justice/Honour/Retribution, M3 = Gain/Profit/Acquisition, M4 = Change/Control/Compliance, M5 = Status/Esteem/Dominance, M6 = Release/Expression/Emotion, M7 = Arousal/Activity/Excitement, M8 = Proximity/Affiliation/Conformity. Ratings were dichotomous (*yes, no*).

Finally, an examination of P factors linked to destabilizing mechanisms revealed P6 (Major Mental Illness), P7 (Personality Disorder), P8 (Substance Use), and P10 (Distorted Thinking about IPV) were more frequently linked to destabilizers that other P factors. P8 (Substance Use) was frequently linked to De4 (Impaired Reasoning) and De6 (Impulsive, Intrusive Thoughts). Unsurprisingly, P7 (Personality Disorder) was most frequently linked to De5 (Obsessive, Perseverative Thoughts). Although the frequency of Major Mental Disorder (P6) was lower in the current sample, when it was coded *Present,* raters linked it to De4 (Impaired Reasoning) and De6 (Obsessive, Perseverative

Thoughts) approximately 20% of the time. P10 (Distorted Thinking about IPV was frequently linked to De5 (Obsessive, Perseverative Thoughts) (88%) and De6 (Impulsive, Intrusive Thoughts) (86%). All frequencies for P factors linked to destabilizers are located in Table 7.

Table 6 Frequency of P Factors Linked to Disinhibitors

SARA-V3 Factors	Distribution, Presence – Yes %							
	D1	D2	D3	D4	D5	D6	D7	D8
P1. Intimate Relatio…	8	2	16	14	22	22	22	20
P2. Non-Intimate Rela…	0	0	26	2	0	0	0	0
P3. Employment/ Fin…	0	0	10	2	0	0	0	0
P4. Trauma/ Victimi…	6	2	2	4	2	0	0	0
P5. General Antisoc…	26	0	0	0	2	2	4	2
P6. Major Mental Dis…	0	4	8	8	10	6	10	8
P7. Personality Disorder	46	0	2	14	52	50	54	52
P8. Substance Use	2	0	2	6	10	6	82	6
P9. Violent/Suicidal Id…	0	2	0	24	0	0	0	0
P10. Distorted Think…	94	2	2	20	98	98	100	98
Other	0	0	0	0	0	0	0	0

Notes. D1 = Negative Attitudes, D2 = Negative Self-Concept, D3 = Alienation, D4 = Nihilism, D5 = Lack of Insight, D6 = Lack of Guilt, D7 = Lack of Anxiety, D8 = Lack of Empathy. Unspecified values are a result of zero variance. Ratings were dichotomous (*yes, no*).

Table 7 Frequency of P Factors Linked to Destabilizers

SARA-V3 Factors	Distribution, Presence – Yes %					
	De1	De2	De3	De4	De5	De6
P1. Intimate Relatio…	0	0	0	4	20	10
P2. Non-Intimate Rela…	0	0	0	0	0	0
P3. Employment/ Fin…	0	0	0	0	2	2
P4. Trauma/ Victimi…	0	0	0	0	4	0
P5. General Antisoc…	0	0	0	2	2	2
P6. Major Mental Dis…	2	6	0	20	14	22
P7. Personality Disorder	0	0	0	18	48	50
P8. Substance Use	2	2	0	70	28	76
P9. Violent/Suicidal Id…	0	0	0	4	6	6
P10. Distorted Think…	0	2	0	32	88	86
Other	0	0	0	0	0	0

Notes. De1 = Disturbed Attention and Concentration, De2 = Disturbed Sensation and Perception, De3 = Impaired Memory, De4 = Impaired Reasoning, De5 = Obsessive, Perseverative Thoughts, De6 = Impulsive, Intrusive Thoughts. Ratings were dichotomous (*yes, no*).

Research Question 2. What is the agreement between raters (i.e., IRR) of closed-ended forensic case formulation mechanism ratings?

Formulation Mechanisms

Motivators

All raters. Given the differences in training level between the Expert and Novice raters and the relatively challenging task of forensic case formulation, it was hypothesized that IRR across formulation mechanism ratings would fall in the moderate range and with some variability across mechanism ratings. As predicted, when all raters were included in the IRR analysis, coefficients for motivating mechanism ratings ranged from $AC_2 = .38$ to .98, or between *slight* and *almost perfect* agreement. The ratings for mechanism M7 (Arousal/Activity/Excitement) had the highest IRR among raters, which fell in the *almost perfect* range, $AC_2 = .98$, $p \leq .001$, 95%CI = [.96, 1.00]. IRR for M1 (Defence/Distance/ Protection) also fell in the almost perfect range, $AC_2 = .95$, $p \leq .001$, 95%CI = [.89, 1.00] as well as ratings for M3 (Gain/Profit/Acquisition), $AC_2 = .89$, $p \leq .001$, 95%CI = [.81, .97]. Raters demonstrated *substantial* agreement in their ratings for M5 (Status/Esteem/ Dominance). Ratings for four mechanisms were less reliable among raters. Rater

agreement for two mechanisms fell in the *fair* range. Raters had some difficulty reliably rating M4 (Change/Control/Compliance), $AC_2 = .59$, $p \leq .001$, 95%CI = [.24, .94] and M8 (Proximity/Affiliation/Conformity), $AC_2 = .47$, $p \leq .001$, 95%CI = [.25, .69]. On the final two mechanisms, raters had difficulty producing reliable ratings. With regard to M2 (Justice/Honour/Retribution), rater agreement was *slight,* $AC_2 = .33$, $p \leq .05$, 95%CI = [.08, .68]. Similarly, ratings for M3 (Release/Expression/Emotion) IRR were *slight*, $AC_2 = .39$, $p \leq .05$, 95%CI = [.06, .72]. Both mechanisms had wide confidence intervals, indicative of less reliable agreement. Results can be reviewed in Table 8. Results for additional IRR coefficients and percent agreement for all raters are presented in Appendix B, Table B19.

Expert versus Novice raters. In order to better understand the IRR of mechanism ratings, coefficients for Expert raters and Novice raters were calculated. It was expected that Experts would demonstrate higher rater agreement across their ratings for mechanisms than Novices. As there is not a well-established statistical test of the difference between two AC_2 coefficients, I evaluated differences between Experts and Novices by examining whether the 95% confidence intervals (CIs) of the AC_2 values observed in each group overlapped. As has been discussed elsewhere (e.g., Austin & Hux, 2002; Schenker & Gentleman, 2001), this is a very conservative procedure. When the 95%CIs of two statistical indexes do not overlap, then an inferential test of the difference between those indexes will be significant at $p < .05$; however, it is entirely possible that an inferential test of the difference between the indexes will be significant at $p < .05$ even when 95%CIs overlap to a limited extent. For this reason, it has been recommended to examine whether the 83%CIs of the indexes overlap. Below, when comparing the AC_2 values for Experts and Novices, I first considered the degree to which 95%CIs overlapped. If the 95%CIs appeared to have some, but limited overlap for ratings on a particular mechanism, I then calculated 83%CIs, assuming mechanism ratings with non-overlapping 83%CIs to be significantly different.

The IRR of Expert and Novice ratings for motivational formulation mechanisms – AC_2 values and their associated 95%CIs – is presented in Table 9. In general, IRR within each group was good: For Experts, AC_2 values for 5 of 8 ratings fell in the *substantial* or *almost perfect* ranges, and for Novices the corresponding figure was 4 of 8 ratings. For most of the ratings, the 95%CIs for Experts and Novices overlapped considerably. The largest difference was for M4 (Change/Control/Compliance). When I calculated the 83%CI for this mechanism, the intervals were overlapping and therefore I considered none of the

differences to be statistically significant. Additional IRR coefficients and percent agreement for Expert and Novice raters are presented in Appendix B, Tables B20 and B21.

Table 8 Reliability (AC_2) of Formulation Mechanisms – All Raters

Mechanism	AC_2 All Raters	Qualitative Interpretation	[95% CI]
Motivators			
M1. Defence/Distance…	.95***	Almost Perfect	[.89, 1.00]
M2. Justice/Honour/Retribution	.38*	Slight	[.08, .68]
M3. Gain/Profit/Acquisition	.89***	Almost Perfect	[.81, .97]
M4. Change/Control/Compli…	.59***	Fair	[.24, .94]
M5. Status/Esteem/Dominance	.77***	Substantial	[.62, .92]
M6. Release/Expression…	.39*	Slight	[.06, .72]
M7. Arousal/Activity/Excite…	.98***	Almost Perfect	[.96, 1.00]
M8. Proximity/Affiliation/Conf…	.47***	Fair	[.25, .69]
Disinihibitors			
D1. Negative Attitudes	.45	Slight	[-.18, 1.00]
D2. Negative Self-Concept	.93***	Almost Perfect	[.84, 1.00]
D3. Alienation	.71***	Moderate	[.50, .91]
D4. Nihilism	.86***	Substantial	[.76, .96]
D5. Lack of Insight	.92***	Substantial	[.84, 1.00]
D6. Lack of Guilt	.95***	Almost Perfect	[.90, .99]
D7. Lack of Anxiety	.95***	Almost Perfect	[.90, 1.00]
D8. Lack of Empathy	.94***	Almost Perfect	[.88, 1.00]
Destabilizers			
De1. Disturbed Attention…	.97***	Almost Perfect	[.91, 1.00]
De 2. Disturbed Sensation…	.93***	Almost Perfect	[.85, 1.00]
De3. Impaired Memory	.98***	Almost Perfect	[.93, 1.00]
De4. Impaired Reasoning	.10	Poor	[-.46, .66]
De5. Obsessive, Persever…	.67***	Moderate	[.49, .85]
D6e. Impulsive, Intrusive…	.74***	Substantial	[.60, .89]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977).

Summary

Rater agreement across motivating mechanisms was mixed. Raters demonstrated high agreement across four motivators, but only *slight* or *fair* agreement on the remaining four motivators. Unexpectedly, the IRR for Expert raters and Novice raters was similar across most motivating mechanism ratings.

Disinhibitors

All raters. Overall, disinhibiting mechanisms were rated somewhat more reliably than were motivating mechanisms (see Table 8). However, as predicted, there was still some variability in IRR. Across ratings for eight disinhibitors, four fell in the *almost perfect* range of rater agreement, two in the *substantial* range, and one in the *moderate* range. Ratings for one mechanism, D1 (Negative Attitudes), fell in the *slight* range and raters demonstrated little agreement on its presence in the sample, $AC_2 = .45$, $p > .05$, 95%CI = [-.18, 1.00]. Rater agreement fell in the *almost perfect* range for mechanisms D2 (Negative Self-Concept) and D5 through D8 (Lack of Insight, Guilt, Anxiety, Empathy). The ratings for these mechanisms fell in the *substantial* and *almost perfect* ranges and resulted in narrow confidence intervals; all were significant at $p \leq .001$. Ratings for D3 (Alienation) were the only to fall in the *moderate* range of agreement, $AC_2 = .71$, $p \leq .001$, 95%CI = [.50, .91]. Results for additional IRR coefficients and percent agreement for all raters are presented in Appendix B, Table B22.

Expert versus Novice raters. The IRR of Expert and Novice ratings for disinhibiting formulation mechanisms is also presented in Table 9. In general, IRR within each group was very good: For Experts, AC_2 values for 7 of 8 ratings fell in the *substantial* or *almost perfect* ranges, and for Novices the corresponding figure was 6 of 8 ratings. The 95%CIs for Experts and Novices overlapped considerably for most of the ratings, with the exception of those for D1 (Negative Attitudes), where the IRR for Experts was *poor* and that for Novices was *fair*. The 83%CIs did not overlap for D1 and therefore I considered the difference between Experts and Novices to be statistically significant. Upon inspection of the data, it was clear one Expert judged this mechanism as *yes* (present) or *possibly or partially present* most of the time, while the other Expert judged it *no* (not present) most of the time. Additional IRR coefficients and percent agreement for Expert and Novice raters are presented in Appendix B, Tables B23 and B24.

Table 9 **Reliability (AC_2) of Formulation Mechanisms – Expert and Novice Raters**

Mechanism	AC_2 Expert Raters	[95% CI]	AC_2 Novice Raters	[95% CI]
Motivators				
M1. Defence/Distance…	.87***	[.75, .98][b]	.96***	[.86, 1.00][a]
M2. Justice/Honour/Retribution	.33*	[.06, .60][e]	.39*	[.03, .76][e]
M3. Gain/Profit/Acquisition	.91***	[.83, .99][a]	.87***	[.59, 1.00][b]
M4. Change/Control/Comp…	.76***	[.59, .92][b]	.39	[-.19, .97][e]
M5. Status/Esteem/Dominance	.77***	[.61, .93][b]	.78***	[.60, .97][b]
M6. Release/Expression…	.43**	[.17, .69][d]	.41	[-.13, .95][e]
M7. Arousal/Activity/Excite…	.96***	[.88, 1.00][a]	.98***	[.94, 1.00][a]
M8. Proximity/Affiliation/Conf…	.44**	[.19, .68][d]	.40	[-.18, .97][e]
Disinihibitors				
D1. Negative Attitudes	.06	[-.27, .39][f]	.68***	[.33, 1.00][d]
D2. Negative Self-Concept	.97***	[.91, 1.00][a]	.88***	[.67, 1.00][b]
D3. Alienation	.87***	[.77, .97][b]	.65***	[.35, .96][d]
D4. Nihilism	.90***	[.81, .99][a]	.82***	[.60, 1.00][b]
D5. Lack of Insight	.95***	[.88, 1.00][a]	.90***	[.74, 1.00][b]
D6. Lack of Guilt	.95***	[.88, 1.00][a]	.94***	[.86, 1.00][a]
D7. Lack of Anxiety	.95***	[.88, 1.00][a]	.95***	[.82, 1.00][a]
D8. Lack of Empathy	.97***	[.92, 1.00][a]	.92***	[.80, 1.00][a]
Destabilizers				
De1. Disturbed Attention…	--	--	.94***	[.80, 1.00][a]
De 2. Disturbed Sensation…	.93***	[.84, 1.00][a]	.92***	[.83, 1.00][a]
De3. Impaired Memory	--	--	.95***	[.84, 1.00][a]
De4. Impaired Reasoning	-.34*	[-.64, -.04][f]	.18	[-.31, .66][f]
De5. Obsessive, Persever…	.70***	[.51, .89][c]	.70***	[.39, 1.00][c]
D6e. Impulsive, Intrusive…	.73***	[.56, .90][c]	.83***	[.60, 1.00][b]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Unspecified values are a result of zero variance. Ordinal weights applied to coefficient calculations; unconditional standard errors applied to Novice calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Summary

Overall, it was clear that all raters demonstrated agreement across disinhibiting mechanisms. The differences in coefficients between Experts and Novices tended to be small and with overlapping 95%CIs, thus most differences were not meaningful, with the exception of D1 (Negative Attitudes), in which Novice raters made more reliable ratings. The findings suggested that rater agreement for disinhibiting mechanisms was generally reliable regardless of experience.

Destabilizers

All raters. The IRR of ratings for destabilizing mechanisms ranged from *poor* to *almost perfect,* but as predicted, generally fell at or above the *moderate* range. Reliability coefficients for destabilizing mechanisms are presented in Table 8. Ratings for De1 (Disturbed Attention and Concentration), De2 (Disturbed Sensation and Perception), and De3 (Impaired Memory) fell in the *almost perfect* range; narrow confidence intervals indicated raters' robust agreement. When examining the IRR of ratings for De6 (Impulsive, Intrusive Thinking), agreement fell in the *substantial* range, $AC_2 = .74$, $p \leq .001$, 95%CI = [.60, .89]. Ratings for De5 (Obsessive, Perseverative Thinking) fell in the *moderate* range, $AC_2 = .67$, $p \leq .001$, 95%CI = [.49, .85]. With regard to ratings for De4 (Impaired Reasoning), raters had difficulty agreeing on the presence of this mechanism, $AC_2 = .10$, $p > .05$, 95%CI = [-.46, .66]. The IRR coefficient fell in the *poor* range. Upon inspection of the data, it was clear that raters could not agree on the presence of this destabilizer. Results for additional IRR coefficients and percent agreement for all raters are presented in Appendix B, Table B25.

Expert versus Novice raters. The IRR of Expert and Novice ratings for destabilizing formulation mechanisms is also presented in Table 9. IRR for Expert ratings of two destabilizing formulation mechanisms, De1 (Disturbed Attention and Concentration) and De3 (Impaired Memory), could not be calculated due to lack of variance. Inspection of the data revealed that Experts were in perfect agreement about the absence of these mechanisms in the cases. In general, IRR within each group was adequate to good: For Experts, AC_2 values for 1 of 4 rating fell in the *substantial* or *almost perfect* ranges, and for Novices the corresponding figure was 4 of 6 ratings. The 95%CIs for Experts and Novices overlapped considerably. The largest difference was for De4 (Impaired Reasoning), although the IRR for both Experts and Novices was *poor.* Even for this

mechanism, however, the 83%CIs overlapped slightly and therefore I considered none of the differences between Expert and Novice raters to be statistically significant. Additional IRR coefficients and percent agreement for Expert and Novice raters are presented in Appendix B, Tables B26 and B27.

Summary

Rater agreement across ratings for destabilizing mechanisms was moderate to high for both Experts and Novices. Interestingly, the mechanism with the poorest agreement across all motivators, disinhibitors, and destabilizers was De4 (Impaired Reasoning). The results indicated that both Experts and Novices were unable to agree on the presence of this mechanism, but in good agreement across other destabilizing mechanisms.

P Factors Linked to Formulation Mechanisms

All raters. When conducting SPJ violence risk assessments for real cases, users are encouraged to consider how present and relevant risk factors motivate, disinhibit, and destabilize the decision to engage in violence. In the current study, in addition to coding the presence of formulation mechanisms, raters simulated this process by indicating which P factors were linked to formulation mechanisms rated *yes* (present) and *possibly or partially present*. Given the training and experience-level differences between the raters and that the task of linking risk factors to formulation mechanisms is a higher-level process within forensic case formulation, raters were expected to achieve *fair* to *moderate* agreement with variability. Raters were surprisingly more reliable in their ratings than predicted. Across mechanisms, raters were in good agreement in terms of identifying linked P factors with motivators. For example, the IRR of P factors linked with M1 (Defence/Distance/Protection) ranged from AC_2 = .88 to .99 and fell in the *substantial* to *almost perfect* ranges of agreement. Similar levels of agreement resulted for ratings associated with M3 (Gain/Profit/Acquisition) and M7 (Arousal/Activity/Excitement). M2 (Justice/Honour/Retribution) linked P factor ratings were somewhat more variable; most fell in the *almost perfect* range, but P1 (Problems in Intimate Relationships) and P10 (Distorted Thinking about IPV) both fell in the *slight* range. A similar pattern was identified with M4 (Change/Control/Compliance), M6 (Release/Expression/Emotion), and M8 (Proximity/ Affiliation/Conformity) in which P1 and P10 ratings were again less reliable

than other linked P factors with this mechanism. The IRR coefficients of these ratings in regard to motivating mechanisms are located in Table 10.

IRR for linked P factors and disinhibitors was similar to that of motivators. Most ratings fell in the *moderate* to *almost perfect* ranges. Ratings for P1 (Problems in Intimate Relationships) across D5 to D8 (Lack of Insight, Guilt, Anxiety, and Empathy) were less reliable, ranging from AC_2 = .57 to .62 and *fair* to *slight* agreement. Reliability was somewhat more variable for P8 (Substance Use) across some disinhibitors. Rater agreement for P8 fell in the *slight* range when linked with D5 (Lack of Insight), AC_2 = .52, $p \leq .05$, 95%CI = [.00, 1.00] and in the *moderate* range when linked with D7 (Lack of Anxiety), AC_2 = .58, $p \leq .001$, 95%CI = [.38, .78]. Generally, raters demonstrated good agreement among linked P factors and disinihibitors. All IRR coefficients for linked P factors and disinihibitors can be reviewed in Table 11.

Raters were quite reliable when linking P factors and destabilizing mechanisms. Given the low frequency of *Present* ratings for some destabilizers, especially De1 (Disturbed Attention and Concentration), De2 (Disturbed Sensation and Perception), and De3 (Impaired Memory), many P factors are unspecified. Interestingly, despite the poor agreement for the presence of De4 (Impaired Reasoning), raters demonstrated *substantial* and *almost perfect* ranges of agreement across a number of linked P factors, including P1 (Problems in Intimate Relationships), P4 (Trauma/Victimization), P5 (General Antisocial Conduct), P6 (Major Mental Disorder), and P9 (Violent/Suicidal Ideation). Raters demonstrated *poor* agreement on ratings for linking P8 (Substance Use) to De4, AC_2 = .33, $p > .05$, 95%CI = [-.12, .78]. Similarly, raters had *poor* agreement in linked ratings for De4 and P10 (Problems in Intimate Relationships), AC_2 = .06, $p > .05$, 95%CI = [-.68, .80]. P1 (Problems in Intimate Relationships) was again problematic in linked ratings for De5 (Obsessive, Perseverative Thinking) and De6 (Impulsive, Intrusive Thinking), wherein raters demonstrated *poor* and *slight* IRR, respectively. All IRR ratings for linked P factors and destabilizing mechanisms are located in Table 12.

Expert versus Novice raters. Across all formulation mechanisms, Expert raters were expected to have higher rater agreement for linked P factors than Novice raters. Overall, given the good agreement across mechanisms and linked P factors, there were few considerable differences between the two rater groups. Examination of 83%CIs indicated there was a statistically meaningful difference between Experts and Novices

when linking P1 (Problems in Intimate Relationships) to D5 (Lack of Insight) and D6 (Lack of Guilt); Expert raters did not agree on if this P factor was linked to these disinhibitors, whereas Novice raters were in high agreement. Results for linked P factors across formulation mechanisms for Expert and Novice raters are presented in Appendix B, Tables B28 through B33.

Summary

Linking P factors to formulation mechanisms is a more advanced skill and thus raters were expected to demonstrate less agreement. However, the results indicated raters agreed which P factors were linked to formulation mechanisms as IRR coefficients generally fell in the *moderate* to *almost perfect* ranges. There were few differences between Expert and Novice raters.

Table 10 Reliability (AC_1) of Perpetrator Risk Factors Linked to Motivators – All Raters

	Motivating Mechanisms, AC_1 [95% CI]							
P Factors	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
P1. Intimate Relation…	.88*** [.75, 1.00][b]	.33* [.04, .61][e]	.87*** [.72, 1.00][b]	.59*** [.28, .89][d]	.70*** [.53, .87][c]	.29 [-.16, .74][c]	--	.47*** [.26, .68][d]
P2. Non-Intimate Re…	.97*** [.90, 1.00][a]	.95*** [.89, 1.00][a]	.96*** [.90, 1.00][a]	.95*** [.88, 1.00][a]	.99*** [.96, 1.00][a]	.94*** [.87, 1.00][a]	--	.91*** [.80, 1.00][a]
P3. Employment/ Fin…	--	.97*** [.90, 1.00][a]	.86*** [.74, .98][b]	.90*** [.78, 1.00][b]	--	.91*** [.83, 1.00][a]	--	.99*** [.96, 1.00][a]
P4. Trauma/ Victimi…	.99*** [.97, 1.00][a]	.96*** [.90, 1.00][a]	--	.97*** [.93, 1.00][a]	.98*** [.92, 1.00][a]	.90*** [.78, 1.00][b]	--	.95*** [.87, 1.00][a]
P5. General Antisoc…	.98*** [.95, 1.00][a]	.98*** .94, 1.00][a]	.98*** [.94, 1.00][a]	--	.96*** [.90, 1.00][a]	.92*** [.83, 1.00][a]	.99*** [.96, 1.00][a]	.98*** [.93, 1.00][a]
P6. Major Mental Dis…	.99*** [.96, 1.00][a]	.98*** [.93, 1.00][a]	--	.93*** [.85, 1.00][a]	.97*** [.90, 1.00][a]	.89*** [.80, .97][a]	--	.93*** [.85, 1.00][a]
P7. Personality Dis…	.95*** [.88, 1.00][a]	.63*** [.45, .81][c]	.94*** [.86, 1.00][a]	.62*** [.43, .81][c]	.84*** [.72, .95][b]	.45** [.16, .73][d]	.99*** [.96, 1.00][a]	.69*** [.53, .86][c]
P8. Substance Use	.98*** [.94, 1.00][a]	.97*** [.90, 1.00][a]	.95*** [.89, 1.00][a]	.91*** [.82, 1.00][a]	.99*** [.97, 1.00][a]	.71*** [.35, 1.00][c]	.99*** [.96, 1.00][a]	.98*** [.93, 1.00][a]
P9. Violent/Suicidal…	--	.97*** [.91, 1.00][a]	--	.97*** [.93, 1.00][a]	--	.94*** [.88, 1.00][a]	--	--
P10. Distorted Think…	.89*** [.76, 1.00][b]	.33* [.07, .59][e]	.82*** [.68, .95][b]	.63*** [.30, .95][d]	.60*** [.34, .87][d]	.44* [.10, .78][e]	.97*** [.92, 1.00][a]	.53*** [.28, .77][d]

Notes. N = 50. M1 = Defence/Distance/Protection, M2 = Justice/Honour/Retribution, M3 = Gain/Profit/Acquisition, M4 = Change/Control/Compliance, M5 = Status/Esteem/Dominance, M6 = Release/Expression/ Emotion, M7 = Arousal/Activity/Excitement, M8 = Proximity/Affiliation/Conformity. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table 11 Reliability (AC_1) of Perpetrator Risk Factors Linked to Disinhibitors – All Raters

	Disinhibiting Mechanisms, AC_1 [95% CI]							
P Factors	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
P1. Intimate Relatio…	.86*** [.71, 1.00][b]	.94*** [.81, 1.00][a]	.73*** [.48, .97][c]	.86*** [.74, .98][b]	.57 [-.11, 1.00][e]	.59 [-.03, 1.00][d]	.62* [.01, 1.00][e]	.59 [-.09, 1.00][e]
P2. Non-Intimate Re…	--	.94*** [.83, 1.00][a]	.62*** [.41, .84][c]	.96*** [.91, 1.00][a]	--	--	--	--
P3. Employment/ Fin…	--	.97*** [.88, 1.00][a]	.83*** [.63, 1.00][c]	.97*** [.91, 1.00][a]	--	--	--	--
P4. Trauma/ Victimi…	.83*** [.60, 1.00][f]	.98*** [.92, 1.00][a]	.96*** [.91, 1.00][a]	.98*** [.94, 1.00][a]	.88*** [.68, 1.00][b]	.97*** [.92, 1.00][a]	.97*** [.92, 1.00][a]	.97*** [.92, 1.00][a]
P5. General Antisoc…	.61*** [.41, .81][e]	.95*** [.83, 1.00][a]	.95*** [.90, 1.00][a]	--	.87*** [.75, .98][b]	.87*** [.75, .99][b]	.84*** [.69, .99][b]	.82*** [.65, 1.00][b]
P6. Major Mental…	.98*** [.95, 1.00][a]	.94*** [.87, 1.00][a]	.93*** [.84, 1.00][b]	.90*** [.80, 1.00][a]	.91*** [.83, .99][a]	.94*** [.87, 1.00][a]	.89*** [.79, .99][a]	.94*** [.87, 1.00][a]
P7. Personality Dis…	.50*** [.26, .74][d]	.92*** [.79, 1.00][a]	.88*** [.73, 1.00][b]	.88*** [.79, .97][a]	.75*** [.60, .90][b]	.79*** [.65, .92][b]	.75*** [.58, .91][b]	.75*** [.62, .89][b]
P8. Substance Use	.98*** [.94, 1.00][a]	--	.97*** [.92, 1.00][a]	.94*** [.88, 1.00][a]	.52* [.00, 1.00][e]	.76*** [.55, .97][c]	.58*** [.38, .78][c]	.81*** [.68, .96][b]
P9. Violent/Suicidal…	.99*** [.97, 1.00][a]	.98*** [.95, 1.00][a]	.99*** [.96, 1.00][a]	.75*** [.59, .90][b]	.96*** [.88, 1.00][a]	.95*** [.87, 1.00][a]	.96*** [.88, 1.00][a]	.94*** [.87, 1.00][a]
P10. Distorted Think…	.58* [.01, 1.00][f]	.94*** [.79, 1.00][a]	.94*** [.86, 1.00][a]	.86*** [.73, .98][b]	.95*** [.89, 1.00][a]	.95*** [.90, 1.00][a]	.91*** [.83, 1.00][a]	.93*** [.87, 1.00][a]

Notes. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. D1 = Negative Attitudes, D2 = Negative Self-Concept, D3 = Alienation, D4 = Nihilism, D5 = Lack of Insight, D6 = Lack of Guilt, D7 = Lack of Anxiety, D8 = Lack of Empathy. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table 12 Reliability (AC_1) of Perpetrator Risk Factors Linked to Destabilizers – All Raters

	Destabilizing Mechanisms AC_1 [95% CI]					
P Factors	**De1**	**De2**	**De3**	**De4**	**De5**	**De6**
P1. Intimate Relatio…	--	.98*** [.93, 1.00][a]	--	.91*** [.70, 1.00][b]	.37 [-.26, 1.00][f]	.65 [-.02, 1.00][e]
P2. Non-Intimate Re…	--	--	--	--	.97*** [.91, 1.00][a]	--
P3. Employment/ Fin…	--	--	--	--	.97*** [.91, 1.00][a]	.99*** [.96, 1.00][a]
P4. Trauma/ Victimi…	--	--	--	.98*** [.94, 1.00][a]	.97*** [.92, 1.00][a]	.99*** [.96, 1.00][a]
P5. General Antisoc…	--	--	--	.98*** [.94, 1.00][a]	.99*** [.96, 1.00][a]	.96*** [.89, 1.00][a]
P6. Major Mental Dis…	.98*** [.94, 1.00][a]	.97*** [.92, 1.00][a]	--	.86*** [.75, .97][b]	.82*** [.65, 1.00][b]	.77*** [.57, .97][c]
P7. Personality Dis…	--	.97*** [.91, 1.00][a]	--	.52*** [.25, .80][d]	.61*** [.37, .85][c]	.62*** [.47, .78][c]
P8. Substance Use	.99*** [.96, 1.00][a]	.93*** [.86, 1.00][a]	.99*** [.96, 1.00][a]	.33 [-.12, .78][f]	.36* [.06, .66][e]	.63*** [.44, .81][c]
P9. Violent/Suicidal…	--	--	--	.92*** [.85, .99][a]	.94*** [.87, 1.00][a]	.91*** [.80, 1.00][b]
P10. Distorted Think…	--	.95*** [.88, 1.00][a]	--	.06 [-.68, .80][f]	.51* [.11, .91][e]	.10 [-.61, .80][f]

Notes. N = 50. De1 = Disturbed Attention and Concentration, De2 = Disturbed Sensation and Perception, De3 = Impaired Memory, De4 = Impaired Reasoning, De5 = Obsessive, Perseverative Thoughts, De6 = Impulsive, Intrusive Thoughts. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Research Question 3. What is the similarity of narrative forensic case formulations?

It is exceedingly challenging to accurately characterize narratives in numeric forms and examine IRR on that number. Thus, to further understand the IRR of forensic case formulations, RAs in Part 2 completed a series of similarity ratings after reviewing paired narrative formulations (see Appendix A.1). The degree of similarity should allow me to capture the magnitude of similarity, or consistency, between ratings. Formulations were paired in two primary ways, by formulation type and by expertise type. The formulation type pairing was *Within Case* or *Across Case* (*Within Case* = two formulations from the same case; *Across Case* = two formulations from different cases). The expertise type pairing was by Expert-Expert, Novice-Novice, and Expert-Novice. Two formulations paired from the same case were expected to have a high degree of similarity, whereas two formulations paired from two different cases were expected to have a low degree of similarity. Also of interest was how rater expertise impacted the judgments of similarity between formulation pairs *Within* and *Across Cases*. Similar to the analytic strategy in Wilson (2013), the subsequent analyses were organized as follows: first by a test of association to understand if a relationship among the variables existed, then, if appropriate, a primary test of a one-way multivariate analysis of variance MANOVA examining the formulation pairing type, then, a secondary test of interaction effects of Formulation Type x Rater Expertise. Finally, when omnibus analyses were significant, univariate analyses of variance (ANOVAs) were conducted to better understand mean differences. Effect sizes are indicated by multivariate η^2 and interpreted per Cohen (1988): η^2 = .01, small effect size; η^2 = .06, medium effect size; η^2 = .14, large effect size.

Overall Similarity

One similarity item was used to judge the overall probability that the two paired formulations were from the same case or different cases. As expected, the mean *Within Case* probability ratings (*M* = 59.12, *SD* = 27.52) were significantly higher than *Across Case* ratings (*M* = 25.95, *SD* = 25.64), $F(1, 140) = 55.29$, $p \leq .001$, multivariate $\eta^2 = .28$. Given this result, it is clear that formulations written for the same case were judged with higher similarity ratings than paired formulations written for different cases.

Three other ratings were examined simultaneously to better understand overall similarity including overall similarity of case formulations, most likely scenario for future violence, and risk management strategies. Mean-interitem correlation (MIC) results indicated a large association between these items (MIC = .68). Thus, MANOVA analyses were conducted to determine whether there was a main effect for formulation type (*Within Case, Across Case)* with respect to overall case formulation, most likely scenario for future violence, and risk management plans. As predicted, an omnibus test for group differences based on formulation type was significant, Wilk's Λ = .57, $F(3, 106) = 27.10$, $p \leq .001$, multivariate η^2 = .43. In addition, a test for an interaction for Formulation Type x Rater Expertise was conducted. Results were not significant, Wilk's Λ = .96, $F(6, 212) = .77$, $p \geq .05$, multivariate η^2 = .02.

Given there was a main effect for formulation type, a post hoc ANOVA was conducted to better understand the result. A Bonferroni corrected $p \leq .017$ was used to interpret statistical significance to account for the multiple comparisons. All pairwise comparisons were significant (see Table 13). As predicted, overall case formulation similarity ratings were higher for *Within Case* pairs, $F(1, 140) = 73.71$, $p \leq .001$, multivariate η^2 = .35. This result suggested RAs considered paired formulations derived from the same case to be more similar than paired formulations from different cases on global indicators of overall similarity for formulations, scenarios, and risk management strategies.

Table 13 Within and Across Case Comparisons – Overall Similarity Ratings

	Within Case *M* (*SD*)	Across Case *M* (*SD*)	*F*	*p*	Effect Size η^2
Case Formulations	51.62 (20.27)	24.86 (16.82)	73.71	≤.001	.35
Most Likely Future Scenario	63.57 (31.65)	29.83 (35.31)	29.02	≤.001	.20
Risk Management Strategies	44.00 (22.62)	29.26 (20.54)	15.49	≤.001	.11

In addition, the overall similarity ratings for most likely scenario was significant, $F(1, 113) = 29.02$, $p \leq .001$, multivariate η^2 = .20. When paired formulations were from the same case, mean similarity ratings for the most likely scenario were 63.57 (*SD* = 31.65) compared with formulations across cases, *M* = 29.83 (*SD* = 35.31). Results suggested that scenario planning for paired formulations from the same case were more similar than

across cases. Finally, the overall similarity of risk management strategies was also significant, $F(1, 131) = 15.49$, $p \leq .001$, multivariate $\eta^2 = .11$. As anticipated, paired case formulations from the same case had higher similarity ratings with regard to stated risk management strategies than paired formulations from different cases.

Similarity of Formulation Mechanisms

Three ratings were used to evaluate the overall similarity of motivators, disinhibitors, and destabilizers in forensic case formulations. The MIC for these three items was .60. Given this large association, MANOVA analyses were conducted to determine if a main effect for formulation type with respect to motivators, disinihibitors, and destabilizers. As predicted, an omnibus test for group differences based on formulation type (*Within Case, Across Case)* was significant, Wilk's $\Lambda = .73$, $F(3, 125) = 15.48$, $p \leq .001$, multivariate $\eta^2 = .27$. In addition, an omnibus test for an interaction for Formulation Type x Rater Expertise and was also significant, Wilk's $\Lambda = .90$, $F(6, 250) = 2.20$, $p \leq .05$, multivariate $\eta^2 = .05$.

Given there was a main effect for formulation type, a post hoc ANOVA was conducted to better understand the result. A Bonferroni correction of $p \leq .017$ was used to interpret significance. After this correction, pairwise differences for *Within Case* and *Across Case* similarity ratings and formulation mechanisms were significant (see Table 14). Mean differences for *Within Case* and *Across Case* pairs were in the expected direction—paired formulations from the same case were judged as more similar with regard to the motivating, disinhibiting, and destabilizing mechanisms than paired formulations from different cases.

Table 14 Within and Across Case Comparisons – Case Formulation Mechanisms

	Within Case *M* (*SD*)	Across Case *M* (*SD*)	*F*	*p*	Effect Size η^2
Motivators	53.09 (26.39)	27.40 (22.05)	39.55	≤.001	.22
Disinhibitors	57.62 (27.87)	39.15 (23.77)	17.13	≤.001	.12
Destabilizers	52.62 (25.82)	30.68 (24.34)	26.36	≤.001	.16

With regard to the Formulation Type x Rater Expertise interaction, an examination of these comparisons showed a significant interaction between Expert-Expert formulation pairings *Within* and *Across Cases* for disinhibiting mechanisms, but not for other rater pairings (Novice-Novice, Expert-Novice) or formulation mechanisms. A test of simple main effects revealed paired Expert-Expert *Within Case* narrative formulations were rated more alike than Novice-Novice or Expert-Novice *Within Case* pairs with regard to disinhibiting mechanisms. On average, RAs rated Expert-Expert discussions of disinhibitors *Within Case* formulations 37.5 percentage points more similar than *Across Case* pairs (see Table 15). Ultimately, these results suggested Expert-Expert paired narrative formulations were driving the difference in similarity ratings *Within Case* and *Across Case* for disinhibiting mechanisms.

Table 15 Simple Main Effects for Rater Pairing Type Within and Across Cases – Disinhibiting Similarity Ratings

	Within Case *M*	Across Case *M*	Mean Difference	*p*
Expert-Expert	68.75	31.25	37.50	≤ .001
Novice-Novice	55.83	42.08	13.75	.063
Expert-Novice	51.74	44.35	7.39	.325

Similarity of Risk Management Strategies

Four items were used to understand the similarity of risk management strategies including monitoring, treatment, supervision, and victim safety planning strategies. When correlated, these items had a moderate to large association, MIC = .43. As a result, MANOVA analyses were conducted to determine if a main effect for formulation type (*Within Case, Across Case)* with respect to risk management strategies (monitoring, treatment, supervision, victim safety). As hypothesized, an omnibus test for significance between *Within Case* and *Across Case* similarity ratings for risk management strategies was significant, Wilk's Λ = .80, $F(4, 60)$ = .3.70, $p \leq .001$, multivariate η^2 = .20. An interaction for Formulation Type x Rater Expertise and risk management strategies was conducted, but was not significant, Wilk's Λ = .96, $F(8, 120)$ = .31, $p \geq .05$, multivariate η^2 = .02.

Univariate analyses to further understand the *Within Case* and *Across Case* group similarity ratings for management strategies were conducted. A Bonferroni correction of $p \leq .013$ was used to interpret significance. Mean similarity ratings for three of four risk management strategies were significantly different, including monitoring, treatment, and supervision strategies. These results indicated that paired formulations for the same case were judged to have more similar monitoring, treatment, and supervision strategies than paired formulations across different cases. Notably, the effect sizes across each were relatively small. Similarity ratings for victim safety strategies were not significant, $F(1, 78) = 4.78$, $p = .03$, multivariate $\eta^2 = .06$. Results for univariate analyses can be reviewed in Table 16.

Table 16 Within and Across Case Comparisons – Risk Management Strategies

	Within Case *M* (*SD*)	Across Case *M* (*SD*)	*F*	*p*	Effect Size η^2
Monitoring	42.80 (25.95)	29.43 (24.53)	7.22	.008	.07
Treatment	45.38 (25.74)	31.62 (24.47)	10.00	.002	.07
Supervision	49.84 (25.49)	32.69 (24.414)	15.36	≤.001	.12
Victim Safety	41.03 (26.14)	28.29 (25.97)	4.78	.032	.06

Summary

Overall, the results of these analyses indicated that paired formulations from the same case were rated as more similar than paired formulations from different cases across a number of indicators. These results were established for overall similarity judgments as well as for specific aspects of the formulations including the type of motivating, disinhibiting, and destabilizing mechanisms selected, scenario planning, and risk management strategies. These findings provide support for the consistency of forensic case formulations in the current sample.

Research Question 4. What is the quality of narrative forensic case formulations?

As forensic case formulations were found to have good IRR, an evaluation of the quality of formulations was conducted. Formulations were expected to be moderate to high in quality overall in light of the study-specific training completed by all raters, but higher for Experts than Novices. To address this issue, three RAs independently rated a total of 127 formulations for 36 cases using the CFQC-R: 57 were Expert formulations (for 21 cases rated by 2 Experts and 15 cases rated by 1 Expert) and 70 were Novice formulations (34 cases rated by 2 Novices and 2 cases rated by 1 Novice). The most appropriate way to analyze these data would have been to use multilevel modeling; however, this was deemed infeasible due to the relatively small number of cases. Instead, I calculated grand mean ratings for each of the 10 CFQC-R items and the CFQC-R total scores, averaged across (a) two different RAs for (b) formulations made by at least three different raters (1 or 2 Experts and 1 or 2 Novices) and for (c) 36 cases. The findings are presented in the second column of Table 17. The grand mean ratings for CFQC-R items were generally high, ranging from a low of 7.96 to a high of 9.35 (out of a maximum of 10 for all items); perhaps most relevantly, the grand mean rating for CFQC-R Item 10 (Overall Quality) was 8.03 (out of a maximum of 10). Finally, the grand mean for the CFQC-R total scores was 84.68 (out a maximum of 100). Overall, the pattern of findings suggests that the raters made case formulations of good quality.

Next, I examined the difference in quality between Expert and Novice formulations. To do this, I calculated grand mean ratings for each of the 10 CFQC-R items and the CFQC-R total scores separately for Experts and Novices, averaged across (a) two different RAs for (b) formulations made by 1 or 2 raters and for (c) 36 cases. The findings are presented in the third and fourth columns of Table 17. As expected, Expert formulations received higher CFQC-R ratings than did Novice formulations on 8 of 10 items, including Item 10 (Overall Quality). In addition, the CFQC-R total score was higher for formulations made by Experts than those made by Novices.

Table 17 CFQC-R – Descriptive Statistics

CFQC-R Item	All Raters *M (SD)*	Experts *M (SD)*	Novices *M (SD)*
1. Narrative	8.45 (.48)	8.68 (.73)	8.28 (.64)
2. External Coherence	8.44 (.64)	8.34 (1.08)	8.57 (.83)
3. Factual Foundation	8.37 (.64)	9.23 (.62)	7.74 (.98)
4. Internal Coherence	9.35 (.41)	9.44 (.60)	9.28 (.57)
5. Completeness	8.35 (.76)	9.33 (.55)	7.63 (1.16)
6. Events are understood by the way they relate over time	8.53 (.72)	9.55 (.41)	7.78 (1.13)
7. Simplicity	8.84 (.43)	8.81 (.80)	8.86 (.55)
8. Predictive	8.34 (1.05)	9.39 (.43)	7.57 (1.65)
9. Action Oriented	7.96 (1.13)	9.07 (.71)	7.14 (1.71)
10. Overall Quality	8.03 (.77)	8.88 (.55)	7.43 (1.17)
Total Score	**84.68 (6.00)**	**90.77 (4.62)**	**80.23 (9.17)**

Notes. All raters, N = 36 cases; N = 127 formulations, n = 57 Expert, n = 70 Novice.

As formulations were made for the same 36 cases, I conducted paired-sample *t* tests to provide a crude evaluation of the statistical significance and magnitude of the differences between Expert and Novice raters. The results of these analyses are presented in Table 18. As the Table indicates, the Expert formulations received higher scores than did Novice formulations on 7 of 10 CFCQ-R items, including Item 10 (Overall Quality). For one of the remaining items, Experts received higher scores than did Novices but the difference was not statistically significant; and the other two remaining items, Novices received higher scores than did Experts but the differences were not statistically significant. Looking at the tests for the 7 items for which the differences between Experts and Novices were statistically significant, 6 had effect sizes were very large (Cohen's $d \geq 1.00$). Overall, then, although the formulations were generally of good quality, the analyses strongly suggest that Experts generated formulations that were higher in quality than those of Novices.

Table 18 Paired-Sample *t* Test Results, Expert Versus Novice CFQC-R Scores

CFQC-R Item	Mean Diff. *(SD)*	*t*	*p*	Cohen's *d*
1. Narrative	.41 (.1.01)	2.44	.02	.41
2. External Coherence	-.23 (1.40)	-.97	.34	.16
3. Factual Foundation	1.49 (1.10)	8.11	≤.001	1.48
4. Internal Coherence	.16 (.83)	1.18	.25	.19
5. Completeness	1.70 (1.19)	8.59	≤.001	1.43
6. Events are understood by the way they relate over time	1.77 (1.19)	8.91	≤.001	1.49
7. Simplicity	-.05 (1.06)	-.26	.80	.05
8. Predictive	1.82 (1.72)	6.34	≤.001	1.06
9. Action Oriented	1.93 (1.75)	6.62	≤.001	1.10
10. Overall Quality	1.45 (1.22)	7.14	≤.001	1.19
Total Score	**10.51 (9.89)**	**6.37**	**≤. 001**	**1.06**

Notes. N = 36 cases; *N* = 127 formulations, *n* = 57 Expert, *n* = 70 Novice.

Chapter 4. Discussion

The theoretical literature establishing the necessity of forensic case formulations is mounting, but empirical evidence is still insufficient. The aim of the current study was to contribute to this much needed area in the empirical literature. This was the first study to evaluate the IRR of case formulations in cases of IPV within the SPJ and Decision Theory approach to forensic case formulation. Expert and Novice raters were used to examine potential differences that experience may have on forensic case formulation skills. First, raters used a closed-ended coding scheme to indicate Presence ratings for formulation mechanisms (e.g., motivators, disinhibitors, destabilizers), which was more structured in nature. Raters also completed narrative case formulations that were more unstructured. The distributions of Presence ratings and the IRR of mechanism ratings was examined. Next, a more thorough analysis of narrative forensic case formulations was conducted. Using a paired *Within Case* and *Across Case* design, the similarity of paired formulations was examined. In addition, the quality of formulations were evaluated using the CFQC-R. Ultimately, the results of this study demonstrated that Presence ratings for most formulation mechanisms were skewed *yes* (present) or *no* (not present). In addition, results suggested that Expert and Novice raters in the current study could reliably formulate cases. Generally, expertise did not impact IRR; raters were in agreement regardless of expertise. The results of similarity ratings when comparing paired *Within Case* and *Across Case* narrative formulations was as expected: formulation pairs from the same case were judged as more alike than formulation pairs from different cases. The results indicated expertise did have more of an impact with regard to writing a narrative case formulation, which required more subjective judgment. Expert raters were better able to explain the etiology of violence in the formulation. Finally, formulations were of high quality, but as predicted, Expert raters produced higher quality formulations than Novice raters. Overall, the results of this study suggested that there is evidence for the use of SARA-V3 and the Decision Theory approach to forensic case formulation as they appeared to be reliable, individualized, and of high quality.

Distribution of Formulation Mechanism Ratings

The distribution of Presence ratings was skewed for some motivating, disinhibiting, and destabilizing formulation mechanisms. It is likely that this distribution pattern can be

attributed to the focus on a homogeneous type of violence in the current sample. In addition, the sample of offenders was homogenous. The frequency of personality disorder or problematic personality disorder traits was high. In addition, few offenders in the sample suffered from major mental illnesses such as psychotic disorders. Given the sample, certain pathways to violence occurred much more frequently than others. Motivators such as M4 (Change/Control/Compliance), M6 (Release/Expression/Emotion), and M8 (Proximity/ Affiliation/Conformity) are theoretically related to problems linked with personality disorder, including symptoms of Borderline Personality Disorder and other problems with interpersonal attachment. Thus, the high presence of these motivators compared to a motivator such as Gain/Profit/Acquisition, is fitting given the sample.

The limited sample also had implications for the distribution of disinhibiting and destabilizing mechanisms. The distributions of Presence ratings for D5 (Lack of Insight), D6 (Lack of Guilt), D7 (Lack of Anxiety), and D8 (Lack of Empathy) were skewed *Present* at extremely high rates. D1 (Negative Attitudes) was also skewed *yes* and *possibly or partially present.* One possible explanation is that because of the high rate of personality disorder in the sample, these disinhibiting mechanisms align with that clinical presentation. When taken into consideration with the skewed destabilizing mechanisms, the picture becomes clearer. De5 (Obsessive, Perseverative Thinking) and De6 (Impulsive, Intrusive Thinking) were both skewed *yes.* Again, these mechanisms logically fit with a sample that is suffering from personality disorder rather than major mental illness.

The prevalence of personality disorder as a primary explanatory factor for violence in the current sample is also evident in the distribution of ratings for linked P factors to formulation mechanisms. Across motivators, disinhibitors, and destabilizers, results show that personality disorder was frequently linked to present mechanisms, in addition to distorted thinking about IPV, which could arguably also be linked to personality disorder. When considering theories of IPV, these results may generalize. For example, the Duluth Model (Pence & Paymar, 1993), in which feminist theory is prominent, highlights issues of societal patriarchy that result in male dominance and the need for power and control over women. Additionally, attachment theory (Bowlby, 1988) also informs the link between personality disorder and IPV (Doumas et al., 2008; McClellan & Killeen, 2000).

IRR - Formulation Mechanism Ratings

Overall, raters agreed on which mechanisms were present or not when formulating cases. There were some differences in the strength of agreement when Expert and Novice ratings were isolated, but across most mechanisms these differences were minimal and not meaningfully different upon examination of overlapping CIs. Considering all mechanism ratings together, it was clear that raters had moderate to good IRR across approximately 75% of formulation mechanism ratings. With regard to mechanisms with lower IRR, there are a number of reasons raters may have had lower agreement. It is possible some mechanisms were just more difficult to code. In general, forensic case formulation is not an easy skill, and some mechanisms may be more difficult to understand when attempting to explain the etiology of violence. It is also possible some mechanisms were just not good explanations of violence generally *or* that the coding manual did a poor job of providing accurate descriptions of some mechanisms. In addition, the file information to reliably code some mechanisms could have been missing or poor. Finally, it was also possible the training related to some items could have been poor. Despite this, given the difficulty of the task, that around 75% of ratings had adequate IRR is still rather impressive. There is a lack of prior research and measurement related to this specific process of forensic case formulation. The results suggested that if we have good case information, train raters well, effectively clarify and operationalize concepts related to Decision Theory, then raters can reliably formulate forensic cases.

When considering IRR and rater type, Experts and Novices did not demonstrate meaningfully different levels of agreement in their ratings on many mechanisms. Even with a relatively junior set of raters and the fact that they were asked to make relatively complex judgments, ratings were made with some real consistency. Follow-up analyses examining overlapping CIs did result in a few differences that may be attributed to expertise. For example, Expert agreement was higher than Novices for M4 (Change/Control/ Compliance), although examination of an 83% CI indicated the difference was likely not meaningful. Still, the Expert IRR coefficient was higher and had a smaller CI. This mechanism can theoretically be linked to personality disorder. It is possible that the two Expert raters, both trained to diagnose personality disorders, had a much better understanding of the presence of these mechanisms than Novices, most of whom were not clinically trained. Again, the “difficulty” of this mechanism may have been discrepant between Experts and Novices. Novice raters were in more agreement with regard to the

presence of D1 (Negative Attitudes) than Experts. There was a true lack of agreement about what this mechanism was trying to capture between Experts, whereas Novices were more reliable in their judgments. All raters disagreed on the presence of De4 (Impaired Reasoning). During consensus meetings, it became clear that there was a lack of understanding among all raters as to what this mechanism was attempting to capture, and this lack of understanding was not reconciled.

With regard to linked P factors and formulation mechanism ratings, rater agreement was good and higher than expected. This is a relatively difficult skill and given the level of rater agreement, it is clear both Experts and Novices were able to understand how perpetrator risk factors impact the decision to engage in violence via motivating, disinhibiting, and destabilizing mechanisms. Across some mechanism ratings, raters were in less agreement about linking mechanisms to P1 (Problems in Intimate Relationships) and P10 (Distorted Thinking about IPV). Reflecting on the consensus process and the overall high agreement for the past and recent presence of these factors, it is unclear why raters were in disagreement. When considering motivators, raters did not agree on linking P1 and P10 to M4 (Change/Control/Compliance), M6 (Release/ Expression/Emotion), and M8 (Proximity/Affiliation/Conformity). This result is perplexing given that relationship and attachment problems are likely central to these motivators. In addition, raters did not agree if P1 was linked to D5 to D8 (Lack of Insight, Guilty, Anxiety, and Empathy). Upon inspection of the coefficients, this disagreement was driven by the Expert raters and not the Novices. It is clear one Expert rater identified P1 to be central to explaining these disinhibitors, while the other did not. Given the lack of agreement overall for De4 (Impaired Reasoning), inspection of the linked P factors could aid in understanding the low IRR of this mechanism. It appears two P factors, P8 (Substance Use) and P10 (Distorted Thinking about IPV), were not reliably linked. Anecdotally, during consensus meetings, there was disagreement about the intent of De4. Some raters considered this mechanism as secondary to substance intoxication or distorted thinking, whereas others disagreed with this conceptualization and considered this mechanism to be related primarily to symptoms of major mental illness such as psychosis. Again, despite the lack of rater agreement on these few linked P factors to formulation mechanisms, overall Experts and Novices were reliable in their judgments.

A clear pattern emerged in the results of various other coefficients included in the supplemental analyses. Percent agreement across most mechanisms was high. Brennan

and Prediger's coefficient tended to fall closer to Gwet's AC and percent agreement, all of which indicated raters were generally reliable in their judgements of formulation mechanisms and linked P factors. Reliability coefficients including Cohen/Conger *κ,* Scott/Fleiss' *κ,* and Krippendorff's α estimated rater agreement to be much lower in some cases. Overall, percent agreement, Gwet's AC, and Brennan and Prediger tended to produce higher IRR coefficients than Cohen/Conger *κ,* Scott/Fleiss' *κ,* and Krippendorff's α*.* This type of discrepancy between the IRR coefficients has been explored empirically (Zhao et al., 2013; Zhao et al., 2018) and many factors including the number of raters, number of categories, "difficulty" of the item, and marginal distributions of ratings can have significant impact on IRR coefficients. In the current sample, it is possible that two primary issues are impacting the discrepancy in IRR coefficients. First, the uneven distributions of formulation mechanism ratings may have resulted in the kappa paradox (Feinstein & Cicchetti, 1990) which impacted certain coefficients more (Cohen/Conger *κ,* Scott/Fleiss' *κ,* and Krippendorff's α*).* In addition, the difficulty of rating an item, or in this case a mechanism, can impact both Gwet's AC and Brennan and Prediger. When mechanisms were "easier" to rate, research has shown the IRR coefficient may be artificially inflated (Zhao et al., 2013).

Given the disparity in the various IRR coefficients for formulation mechanisms in the current sample, it is possible that researchers in the field should consider how we present IRR statistics in our work. One would likely come to entirely different conclusions about the level of agreement in the current study if I had presented only Cohen's *κ.* Given the lack of consensus around which IRR coefficients most accurately estimate and describe rater agreement, the field could consider standards such as reporting multiple coefficients. Additionally, some discussion around psychological research and the most appropriate coefficient given our methods compared to other sciences and medicine could be beneficial. Ultimately, in the current study, Generalizability Theory analyses could help identify the source of disagreement amongst raters to better understand IRR.

Formulation Similarity

Results indicated that two paired formulations from the same case were judged as more similar than two paired formulations from different cases across a number of variables. This design and the results provided evidence that raters wrote individualized

formulations that were specific to the case, rather than providing generalized or unspecific explanations of violence. In addition, results also suggested raters thought carefully about the specific needs of the individual offender and recommended risk management strategies that matched the needs. This result provided evidence that the Decision Theory method of forensic case formulation had utility when formulating IPV in the current study. Interestingly, interactions between pair type (*Within Case, Across Case*) and rater type (Expert-Expert, Novice-Novice, Expert-Novice) were not significant with the exception of mean differences for formulation mechanisms. This suggested that both Experts and Novices were equally similar or consistent in their narrative formulations, but that Experts may have provided better explanations of the etiology of violence.

With regard to comparisons *Within Case* and *Across Case,* paired formulations and overall similarity, RAs judged *Within Case* pairs to be much more alike than *Across Case* pairs. In addition, three similarity items measuring global aspects of the formulations were combined. These items included judgments for overall similarity of case formulations, the most likely future scenario, and risk management strategies. Again, findings indicated that RAs judged *Within Case* paired formulations as more similar than *Across Case* paired formulations. An examination of univariate results showed that across all three global ratings, *Within Case* pairs had higher mean similarity ratings with small to medium effect sizes, the largest effect being for the overall case formulation ratings. Given that there were many similarities across the cases that comprised this sample, this result is further indication that raters in Part 1 used the process of case formulation to think carefully and write a unique explanation of violence for offenders.

When considering each of the three mechanisms, RAs judged *Within Case* pairs as more similar than *Across Case* pairs. The effect sizes were relatively small, but statistically significant. Again, given the homogeneity of the cases and the type of violence in the sample, some similarities were expected across cases. However, an examination of the similarity ratings for the three primary formulation mechanisms also showed raters wrote distinct narrative formulations. An examination of pairwise comparisons for rater type *Within Case* and *Across Case* resulted in an interesting finding. When cases were paired with Expert-Expert formulations, the mean difference between *Within Case* pairs and *Across Case* pairs was significant. Given that both Experts had much more experience than the Novice raters, this result provides evidence that their formulation skills

were better. This result was not unexpected and ultimately suggests paired Expert formulations were easier to distinguish *Within Case* and *Across Case.*

The similarity of risk management strategies also resulted in some interesting findings. Across monitoring, treatment, and supervision strategies, RAs judged paired *Within Case* formulations as more alike than *Across Case* formulations. Although the mean similarity scores *Within Case* and *Across Case* for victim safety strategies were not statistically significant, the mean similarity scores were still in the expected direction. Given the lack of victim information contained within the files, this result suggested raters were still able to glean and formulate some information to recommend safety strategies.

Wilson (2013), the only other study to examine the IRR of formulation mechanisms from an SPJ and Decision Theory framework, used an analogous design of paired *Within Case* and *Across Case* formulations and then examined the similarity of pairings. The results of that study revealed less similarity for case formulations and formulation mechanisms than the current study. The participants in Wilson were recruited online and completed all violence risk assessments remotely following an online training for the assessment of sexual violence risk. It is possible that raters in the current study benefitted from the live, team approach to consensus meetings and receiving real-time feedback about their formulation skills from other raters, including two more experienced raters. In addition, the current study may have been more controlled given there were less raters and it occurred in person.

Formulation Quality

The forensic case formulations in the current study were of high quality; as expected Expert formulations were of higher quality overall than Novice formulations. Examination of the quality judgments at the item level, Experts outperformed Novices across most aspects of quality measured on the CFQC-R. Expert formulations were judged to be written more clearly and coherently, providing more of a factual foundation, more complete, more clearly tying together information across time in their formulations, and providing more predictive information such as future scenarios and management strategies as well as matching risk and responsivity appropriately. The largest and most important differences between Experts and Novices were related to the predictive quality of the formulations as well as the matched risk management strategies. That Expert

formulations were rated higher in quality across these two aspects is no surprise. Both Experts had conducted IPV risk assessments with the SARA-V3 in supervised clinical practice and were likely much more aware of the various supervision, monitoring, treatment, and victim safety strategies available and thus more specific about these strategies in their formulations. In addition, this familiarity of resources available to target offender risks and needs also meant Experts were better at matching appropriate recommendations to risk level. The skill of using an RNR approach when considering risk management strategies can be considered a higher-level skill. Novices could likely have benefitted from more specific training related to risk management strategies and local programming available.

Despite some of the discrepancies in quality, overall Novices produced high-quality formulations and outperformed Expert formulations on a few aspects of quality. Novice quality scores were slightly higher for items measuring external and internal coherence (using Decision Theory as the basis for the formulation and making non-contradictory assumptions within the formulation) as well as in the simplicity of their writing. In reviewing some Novice formulations, most used more of a formulaic approach to explaining how the offender's decision was motivated, disinhibited, and destabilized by various present and relevant risk factors. Depending on the audience, this method and style of writing may be preferred.

Strengths and Limitations

Strengths. This study has a number of strengths that help inform the forensic case formulation literature. First, this was a relatively intensive empirical study of forensic case formulation. Few studies of this kind have been conducted and this is the only study to evaluate the Decision Theory method of forensic case formulation for IPV. In addition, raters completed all steps of SARA-V3. Commonly, research is conducted only on Steps 2 and 3 (rating the presence of N, P, and V factors and the relevance of P factors) and Step 6 (Conclusory Opinions). Completing all steps is better aligned with best practice guidelines and expectations in the field when conducting violence risk assessment in the SPJ framework. Thus, the data collection for this study was truer to how users conduct violence risk assessment and therefore results may be more generalizable. Having multiple raters complete each case allowed for a more robust examination of IRR. In addition, using Expert and Novice raters allowed for comparisons across skill level. Rater

agreement was evaluated with traditional IRR coefficient calculations, but also with a novel coefficient that was better suited for the data in the current study. The presentation of various IRR coefficients and percent agreement also allowed for an examination of the advantages and disadvantages of some coefficients available to researchers. With regard to the IRR of forensic case formulations, an additional method for exploring IRR was employed using a *Within Case* and *Across Case* paired design, which provided converging evidence of adequate IRR. Finally, the CFQC-R, a relatively new and untested measure of case formulation quality, was used in the current study. This provided for some brief analyses of the measure (i.e., IRR) and also allowed for an examination of the quality of forensic case formulations derived from a Decision Theory approach.

Limitations. There are many ways in which this study could be improved. With a larger sample size of forensic cases, a number of design and data issues could be addressed. Despite stratified sampling efforts, the lack of variability in the sample remained. This was a relatively homogenous sample. The offender files included were generally moderate-to-high risk offenders who suffered predominantly from personality disorders, not major mental illness. In addition, all offenders included in the sample were men. There was also a lack of indigenous offenders, despite their overrepresentation in Canadian correctional systems. It was suspected that the homogeneity of the sample impacted the frequency of present formulation mechanisms. Raters repeatedly encountered files in which the following general forensic case formulation applied:

> Mr. X's present and relevant risk factors included P1 (Problems in Intimate Relationships), P5 (General Antisocial Conduct), P7 (Personality Disorder), P8 (Substance Abuse), and P10 (Distorted Thinking about IPV). His decision to engage in IPV was motivated by problems in his intimate relationships, personality disorder (Borderline Personality Traits) and general antisocial conduct which led to his urge to express his anger, gain control over his partner and the situation, and remain in a relationship with her. The perceived costs of violence were decreased by his substance use and personality disorder which led to a lack of insight, guilt, anxiety, and empathy regarding his decision to engage in violence. His thinking was destabilized further by his personality disorder and substance use resulting in some impaired reasoning, obsessive and perseverative thinking, and impulsive thinking.

These common present and relevant risk factors as well as formulation mechanisms mean raters in the current study did not have a chance to really evaluate other pathways to IPV, including major mental illness. In addition, because IPV was the predominant type of violence in the current sample, a true evaluation of some formulation mechanisms that were rated as present less frequently was not possible. In addition, the nested design of this study created a number of problems with regard to data analysis. Although this type of design is truer to the resources more typically available to researchers in psychology, it impacted the types of analyses that could be performed. For example, because not all raters coded every case, some IRR coefficients could not be used (e.g., ICCs). Further, only one measure of quality, the CFQC-R, was used in this study. It is possible other aspects of quality, specifically characteristics identified as important by professionals outside of the clinical psychology field, may be essential. In addition, it was difficult to conduct inferential statistical analyses with quality data given the nested design.

Given that this was a file review study, there are inherent concerns about field reliability. Files often also lacked victim information and so a full evaluation of SARA-V3 including victim vulnerability factors was not possible. In addition, there have been numerous attempts to obtain follow-up data on this sample, which have not been successful. Predictive validity analyses would help us gain a more thorough understanding of SARA-V3 and forensic case formulation in cases of IPV given the lack of research currently. Finally, SARA-V3 and forensic case formulation from an SPJ and Decision Theory perspective need further evaluation outside of researchers associated with the guide's two co-authors.

Implications for Practice and Policy

Although forensic case formulation is an advanced skill, the results of this study demonstrated that those conducting violence risk assessments can learn to formulate through training and that formulation skills are likely improved through practice. Those conducting risk assessments could benefit from a manual or companion guide specific to forensic case formulation as well as training specific to case formulation skills. In the current study, although Novice raters in Part 1 completed a three-day SARA-V3 training with practice cases, the specific online formulation training was only one-hour. They were also provided with a case formulation code book that defined and provided exemplars for each of the mechanisms and completed additional practice cases for which they received

feedback. This training, the code book, and practice with feedback likely contributed greatly to the consistency and high quality of formulations produced in Part 1. In addition, it is also likely that Novice raters in this study benefitted from feedback and observing the process of formulation by the Expert raters during consensus meetings. Practice and exposure to many different types of cases helps increase the "mental set" of cases available to the individual conducting the risk assessment to better ground decisions around risk factor ratings and the presence of motivators, disinhibitors, and destabilizers. For example, during the data collection for this study, there were often discussions around whether ratings were *yes* (present) and p*ossibly or partially present.* The Expert raters tended to rate some behaviours, especially related to antisocial conduct, more liberally than Novice raters. Part of this discrepancy may be attributed to the mental set of cases available to the Expert raters. In addition, Expert raters may have been better at considering the ratings across their mental set as well as within the case, as both are important when conducting risk assessment. Anecdotally, the quality of coding improved as raters completed cases. Consensus meetings moved faster and during those later meetings, raters seemed to be in agreement more often. This could have just been an improved consensus meeting process, but also an indication that all raters better understood formulation. Specifically, in the cases included in this sample, raters may have picked up on patterns (e.g., IPV offender typologies) and associated motivated, disinhibitors, destabilizers. Practice over time, exposure to more cases, and a good theoretical understanding of pathways to violence may improve the process of violence risk assessment and forensic case formulation. Finally, results from this study further support findings from prior research that conducting violence risk assessment in teams is helpful. In this study, raters reached some clarity regarding the formulation of cases when they were able to think aloud with others during consensus meetings. Given that case formulation is an iterative and generally non-linear process, hypothesizing aloud with others may aid in formulation. Ultimately, forensic case formulation is a challenging skill to teach and can be difficult to master. Training, feedback, a theoretical process (in this case, Decision Theory), and working in teams may lead to improved formulation skills and a quicker trajectory to obtaining a reasonable level of proficiency.

Given that good rater agreement for formulating forensic cases was established, this is one piece of evidence that forensic case formulation is a reliable process. In addition, the consistency of formulations was established via the *Within Case* and *Across*

Case paired design and similarity ratings. These results showed that formulations were written in an individualized manner that told a narrative about the offender and his decision to engage in violence. Finally, formulations across both Experts and Novices were judged to be of high quality. Taken together, these findings suggested that forensic case formulation is one way to understand and communicate about an individual's decision to engage in violence as well as how to manage potential future violence. Although some are seemingly sprinting toward algorithmic or artificial intelligence tools to assess violence risk in various contexts, we as a society need to carefully consider what we might lose (and gain) by these approaches. Indeed, if we want to reorient toward rehabilitation and move away from tough on crime policies, it is likely important to understand the risks and needs of each unique offender. Providing matched and appropriate resources to reduce their propensity to engage in future violence should be the goal. Forensic case formulation is a process that promotes an idiosyncratic understanding of why an individual engaged in past violence and what strategies may help them desist in the future. It is also a way to communicate to others, including the offender, a comprehensive understanding of their violence. Algorithmic approaches lack this idiographic process and thus may result in less effective management strategies. That is not to say that risk factors that comprise SPJ risk assessment guides are not also biased in some way. Although the method by which risk factors get included in SPJ guides is comprehensive (e.g., thorough literature reviews, input from subject matter experts), historically, our literature and Experts are skewed white, and until recently, male (i.e., WEIRD). Moving forward it will also be important to consider how some risk factors are derived and used when conducting a violence risk assessment and if that type of factor is discriminatory or biasing in some way (Skeem et al, 2019).

Forensic case formulation contextualized within theories of IPV perpetration could help inform the formulation process and ensure sound conceptualization. In other words, clinically, it is likely important to ground formulations within a theory of IPV perpetration. Of course, both general theories of IPV, such as feminist perspectives, attachment, and biopsychosocial, should be considered in parallel with idiosyncratic factors to explain the individual's decision to engage in violence. Chesworth (2018) suggested an ecological model that captures systems and multiple theories to more completely explain IPV perpetration. Although the sample in the current study has been described as limited given the number of cases included and similarity of the demographic and risk factors, it is

possible that it is actually more representative of IPV offenders given what theory and prior research would suggest about typologies, pathways to IPV offending, substance use, and personality disorder (Fals-Stewart et al., 2005; Johnson, 2008; Machisa et al., 2016; Mauricio et al., 2007; Messinger et al., 2014; Pence & Paymar, 1993; White & Widom, 2003). In terms of implications for the SARA-V3 specifically, this study affirms that forensic case formulation may help users further individualize violence risk assessments and recommend risk management strategies tailored to the evaluee. In practice, it is likely that personality disorder, substance abuse, and cognitive distortions related to the perpetration of IPV are likely to occur frequently. Having a strong knowledge base for IPV pathways and theory could increase IRR and quality of violence risk assessments more generally when using SARA-V3.

Implications for Theory and Future Research

Given the lack of research investigating forensic case formulations, there are many avenues for future investigations. One primary area is the predictive validity of forensic case formulations. This includes the investigation of how risk management strategies impact future acts of violence. If the goal of violence risk assessment is ultimately to reduce the occurrence of future violence, how do we measure predictive validity and recidivism rates when risk management strategies, designed to prevent violence, are implemented? This type of research is inherently challenging and complex given the many confounding variables (e.g., quality and availability of risk management strategy interventions, how recidivism is measured, if judges order recommended management strategies, if offenders violate or follow a judge's order).

A better understanding the process by which those who conduct violence risk assessment learn how to formulate cases is important. Research directed at surveying or interviewing professionals about the experience of learning forensic case formulation could help us recognize areas of case formulation that are more and less challenging in order to target training. Equally as interesting would be to examine the process by which seasoned professionals formulate cases. This type of exploratory research could help us understand how proficiency in the skills necessary for case formulation are gained and the learning experience of seasoned professionals. Again, this information could help inform training curriculum and a better understanding of knowledge and skill acquisition. With regard to the raters recruited for this study, Expert raters were relatively less experienced

than registered clinicians who have years of practical experience. In addition, Novice raters, although lacking in the same level of training and supervised clinical experiences of the Experts, were all graduate students in fields relevant to violence risk assessment. Thus, it is possible that the results of this study represent an expertise level that is quite similar and also limited in terms of the floor and ceiling in comparison to other types of raters who would potentially be trained in the use of violence risk assessment tools. Thus, a study with more clearly differentiated levels of expertise is necessary to truly test skill acquisition and ability.

It is important to test the Decision Theory method of forensic case formulation in more heterogenous samples. Although this study and Wilson (2013) specifically examined Decision Theory, these samples included IPV offenders and sex offenders. In the future, evaluation the Decision Theory method of forensic formulation in a mixed sample, one which includes all types of violence, would provide an interesting analysis of IRR. In addition, more diverse offender characteristics including the presence of major mental disorder need to be included in future tests of the Decision Theory method. Other demographic characteristics such as gender identity and expression and sexual orientation should also be empirically investigated with the SARA-V3 and forensic case formulation. The lack of victim information in the files included in the current study limited the investigation of victim factors. Victim factors could impact forensic case formulations and thus future studies should focus on samples in which robust victim information is included. A fully crossed design would rid some of the data analytic problems encountered in the current study and permit the use of generalizability theory analyses.

Specific to the SPJ approach to violence risk assessment, further development of the Decision Theory model is warranted. Moving forward, it will likely be important to assess the content validity of the theory and individual mechanisms. This could be achieved through further literature review, discussions with subject matter experts, and survey research to ensure mechanisms that motivative, disinhibit, and destabilize the decision to engage in violence are exhaustive. Additionally, work will need to be done to establish consensus around the operationalization of these mechanisms. This work will then allow for further empirical testing of the Decision Theory approach with regard to IRR, utility, and quality.

Conclusion

This was the first study to examine forensic case formulation using a Decision Theory and SPJ approach to IPV risk assessment. Forensic case formulation is an advanced skill and the results of the current study provided empirical support that the Decision Theory method of forensic case formulation. Adequate IRR was established with both Expert and Novice raters. In addition, the formulations written using the Decision Theory approach were judged to be of high quality. Overall, forensic case formulation is an area of increased interest in our field. This study provided some empirical evidence that even Novice raters can be trained in forensic case formulation and produce high quality, reliable, and consistent formulations using the Decision Theory approach.

References

Alhabib, S., Nur, U., & Jones, R. (2010). Domestic violence against women: Systematic review of prevalence studies. *Journal of Family Violence*, *25*, 369-382.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, 46(2), 293-302.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association. (2006). Guidelines and Principles for Accreditation of Programs in Professional Psychology (G&P). Retrieved from http://www.apa.org/ed/accreditation/about/policies/guiding-principles.pdf

American Board of Professional Psychology. (n.d.). Clinical psychology. Retrieved from https://legacy.abpp.org/i4a/pages/index.cfm?pageid=3307.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, *17*(1), 19-52.

Andrews, D. A., & Bonta, J. (2001). *Level of Service Inventory–Revised (LSI-R): User's manual.* North Tonawanda, NY: Multi-Health Systems.

Austin, P. C., & Hux, J. E. (2002). A brief note on overlapping confidence intervals. *Journal of Vascular Surgery*, *36*(1), 194-195.

Banasik, M., Nowopolski, M., & Gierowski, J. (2018). Formulation in forensic psychiatry: Problems, challenges and the usefulness of psychocriminological concepts. *Problems of Forensic Sciences, 111*(1), (79-96).

Beaupré, P. (2015). *Cases in adult criminal courts involving intimate partner violence* (Statistics Canada – Catalogue no. 85-002-X). Statistics Canada.

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*(1), 60-67.

Bennett, C. A., & Franklin, N. L. (1954). *Statistical analysis in chemistry and the chemical industry*. Hoboken: John Wiley & Sons.

Bjørkly, S. E., Eidhammer, G., & Selmer, L. (2014). Concurrent validity and clinical utility of the HCR-20^{V3} compared with the HCR-20 in forensic mental health nursing: Similar tools but improved method. *Journal of Forensic Nursing, 10*(4), 234-242.

Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation* (User Report 2007– 06). Ottawa, Ontario: Public Safety Canada.

Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development.* New York, NY: Basic Books.

British Psychological Society. (2017). *Practice guidelines, Third edition (*BPS Publication no. INF115/08.17). Leichester, UK.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687-699.

Brown, S., & Völlm, B. (2013). Case formulation in personality disordered offenders: Views from the front line. *Criminal Behaviour and Mental Health*, *23*(4), 263-273.

Brown, S., & Völlm, B. (2016). The implementation of case formulation by probation officers: Service user and carer views. *The Journal of Forensic Psychiatry & Psychology, 27*(2)*,* 215-231.

Bucci, S., French, L., & Berry, K. (2016). Measures assessing the quality of case conceptualization: A systematic review. *Journal of Clinical Psychology, 72*(6), 517-533.

Butler, G. (1998) Clinical formulation, in A.S. Bellack and M. Hersen (eds.) *Comprehensive Clinical Psychology*, Oxford, UK: Pergamon.

Campbell, J. C. (1986). Nursing assessment of risk of homicide for battered women. *Advances in Nursing Science, 8*(4), 36-51.

Campbell, A. M. (2020). An increasing risk of family violence during the Covid-19 pandemic: Strengthening community collaborations to save lives. *Forensic Science International: Reports*, *2,* 1-3

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*(7)*,* 1089-1097.

Canadian Psychological Association. (2012). *Evidence-based practice of psychological treatments: A Canadian perspective. Report of the CPA task force on evidence-based practice of psychological treatments.* Retrieved from https://www.cpa.ca/docs/File Practice/Report_of_the_EBP_Task_Force_FINAL_Board_Approved_2012.pdf

Catalano, S. (2015). *Intimate Partner Violence, 1993–2010* (Report No. NCJ 239203)*.* US Department of Justice, Bureau of Justice, Washington, DC.

Chesworth, B. R. (2018). Intimate partner violence perpetration: Moving toward a comprehensive conceptual framework. *Partner abuse*, *9*(1), 75-100.

Cicchetti, D. V., Klin, A., & Volkmar, F. R. (2017). Assessing binary diagnoses of bio-behavioral disorders. *The Journal of Nervous and Mental Disease, 205*(1), 58-65.

Cicchetti, D. V., & Sparrow, S. S. (1990). Assessment of adaptive behavior in young children. In J. H. Johnson & J. Goldman (Eds.), *Pergamon general psychology series, 163. Developmental assessment in clinical child psychology: A handbook* (p. 173–196). Pergamon Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

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

Conger, A. J. (1985). Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement*, *45*(4), 861-868.

Connell, C. (2015). An integrated case formulation approach in forensic practice: the contribution of Occupational Therapy to risk assessment and formulation. *The Journal of Forensic Psychiatry & Psychology*, *26*(1), 94-106.

Daffern, M., Ferguson, M., Ogloff, J., Thomson, L., & Howells, K. (2007). Appropriate treatment targets or products of a demanding environment? The relationship between aggression in a forensic psychiatric hospital with aggressive behaviour preceding admission and violent recidivism. *Psychology, Crime & Law*, *13*(5), 431-441.

Daffern, M., Howells, K., Mannion, A., & Tonkin, M. (2009). A test of methodology intended to assist detection of aggressive offence paralleling behaviour within secure settings. *Legal and Criminological Psychology*, *14*(2), 213-226.

Daffern, M., Howells, K., Stacey, J., Hogue, T., & Mooney, P. (2008). Is sexually abusive behaviour in personality disordered inpatients analogous to sexual offences committed prior to hospitalization?. *Journal of Sexual Aggression*, *14*(2), 123-133.

Davies, J., Black, S., Bentley, N., & Nagi, C. (2013). Forensic case formulation: Theoretical, ethical and practical issues. *Criminal Behaviour and Mental Health*, *23*(4), 304-314.

Delle-Vergini, V. & Day, A. (2016). Case formulation in forensic practice: Challenges and opportunities. *The Journal of Forensic Practice, 18*(3)*,* 240–250.

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20: Assessing Risk for Violence* (3rd ed.). Vancouver, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Doumas, D. M., Pearson, C. L., Elgin, J. E., & McKinley, L. L. (2008). Adult attachment as a risk factor for intimate partner violence: The “mispairing” of partners' attachment styles. *Journal of Interpersonal Violence*, *23*(5), 616-634.

Dugan, L., Nagin, D. S., & Rosenfeld, R. (1999). Explaining the decline in intimate partner homicide: The effects of changing domesticity, women's status, and domestic violence resources. *Homicide Studies*, *3*, 187-214.

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*(1), 6-10.

Eells, T. D. (2007). History and current status of psychotherapy case formulation. In T. D. Eells (Ed.), *Handbook of psychotherapy case formulation, 2nd ed.* (pp. 3–32). New York: Guilford.

Eells, T. D. (Ed.). (2014). *Handbook of psychotherapy case formulation*. New York: Guilford Press.

Eells, T. D., & Lombart, K. G. (2011). Theoretical and evidence-based approaches to case formulation. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation (pp. 3-32).* Hoboken, NJ: Wiley.

Fals-Stewart, W., Leonard, K. E., & Birchler, G. R. (2005). The occurrence of male-to-female intimate partner violence on days of men's drinking: The moderating effects of antisocial personality disorder. *Journal of Consulting and Clinical Psychology*, *73*(2), 239.

Farmer, A., & Tiefenthaler, J. (2003). Explaining the recent decline in domestic violence. *Contemporary Economic Policy*, *21*(2), 158-172.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ: British Medical Journal*, *345*(7868), 1-12.

Federal Bureau of Investigation. (2017). *Crime in the United States by volume and rate per 100,000 inhabitants, 1997-2016.* Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/tables/table-1.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549.

Felson, R. B. (2009). Violence, crime, and violent crime. *International Journal of Conflict and Violence*, *3*(1), 23-39.

[a]Feng, G. (2013). Underlying determinants driving agreement among coders. *Quality & Quantity, 47*(5), 2983-2997.

[b]Feng, G. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity, 47*(5), 2959-2982.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* John Wiley & Sons.

Flinn, L., Braham, L., & das Nair, R. (2014). How reliable are case formulations? A systematic literature review. *British Journal of Clinical Psychology*, *54*(3), 266-290.

Garcia-Moreno, C., Jansen, H. A., Ellsberg, M., Heise, L., & Watts, C. H. (2006). Prevalence of intimate partner violence: Findings from the WHO multi-country study on women's health and domestic violence. *The Lancet*, *368*(9543), 1260-1269.

Gatner, D. T., Ryan, T. J., Douglas, K. S., & Hart, S. D. (2017, June). *A conceptual analysis of violence risk formulation for Psychopathic Personality Disorder*. Paper presented at the annual meeting of the International Association of Forensic Mental Health Services, Split, Croatia.

Gatner, D. T. (2019). *How much does that cost? Examining the economic costs of crime in North America attributable to people with psychopathic personality disorder.* (Unpublished doctoral dissertation). Simon Fraser University, Burnaby, BC

Glaser, B. (2011). Paternalism and the good lives model of sex offender rehabilitation. *Sexual Abuse*, *23*(3), 329-345.

Graham, L. M., Sahay, K. M., Rizo, C. F., Messing, J. T., & Macy, R. J. (2019). The validity and reliability of available intimate partner homicide and reassault risk assessment tools: A systematic review. *Trauma, Violence & Abuse.* Advance online publication. doi: 10.1177/1524838018821952

Grann, M., & Wedin, I. (2002). Risk factors for recidivism among spousal assault and spousal homicide offenders. *Psychology, Crime and Law*, *8*(1), 5-23.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255-282.

Gwet, K. L. (2002). Kappa statistics is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability*, *1*(6), 1-6.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

Gwet, K. L. (2014). *Handbook of Interrater Reliability, Fourth Edition.* Advanced Analytics, LLC.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior*, *36*(9), 865-891.

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

Harris, G., Rice, M., Quinsey, V., & Cormier, C. (2015). *Violent offenders: Appraising and managing risk (3rd ed.).* American Psychological Association.

Hart, S. D. (2015). *Assessing and managing violence risk: Advanced workshop.* Professional Training, Vancouver, British Columbia, Canada.

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, *3*(1), 121-137.

Hart, S. D., & Boer, D. P. (2010). Structured professional judgment guidelines for sexual violence risk assessment: The Sexual Violence Risk–20 (SVR-20) and Risk for Sexual Violence Protocol (RSVP). In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment tools* (pp. 269-294). Milton Park, UK: Routledge.

Hart, S. D, & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences & the Law, 31*(1), 81-102.

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., & Logan, C. (2011). Formulation of violence risk using evidence-based assessments: The Structured Professional Judgment approach. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation* (pp. 83-107). Hoboken, NJ: Wiley.

Hart, S. D., Kropp, P. R., Laws, D. R., Klaver, J., Logan, C., & Watt, K. A. (2003). *The Risk for Sexual Violence Protocol (RSVP): Structured professional guidelines or assessing risk of sexual violence.* Vancouver, Canada, BC: The Institute Against Family Violence.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments. *British Journal of Psychiatry*, *190* (suppl. 49), s60-s65.

Hart, S. D., Sturmey, P., Logan, C., & McMurran, M. (2011). Forensic case formulation. *International Journal of Forensic Mental Health*, *10*(2), 118-126.

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*(7), 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*(1), 64-75.

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2011). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment, 24*(1)*,* 64-101.

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*(3), 267-275.

Hilton, N. Z., Pham, A. T., Jung, S., Nunes, K., & Ennis, L. (2020). Risk scores and reliability of the SARA, SARA-V3, B-SAFER, and ODARA among Intimate Partner Violence (IPV) cases referred for threat assessment. *Police Practice and Research.* Advance online publication. doi: 10.1080/15614263.2020.1798235

Hopton, J., Cree, A., Thompson, S., Jones, R., & Jones, R. (2018). An evaluation of the quality of HCR-20 risk formulations: A comparison between HCR-20 Version 2 and HCR-20 Version 3. *International Journal of Forensic Mental Health*, *17*(2)*,* 195-201.

Johnson, M. P. (2008). *A typology of domestic violence: Intimate terrorism, violent resistance, and situational couple violence*. Boston, MA: Northeastern University Press

Jones, L. F. (1997) Developing models for managing treatment integrity and efficacy in a prison based TC: The Max Glatt Centre. In E. Cullen, L. Jones, & R. Woodward (Eds.) *Therapeutic Communities for Offenders*. Wiley.

Jones L. F. (2004). Offence Paralleling Behaviour (OPB) as a framework for assessment and interventions with offenders. In A. Needs & G. Towl (Eds.) *Applying Psychology to Forensic Practice*. British Psychological Society.

Kapoor, R., & Williams, A. (2012). An unwelcome guest: The unconscious mind in the courtroom. *Journal of the American Academy of Psychiatry and the Law, 40*(4)*,* 456-461.

Klein, D. (2018). Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal*, *18*(4), 871-901.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61-70.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411-433.

Krippendorff, K. H. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. In C. T. Salmon (Ed.), *Communication Yearbook 36* (pp. 481-499). Routledge.

Krippendorff, K. H. (2016). Misunderstanding reliability. *Methodology, 12*(4), 139-144.

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*(1), 101-118.

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., 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. (2008). *Manual for the Spousal Assault Risk Assessment Guide,* 2nd ed*.* Vancouver, Canada: British Columbia Institute on Family Violence.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1)*,* 159-174.

Laws, D. R., & Ward, T. (2011). Desistance and sexual offending: Alternatives to throwing away the keys. Guildford Press.

Lewis, G., & Doyle, M. (2009). Risk formulation: What are we doing and why?. *International Journal of Forensic Mental Health*, *8*(4), 286-292.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365-377.

Logan, C. (2014). The HCR-20 Version 3: A case study in risk formulation. *International Journal of Forensic Mental Health*, *13*(2), 172-180.

Logan, C., & Johnstone, L. (2010). Personality disorder and violence: Making the link through risk formulation. *Journal of Personality Disorders*, *24*(5), 610-633.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587-604.

Machisa, M. T., Christofides, N., & Jewkes, R. (2016). Structural pathways between child abuse, poor mental health outcomes and male-perpetrated intimate partner violence (IPV). *PLoS One*, *11*(3), 1-15.

Mappin, L., Dawson, D. L., Gresswell, D. M., & Beckley, K. (2013). Female-perpetrated intimate partner violence: An examination of three cases using multiple sequential functional analysis. *Criminal Behaviour and Mental Health*, *23*(4), 290-303.

Mauricio, A. M., Tein, J. Y., & Lopez, F. G. (2007). Borderline and antisocial personality scores as mediators between attachment and intimate partner violence. *Violence and Victims*, *22*(2), 139-157.

Maxwell, A.E. (1977). *Multivariate analysis in behavioural research*. London: Chapman & Hall.

Maxwell, C. D., Garner, J. H., & Fagan, J. A. (2002). The preventive effects of arrest on intimate partner violence: Research, policy and theory. *Criminology and Public Policy, 2*, 51-80.

McClellan, A. C., & Killeen, M. R. (2000). Attachment theory and violence toward women by male intimate partners. *Journal of Nursing Scholarship*, *32*(4), 353–360.

McMurran, M., & Bruford, S. (2016). Case formulation quality checklist: A revision based upon clinicians' views. *Journal of Forensic Practice*, *18*(1), 31-38.

McMurran, M., Logan, C., & Hart, S. D. (2012). *Case Formulation Quality Checklist*. Institute of Mental Health: Nottingham.

Messing, J. T., Thaller, J. (2013). The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence, 28*(7), 1537–1558.

Messinger, A. M., Fry, D. A., Rickert, V. I., Catallozzi, M., & Davidson, L. L. (2014). Extending Johnson's intimate partner violence typology: Lessons from an adolescent sample. *Violence Against Women*, *20*(8), 948-971.

Minoudis, P., Craissati, J., Shaw, J., McMurran, M., Freestone, M., Chuan, S., & Leonard, A. (2013). An evaluation of case formulation training and consultation with probation officers. *Criminal Behaviour and Mental Health, 23*(4), 252-262.

Mumma, G. H. (2011). Current issues in case formulation. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation* (pp. 33-60). Wiley-Blackwell.

Nicholls, T. L., Pritchard, M. M., Reeves, K. A., Hilterman, E. (2013). Risk assessment in intimate partner violence: A systematic review of contemporary approaches. *Partner Abuse, 4*(1), 76–168.

Ogloff, J. R. P., & Davis, M. R. (2004). Advances in offender assessment and rehabilitation: Contributions of the risk-needs-responsivity approach. *Psychology, Crime & Law, 10*(3), 229-242.

Palmer, L. (2016). *Dynamic risk factors and their utilisation in case formulation: A new conceptual framework* (Unpublished master's thesis). Victoria University of Wellington, Kelburn, Wellington, New Zealand.

Pardo, M. S., & Allen, R. J. (2008). Juridical proof and the best explanation. *Law and Philosophy*, *27*(3), 223-268.

Pence, E., & Paymar, M. (1993). *Education groups for men who batter: The Duluth model.* New York: Springer.

Persons, J. B., & Tompkins, M. A. (1997). Cognitive behavioral case formulation. In T. Eels (Ed.), *Handbook of psychotherapy case formulation* (pp. 314–339). New York: Guilford Press.

Polaschek, D. L. (2012). An appraisal of the risk–need–responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and Criminological Psychology*, *17*(1), 1-17.

Quarfoot, D., & Levine, R. A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution?. *The American Statistician*, *70*(4), 373-384.

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

Ridley, C. R., Jeffrey, C. E., & Robertson III, R. B. (2017) Case mis-conceptualization in psychological treatment: An enduring clinical problem. *Journal of Clinical Psychology, 73*, 359-375.

Ryan, T. J. (2016). *An examination of the interrater reliability and concurrent validity of the Spousal Assault Risk Assessment Guide – Version 3 (SARA-V3)* (Unpublished master's thesis). Simon Fraser University, Burnaby, British Columbia, Canada.

Ryan, T. J., Gatner, D. T., Slaney, K. L., & Hart, S. D. (2019). *The SPJ approach to violence risk case formulation: Philosophical underpinnings.* Paper presented at the 2019 annual meeting of the American Psychology-Law Society (AP-LS), Portland, OR.

Schafers, C. L. (2019). *Risk, responsivity, and the treatment process in an intimate partner violence group program.* (Unpublished doctoral dissertation). University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182-186.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly, 19*(3), 321-325.

Sher, A. M., & Gralton, E. (2014). Implementation of the START: AV in a secure adolescent service. *Journal of Forensic Practice*, *16*(3), 184-193.

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

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., Doyle, M., Folino, J. O., Godoy-Cervera, V., Grann, M., Ho, R. M. Y., Large, M. M., Hjort Nielsen, L., Pham, T. H., Francisca Rebocho, M., Reeves, K. A., Rettenberger, M., de Ruiter, C., Seewald, K., & Otto, R. K., (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*(3), 193-206.

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice and Behavior*, *37*(9), 965-988.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*(3), 499-513.

Sinha, M. (2013). *Family violence in Canada: A statistical profile, 2011. Section 3: Intimate partner violence*. Retrieved from https://www150.statcan.gc.ca/n1/pub/85-002-x/2013001/article/11805-eng.pdf.

Skeem, J. L., Scurich, N., & Monahan, J. (2019). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior, 44*(1), 51-59.

Statistics Canada. (2017). *Canada's crime rate: Two decades of decline.* Retrieved from http://www.statcan.gc.ca/pub/11-630-x/11-630-x2015001-eng.htm.

Storey, J. E., & Hart, S. D. (2014). An examination of the Danger Assessment as a victim-based risk assessment instrument for lethal intimate partner violence. *Journal of Threat Assessment and Management*, *1*(1), 56-66.

Straus, M. A. (2006). Future research on gender symmetry in physical assaults on partners. *Violence Against Women*, *12*(11), 1086-1097.

Sturmey, P., & McMurran, M. (2011). *Forensic case formulation.* Wiley-Blackwell.

Sugerman, D. B., & Boney-McCoy, S. (2000). Research synthesis in family violence: The art of reviewing the research. *Journal of Aggression, Maltreatment and Trauma, 4*, 55-82.

Sutherland, A. A., Johnstone, L., Davidson, K. M., Hart, S. D., Cooke, D. J., Kropp, P. R., ... & 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*(2), 119-133.

Van der Put, C. E., Gubbels, J., and Assink, M. (2019). Predicting domestic violence: A meta-analysis on the predictive validity of risk assessment tools. *Aggression and Violent Behavior, 47*(1), 100-116.

van Gelder, N., Peterman, A., Potts, A., O'Donnell, M., Thompson, K., Shah, N., & Oertelt-Prigione, S. (2020). COVID-19: Reducing the risk of infection might increase the risk of intimate partner violence. *EClinicalMedicine*, *21,* 1-2.

Vess, J., Ward, T., & Collie, R. (2008). Case formulation with sex offenders: An illustration of individualized risk assessment. *The Journal of Behavior Analysis of Offender and Victim Treatment and Prevention*, *1*(3), 284-293.

Ward, T. (2002). The management of risk and the design of good lives. *Australian Psychologist*, *37*(3), 172-179.

Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime, and Law, 10*(3), 243−257.

Ward, T., & Stewart, C. (2003). Criminogenic needs and human needs: A theoretical model. *Psychology, Crime & Law, 9*(2), 125−143.

Webster, C. D., Martin, M., Brink, J., Nicholls, T. L. & Desmarais, S. L. (2009). *Manual for the Short-term Assessment of Risk and Treatability (START) (Version 1.1)*. BC Mental Health St. Joseph's Healthcare: British Columbia.

Weerasekera, P. (1993). Formulation: A multiperspective model. *The Canadian Journal of Psychiatry*, *38*(5), 351-358.

Weerasekera, P. (1996). *Multiperspective case formulation: A step towards treatment integration*. Krieger: Malabar, FL.

White, H. R., & Widom, C. S. (2003). Intimate partner violence among abused and neglected children in young adulthood: The mediating effects of early aggression, antisocial personality, hostility and alcohol problems. *Aggressive Behavior*, *29*(4), 332-345.

Whitehead, P. R., Ward, T., & Collie, R. M. (2007). Time for a change: Applying the Good Lives Model of rehabilitation to a high-risk violent offender. *International Journal of Offender Therapy and Comparative Criminology*, *51*, 578-598.

Wikström, P. O. H., & Treiber, K. H. (2009). Violence as situational action. *International Journal of Conflict and Violence, 3*(1), 75-96.

Williams, K. R., & Houghton, A. B. (2004). Assessing the risk of domestic violence reoffending: A validation study. *Law and Human Behavior*, *28*(4), 437-455.

Williams, K., Wormith, J., Bonta, J., & Sitarenios, G. (2017). The use of meta-analysis to compare and select offender risk instruments: A commentary on Singh, Grann, and Fazel (2011). *International Journal of Forensic Mental Health, 16*(1), 1-15.

Willis, G. & Ward, T. (2013). The good lives model: Evidence that it works. In L. Craig, L. Dixon, & T.A. Gannon (2013), *What Works in Offender Rehabilitation: An evidence based approach to assessment and treatment* (pp. 305-318). John Wiley & Sons.

Wilson, C. (2013). *Reliability and consistency of risk formulations in assessments of sexual violence risk.* (Unpublished doctoral dissertation). Simon Fraser University, Burnaby, BC.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC_1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 1-7.

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*(5), 740-767.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate?. *BMC Medical Research Methodology*, *16*(1), 1-10.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. In C. T. Salmon (Ed.), *Communication Yearbook 36 (2012)* (pp. 419–480). New York and London: Routledge.

Zhao, X., Feng, G. C., Liu, J. S., & Deng, K. (2018). We agreed to measure agreement-redefining reliability de-justifies Krippendorff's Alpha. *China Media Research*, *14*(2), 1-15.

Appendix A. Coding Forms

Similarity Ratings Coding Form

Overall Similarity Ratings

1. What is the probability that these case formulations were written by two different raters looking at the ***same*** case?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not likely at all

100% = Extremely likely

2. How confident or sure are you in your probability estimate above?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not at all confident

100% = Extremely confident

Part 1: Case Formulation Similarity Ratings

1. Rate the *overall similarity* between the two cases.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

1a. Rate the *similarity of the motivators* described in the two cases, also considering the specific motivating mechanism(s) as well as any explanations offered.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify motivators

1b. Rate the *similarity of the disinhibitors* described in the two cases, also considering the specific disinhibiting mechanism(s) as well as any explanations offered.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify disinhibitors

1c. Rate the *similarity of the destabilizers* described in the two cases, also considering the specific destabilizing mechanism(s) as well as any explanations offered.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify destabilizers

Part 2: Case Scenarios

2. Both coders identified _____ as the most likely scenario:
   - ☐ Repeat
   - ☐ Twist
   - ☐ Escalation
   - ☐ Desistance
   - ☐ Different scenarios

☐ - Cannot rate this item, one or both formulations did not identify the most likely scenario

2a. Rate the similarity of *the most likely scenario* described for the two cases.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify the type of violence in a most likely scenario

Part 3: Management Strategies

3. Rate the *overall similarity of the management strategies* provided.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify management strategies

3a. Rate the similarity of the *monitoring* strategies provided.

(Monitoring strategies include face to face interviews, telephone interviews, visits with the POI, phone calls with mental health providers, phone calls with family, etc.)

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify monitoring strategies

3b. Rate the similarity of the *treatment* strategies provided.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify treatment strategies

3c. Rate the similarity of the *supervision* strategies provided.
(Supervision strategies include remand in custody, restraining order, report as directed, reside as directed, no weapons, no alcohol/drugs, no contact, don't associated, don't travel/no go, etc.)

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Both coders recommended no contact with victim(s)
☐ - Both coders recommended abstinence from substances
☐ - Cannot rate this item, one or both formulations did not identify supervision strategies

3d. Rate the similarity of the *victim safety planning* strategies provided.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

0% = Not similar at all

40% = Different in many important respects, but similar in a few

70% = Similar in many important respects, but different in a few

100% = Very similar

☐ - Cannot rate this item, one or both formulations did not identify victim safety planning strategies (note: you should not select this option if both formulations indicated victim safety planning strategies could not be identified due to a lack of information about the victim OR if few specific victim safety planning strategies are listed due to a paucity of victim information. You should only select this option if one or both formulations fail to mention victim safety planning strategies at all.)

Case Formulation Quality Checklist-Revised

Case Formulation Quality Checklist (CFQC; McMurran & Bruford, 2016)		
Rating scale:		
0 Does not meet this criterion at all	5 Meets this criterion somewhat	10 Meets this criterion exceptionally well
Criterion	**Definition**	**Rating (0 – 10)**
1. Narrative	The formulation is presented in everyday language that tells a coherent, ordered, and meaningful story • Evaluate the clarity of the writing. • How much logical linguistic sense does the formulation make?	
2. External Coherence	The formulation is explicitly consistent with an empirically supported theory • Uses Decision Theory – identifies motivators, disinihibitors, and destabilizers (at least one from each category) • Links the motivators, disinihibitors, and destabilizers to risk factors	
3. Factual Foundation	The formulation is based on relevant information about the case that is adequate in terms of quantity and quality • Does the formulation have utility? • Is it detailed? Are the included details relevant to the formulation? • Is there enough context to understand the case? Is it individualized enough? • How individualized and useful is the formulation from a violence risk point of view?	
4. Internal Coherence	The formulation rests on propositions or makes assumptions that are compatible or non-contradictory	
5. Completeness	The formulation has a plot that ties together as much of the relevant information as possible	

6. Events are understood by the way relate over time	The formulation ties together information about the past, present, and future of the case	
7. Simplicity	The formulation is free from unnecessary details	
8. Predictive	The formulation goes beyond description, statement of facts, or classification to make detailed and testable predictions. The key predictions are those about which strategies will be most effective in treating and managing harmful behaviour • Formulation includes most likely future scenario • Includes monitoring, treatment, and supervision strategies	
9. Action Oriented	The formulation prioritizes and plans treatments • Identifies appropriate monitoring, treatment, and supervision strategies specific to the needs of the individual • Risk and responsivity are matched appropriately (e.g., generally a higher risk client has more intensive recommendations)	
10. Overall Quality	The formulation is comprehensive, logical, coherent, focused, and informative	

Appendix B. Formulation Mechanisms – Supplemental Interrater Reliability Analyses

Table B.1 Reliability (Various Coefficients) of Motivating Mechanisms – All Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
M1. Defence…	.95*** [.91, 1.00]	.85*** [.71, .99][b]	.06 [-.44, .56][f]	.04 [-.15, .23][f]	.05 [-.14, .24][f]
M2. Justice…	.72*** [.60, .85]	.25 [-.09, .59][f]	.33* [.02, .64][e]	.33* [.02, .64][e]	.33* [.05, .62][e]
M3. Gain…	.91*** [.86, .97]	.77*** [.62, .92][b]	.42* [.04, .79][e]	.47** [.13, .81][e]	.48** [.14, .82][e]
M4. Change…	.73*** [.56, .90]	.27 [-.20, .74][f]	.10 [-.12, .33][f]	.02 [-.19, .24][f]	.03 [-.17, .23][f]
M5. Status…	.85*** [.79, .93]	.59*** [.36, .82][d]	.27 [-.02, .56][e]	.26 [-.04, .55][e]	.25 [-.02, .52][e]
M6. Release…	.70*** [.58, .82]	.18 [-.15, -.50][f]	.14 [-.13, -.41][f]	.11 [-.15, .37][f]	.10 [-.15, .35][f]
M7. Arousal…	.98*** [.96, 1.00]	.96*** [.90, 1.00][a]	.45 [-.27, 1.00][f]	.40 [-.23, 1.00][f]	.40 [-.24, 1.00][f]
M8. Proximity…	.77*** [.67, .87]	.37** [.10, .65][e]	.50*** [.29, .71][d]	.48*** [.26, .69][d]	.49*** [.27, .72][d]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.2 Reliability (Various Coefficients) of Motivating Mechanisms – Expert Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
M1. Defence…	.89*** [.80, .97]	.69*** [.46, .93][c]	.04 [-.17, .25][f]	.03 [-.19, .25][f]	.04 [-.18, .26][f]
M2. Justice…	.71*** [.60, .81]	.21 [-.07, .49][f]	.28* [.02, .54][e]	.28* [.02, .54][e]	.29* [.03, .55][e]
M3. Gain…	.93*** [.88, .98]	.82*** [.68, .96][b]	.65*** [.39, .90][c]	.64*** [.39, .90][c]	.65*** [.39, .90][c]
M4. Change…	.82*** [.72, .92]	.51*** [.24, .78][d]	.13 [-.20, 046][f]	.12 [-.21, .46][f]	.13 [-.20, .47][f]
M5. Status…	.84*** [.75, .94]	.58*** [.33, .82][d]	.30 [-.05, .65][e]	.29 [-.07, .65][f]	.31 [-.05, .66][e]
M6. Release…	.73*** [.62, .83]	.27 [-.01, .54][e]	.18 [-.10, .45][f]	.15 [-.14, .45][f]	.17 [-.12, .46][f]
M7. Arousal…	.96*** [.89, 1.00]	.92*** [.80, 1.00][a]	.00 [.00, .00][f]	-.04 [-.09, .00][f]	-.01 [-.04, .02][f]
M8. Proximity…	.77*** [.66, .87]	.38** [.12, .65][e]	.43*** [.18, .68][d]	.42*** [.17, .67][d]	.44*** [.19, .68][d]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.3 Reliability (Various Coefficients) of Motivating Mechanisms – Novice Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
M1. Defence…	.96*** [.87, 1.00]	.87*** [.59, 1.00][b]	. 14 [-1.00, 1.00][f]	. 00 [-1.00, 1.00][f]	. 02 [-.38, .42][f]
M2. Justice…	.73*** [.59, .87]	. 27 [-.10, .65][f]	. 35* [.05, .66][e]	. 35* [.02, .68][e]	. 39** [.11, .67][e]
M3. Gain…	.89*** [.66, 1.00]	. 71* [.09, 1.00][e]	. 14 [-1.00, 1.00][f]	. 20 [-1.00, 1.00][f]	. 22 [-.64, 1.00][f]
M4. Change…	.64*** [.42, .86]	. 03 [-.58, .63][f]	-. 04 [-.43, .06][f]	-.07 [-.42, .27][f]	-.07 [-.36, .22][f]
M5. Status…	.86*** [.77, .95]	.61*** [.37, .85][c]	.29 [-.06, .65][f]	.22 [-.13, .56][f]	.25 [-.19, .68][f]
M6. Release…	.69*** [.46, .91]	.15 [-.46, .77][f]	.16 [-.48, .80][f]	.41 [-.13, .95][f]	.11 [-.62, .85][f]
M7. Arousal…	.98*** [.95, 1.00]	.95*** [.86, 1.00][a]	.44 [-.28, 1.00][f]	.48 [-.07, 1.00][e]	.49 [-.75, 1.00][f]
M8. Proximity…	.71*** [.44, .98]	.23 [-.50, .95][f]	.41 [-.15, .97][f]	.40 [-.18, .97][f]	.42 [-.15, .99][f]

Notes. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.4 Reliability (Various Coefficients) of Disinhibiting Mechanisms – All Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
D1. Neg. Att…	.71*** [.46, .96]	.22 [-.45, .89][f]	.10 [-.31, .50][f]	.11 [-.13, .36][f]	.12 [-.12, .36][f]
D2. Neg. Self…	.94*** [.87, 1.00]	.84*** [.65, 1.00][b]	.20 [-.05, .45][e]	.05 [-.23, .33][f]	.05 [-.24, .34][f]
D3. Alienation	.81*** [.72, .91]	.50*** [.25, .75][d]	.31* [.06, .56][e]	.25 [-.01, .50][e]	.25* [.02, .47][e]
D4. Nihilism	.90*** [.84, .96]	.73*** [.57, .89][c]	.34 [-.02, .71][e]	.34* [.03, .66][e]	.36** [.09, .64][e]
D5. Lack Insight	.93*** [.86, 1.00]	.82*** [.63, 1.00][b]	.38** [.12, .64][e]	.17 [-.10, .45][f]	.20 [-.09, .49][f]
D6. Lack Guilt	.96*** [.92, .99]	.88*** [.79, .98][a]	.67*** [.38, .97][c]	.54** [.14, .95][d]	.58** [.19, .97][d]
D7. Lack Anxiety	.95*** [.92, .99]	.88*** [.78, .98][b]	.49** [.14, .84][e]	.31 [-.03, .65][e]	.33 [-.02, .67][e]
D8. Lack Empathy	.95*** [.90, .99]	.86*** [.74, .98][b]	.50*** [.20, .81][d]	.33 [-.07, .72][f]	.37 [-.02, .75][e]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.5 Reliability (Various Coefficients) of Disinhibiting Mechanisms – Expert Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
D1. Neg. Att…	.57*** [.44, .70]	-.16 [-.50, .18][f]	.14* [.01, .26][e]	-.10 [-.35, .16][f]	-.08 [-.33, .18][f]
D2. Neg. Self…	.97*** [.91, 1.00]	.93*** [.80, 1.00][a]	.00 [.00, .00][f]	-.04 [-.09, .00][f]	-.01 [-.03, .02][f]
D3. Alienation	.90*** [.84, .90]	.74*** [.59, .90][b]	.21 [-.10, .52][f]	.19 [-.13, .52][f]	.22 [-.10, .53][f]
D4. Nihilism	.93*** [.87, 1.00]	.82*** [.67, .96][b]	.57*** [.27, .87][d]	.57*** [.27, .87][d]	.58*** [.29, .87][d]
D5. Lack Insight	.95*** [.89, 1.00]	.87*** [.74, 1.00][b]	.20 [-.17, .56][f]	.15 [-.26, .57][f]	.18 [-.23, .58][f]
D6. Lack Guilt	.96*** [.90, 1.00]	.89*** [.76, 1.00][b]	.54* [.03, 1.00][e]	.53* [.00, 1.00][e]	.54 [.03, 1.00][e]
D7. Lack Anxiety	.95*** [.89, 1.00]	.87*** [.73, 1.00][b]	.20 [-.17, .56][f]	.15 [-.26, .57][f]	.18 [-.23, .58][f]
D8. Lack Empathy	.97*** [.92, 1.00]	.93*** [.85, 1.00][a]	.64** [.19, 1.00][d]	.63** [.17, 1.00][d]	.64** [.20, 1.00][d]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.6 Reliability (Various Coefficients) of Disinhibiting Mechanisms – Novice Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
D1. Neg. Att…	.81*** [.64, .97]	.49* [.04, .93][e]	.26 [-.39, .90][f]	.20 [-.53, .94][f]	.20 [-.48, .87][f]
D2. Neg. Self…	.91*** [.77, 1.00]	.76*** [.38, 1.00][c]	.08 [-.82, .98][f]	.10 [-.89, 1.00][f]	.11 [-.72, .93][f]
D3. Alienation	.79*** [.65, .94]	.44* [.06, .83][e]	.38 [-.03, .80][e]	.40* [.02, .77][e]	.40 [-.10, .89][f]
D4. Nihilism	.87*** [.73, 1.00]	.66*** [.28, 1.00][d]	.16 [-.55, .88][f]	.15 [-.77, 1.00][f]	.18 [-.38, .74][f]
D5. Lack Insight	.92*** [.79, 1.00]	.78*** [.43, 1.00][c]	.36 [-.07, .79][f]	.14 [-.70, .97][f]	.20 [-.26, .65][f]
D6. Lack Guilt	.95*** [.89, 1.00]	.86*** [.70, 1.00][b]	.66** [.25, 1.00][d]	.48 [-.24, 1.00][f]	.55* [.11, .98][e]
D7. Lack Anxiety	.95*** [.83, 1.00]	.87*** [.55, 1.00][c]	.51 [-.49, 1.00][f]	.29 [-1.00, 1.00][f]	.35 [-.38, 1.00][f]
D8. Lack Empathy	.94*** [.84, 1.00]	.83*** [.58, 1.00][b]	.46 [-.13, 1.00][f]	.19 [-.95, 1.00][f]	.25 [-.35, .84][f]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.7 Reliability (Various Coefficients) of Destabilizing Mechanisms – All Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
De1. Disturb. Attn…	.97*** [.91, 1.00]	.94*** [.82, 1.00][a]	.16 [-.30, .63][f]	-.02 [-.04, .01][f]	-.01 [-.04, .02][f]
De2. Disturb. Sen…	.94*** [.87, 1.00]	.83*** [.66, 1.00][b]	.46* [.06, .86][e]	.44 [.00, .89][e]	.43* [.02, .85][e]
De3. Impair. Mem…	.98*** [.94, 1.00]	.95*** [.87, 1.00][a]	.34 [.00, .68][e]	-.01 [-.03, .01][f]	-.01 [-.04, .17][f]
De4. Impair. Rea…	.63*** [.46, .79]	-.01 [-.44, .42][f]	.10 [-.20, .40][f]	.07 [-019, .34][f]	.08 [-.18, .33][f]
De5. Obsessive…	.83*** [.76, .89]	.53*** [.34, .71][d]	.45*** [.23, .68][d]	.42*** [.22, .63][d]	.46*** [.27, .64][d]
D6e. Impulsive…	.83*** [.75, .92]	.54*** [.31, .77][d]	.49** [.19, .80][d]	.38 [-.03, .78][e]	.41 [-.02, .84][e]

Notes. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.8 Reliability (Various Coefficients) of Destabilizing Mechanisms – Expert Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
De1. Disturb. Attn…	--	--	--	--	--
De2. Disturb. Sen…	.94*** [.90, 1.00]	.83*** [.67, 1.00][b]	.39 [-.15, .92][f]	.41 [-.09, .92][f]	.37 [-.21, .94][f]
De3. Impair. Mem…	--	--	--	--	--
De4. Impair. Rea…	.47*** [.35, .59]	-.43* [-.75, -.11][f]	-.06 [-.23, .11][f]	-.33* [-.57, -.10][f]	-.32* [-.56, -.08][f]
De5. Obsessive…	.84*** [.75, .93]	.58*** [.35, .80][d]	.51*** [.25, .77][d]	.51*** [.25, .77][d]	.52*** [.26, .78][d]
D6e. Impulsive…	.80*** [.68, .91]	.45** [.16, .74][d]	-.04 [-.10, .74][f]	-.16 [-.25, -.08][f]	-.13*** [-.21, -.06][f]

Notes. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.9 Reliability (Various Coefficients) of Destabilizing Mechanisms – Novice Raters

	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
De1. Disturb. Attn…	.94*** [.82, 1.00]	.88*** [.63, 1.00][b]	-.12 [-1.00, 1.00][f]	-.03 [-1.00, 1.00][f]	-.02 [-.08, .04][f]
De2. Disturb. Sen…	.93*** [.86, 1.00]	.82*** [.62, 1.00][b]	.46 [-.12, 1.00][f]	.44 [-.17, 1.00][f]	.41 [-.38, 1.00][f]
De3. Impair. Mem…	.96*** [.85, 1.00]	.92*** [.70, 1.00][b]	.24 [-.76, 1.00][f]	-.02 [-1.00, 1.00][f]	-.02 [-.08, .05][f]
De4. Impair. Rea…	.63*** [.44, .82]	.00 [-.50, .50][f]	.10 [-.35, .55][f]	.08 [-.41, .57][f]	.08 [-.32, .49][f]
De5. Obsessive…	.84*** [.70, .98]	.57** [.18, .95][d]	.50** [.15, .86][d]	.43* [.04, .85][e]	.51** [.91, .82][d]
D6e. Impulsive…	.90*** [.77, 1.00]	.72*** [.37, 1.00][c]	.73*** [.42, 1.00][c]	.69*** [.31, 1.00][d]	.75*** [.48, 1.00][c]

Notes. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect*, b = *substantial*, c = *moderate*, d = *fair*, e = *slight*, f = *poor.*

Table B.10 Reliability (AC1) of Perpetrator Risk Factors Linked to Motivators – Expert Raters

	Motivating Mechanisms AC_1 [95% CI]							
P Factors	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
P1. Intimate Relation…	.83*** [.70, .97][b]	.34* [.06, .62][e]	.83*** [.69, .97][b]	.75*** [.58, .93][b]	.82*** [.66, .97][b]	.49*** [.23, .76][d]	.98*** [.94, 1.00][a]	.50*** [.25, .76][d]
P2. Non-Intimate Re…	.94*** [.86, 1.00][a]	.94*** [.86, 1.00][a]	.91*** [.82, 1.00][a]	.93*** [.84, 1.00][a]	--	.91*** [.80, 1.00][b]	--	.94*** [.86, 1.00][a]
P3. Employment/ Fin…	--	.94*** [.86, 1.00][a]	.92*** [.82, 1.00][b]	.86*** [.73, .99][b]	--	.86*** [.73, .99][b]	--	--
P4. Trauma/ Victimi…	--	.94*** [.86, 1.00][a]	--	.96*** [.88, 1.00][a]	--	.88*** [.77, 1.00][b]	--	.94*** [.86, 1.00][a]
P5. General Antisoc…	.98*** [.94, 1.00][a]	.96*** [.90, 1.00][a]	.98*** [.94, 1.00][a]	--	.96*** [.88, 1.00][a]	.91*** [.80, 1.00][b]	.98*** [.94, 1.00][a]	.96*** [.90, 1.00][a]
P6. Major Mental Dis…	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]	--	.91*** [.81, 1.00][b]	--	.88*** [.77, .99][b]	--	.94*** [.86, 1.00][a]
P7. Personality Dis…	.96*** [.90, 1.00][a]	.56*** [.32, .80][d]	.91*** [.82, 1.00][a]	.68*** [.46, .89][c]	.93*** [.83, 1.00][a]	.60*** [.36, .83][c]	.96*** [.90, 1.00][a]	.61*** [.37, .84][c]
P8. Substance Use	.98*** [.94, 1.00][a]	--	.93*** [.85, 1.00][a]	.91*** [.80, 1.00][b]	--	.62*** [.40, .84][c]	.98*** [.94, 1.00][a]	--
P9. Violent/Suicidal…	--	.98*** [.94, 1.00][a]	--	.98*** [.92, 1.00][a]	--	.96*** [.88, 1.00][a]	--	--
P10. Distorted Think…	.83*** [.70, .97][b]	.21 [-.08, .49][f]	.79*** [.63, .95][b]	.85*** [.72, .98][b]	.68*** [.47, .88][c]	.57*** [.32, .81][d]	.96*** [.90, 1.00][a]	.65*** [.42, .87][c]

Notes. N = 50. M1 = Defence/Distance/Protection, M2 = Justice/Honour/Retribution, M3 = Gain/Profit/Acquisition, M4 = Change/Control/Compliance, M5 = Status/Esteem/Dominance, M6 = Release/Expression/ Emotion, M7 = Arousal/Activity/Excitement, M8 = Proximity/Affiliation/Conformity. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.11 Reliability (AC1) pf Perpetrator Risk Factors Linked to Motivators – Novice Raters

	Motivating Mechanisms AC_1 [95% CI]							
P Factors	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
P1. Intimate Relation…	.91*** [.74, 1.00][b]	.34* [.03, .66][e]	.94*** [.77, 1.00][b]	.42 [-.11, .95][f]	.63*** [.33, .94][d]	.12 [-.44, .68][f]	--	.40 [-.32, 1.00][f]
P2. Non-Intimate Re…	--	.96*** [.85, 1.00][a]	--	.98*** [.90, 1.00][a]	.98*** [.91, 1.00][a]	.94*** [.80, 1.00][a]	--	.86*** [.62, 1.00][b]
P3. Employment/ Fin…	--	--	.88*** [.55, 1.00][c]	.96*** [.82, 1.00][a]	--	.93*** [.77,1.00][b]	--	.98*** [.91, 1.00][a]
P4. Trauma/ Victimi…	.99*** [.95, 1.00][a]	.99*** [.90, 1.00][a]	--	.99*** [.95, 1.00][a]	.96*** [.90, 1.00][a]	.91*** [.75, 1.00][b]	--	.95*** [.81, 1.00][a]
P5. General Antisoc…	.99*** [.95, 1.00][a]	--	.98*** [.91, 1.00][a]	.95*** [.87, 1.00][a]	.97*** [.90, 1.00][a]	.94*** [.81, 1.00][a]	--	--
P6. Major Mental Dis…	--	.96*** [.83, 1.00][a]	--	.93*** [.82, 1.00][a]	.94*** [.80, 1.00][a]	.89*** [.76, 1.00][b]	--	.90*** [.70, 1.00][b]
P7. Personality Dis…	.95*** [.82, 1.00][a]	.55** [.19, .91][d]	.94*** [.77, 1.00][b]	.49*** [.20, .78][d]	.85*** [.63, 1.00][b]	.27 [-.19, .74][f]	--	.68*** [.44, .93][c]
P8. Substance Use	.98*** [.93, 1.00][a]	.94*** [.77, 1.00][a]	.93*** [.76, 1.00][b]	.92*** [.74, 1.00][b]	.99*** [.95, 1.00][a]	.76*** [.58, .94][b]	--	.96*** [.82, 1.00][a]
P9. Violent/Suicidal…	--	.98*** [.90, 1.00][a]	--	.96*** [.90, 1.00][a]	--	.93*** [.76, 1.00][b]	--	--
P10. Distorted Think…	.92*** [.79, 1.00][b]	.42 [.16, .69][d]	.86*** [.60, 1.00][b]	.43 [-.04, .91][e]	.50* [.01, .99][e]	.40 [-.15, .94][f]	.96*** [.84, 1.00][a]	.44 [-.11, 1.00][f]

Notes. N = 50. M1 = Defence/Distance/Protection, M2 = Justice/Honour/Retribution, M3 = Gain/Profit/Acquisition, M4 = Change/Control/Compliance, M5 = Status/Esteem/Dominance, M6 = Release/Expression/ Emotion, M7 = Arousal/Activity/Excitement, M8 = Proximity/Affiliation/Conformity. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.12 Reliability (AC_1) of Perpetrator Risk Factors Linked to Disinhibitors – Expert Raters

	Disinhibiting Mechanisms AC_1 [95% CI]							
P Factors	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
P1. Intimate Relatio…	.81*** [.67, .96][b]	.98*** [.94, 1.00][a]	.84*** [.71, .97][b]	.85*** [.73, .98][b]	.06 [-.27, .38][f]	.10 [-.22, .42][f]	.15 [-.17, .47][f]	.06 [-.27, .38][f]
P2. Non-Intimate Re…	--	.98*** [.94, 1.00][a]	.73*** [.54, .91][c]	.98*** [.94, 1.00][a]	--	--	--	--
P3. Employment/ Fin…	--	--	.91*** [.82, 1.00][a]	.98*** [.94, 1.00][a]	--	--	--	--
P4. Trauma/ Victimi…	--	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]	.96*** [.90, 1.00][a]	.94*** [.86, 1.00][a]	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]
P5. General Antisoc…	.52*** [.27, .77][d]	--	.98*** [.94, 1.00][a]	--	.86*** [.74, .98][b]	.89*** [.78, .99][b]	.83*** [.70, .97][b]	.78*** [.62, .94][b]
P6. Major Mental…	--	.94*** [.86, 1.00][a]	.91*** [.82, 1.00][a]	.95*** [.89, 1.00][a]	.88*** [.77, .99][b]	.89*** [.78, .99][b]	.86*** [.74, .98][b]	.89*** [.78, .99][b]
P7. Personality Dis…	.47*** [.21, .73][d]	.98*** [.94, 1.00][a]	.94*** [.86, 1.00][a]	.88*** [.76, .99][b]	.80*** [.63, .97][b]	.84*** [.68, 1.00][b]	.84*** [.68, 1.00][b]	.80*** [.63, .97][b]
P8. Substance Use	.98*** [.94, 1.00][a]	--	.98*** [.94, 1.00][a]	.93*** [.85, 1.00][a]	.47** [.21, .74][d]	.57*** [.33, .81][d]	.55*** [.30, .79][d]	.78*** [.62, .95][b]
P9. Violent/Suicidal…	--	.98*** [.94, 1.00][a]	--	.81*** [.65, .96][b]	.98*** [.94, 1.00][a]	.96*** [.90, 1.00][a]	.98*** [.94, 1.00][a]	.91*** [.82, 1.00][a]
P10. Distorted Think…	.26* [-.05, .57][e]	.98*** [.94, 1.00][a]	.91*** [.82, 1.00][a]	.79*** [.63, .95][b]	.96*** [.90, 1.00][a]	.94*** [.86, 1.00][a]	.96*** [.90, 1.00][a]	.93*** [.86, 1.00][a]

Notes. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. D1 = Negative Attitudes, D2 = Negative Self-Concept, D3 = Alienation, D4 = Nihilism, D5 = Lack of Insight, D6 = Lack of Guilt, D7 = Lack of Anxiety, D8 = Lack of Empathy. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.13 Reliability (AC_1) of Perpetrator Risk Factors Linked to Disinhibitors – Novice Raters

	Disinhibiting Mechanisms AC_1 [95% CI]							
P Factors	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
P1. Intimate Relatio…	.85*** [.68, 1.00][b]	.86*** [.65, 1.00][b]	.63*** [.23, 1.00][e]	.89*** [.64, 1.00][b]	.95*** [.85, 1.00][a]	.96*** [.85, 1.00][a]	--	--
P2. Non-Intimate Re…	--	.89*** [.66, 1.00][b]	.57*** [.18, .96][d]	.95*** [.87, 1.00][a]	--	--	--	--
P3. Employment/ Fin…	--	.94*** [.77, 1.00][b]	.75*** [.36, 1.00][c]	.94*** [.83, 1.00][a]	--	--	--	--
P4. Trauma/ Victimi…	.72*** [.46, .98][c]	.96*** [.85, 1.00][a]	.95*** [.83, 1.00][a]	.99*** [.95, 1.00][a]	.79*** [.37, 1.00][c]	.95*** [.82, 1.00][a]	.95*** [.82, 1.00][a]	.95*** [.82, 1.00][a]
P5. General Antisoc…	.52*** [21, .84][d]	.90*** [.64, 1.00][b]	.94*** [.80, 1.00][a]	--	.88*** [.67, 1.00][b]	.88*** [.67, 1.00][b]	.85*** [.57, 1.00][b]	.88*** [.69, 1.00][b]
P6. Major Mental…	.97*** [.89, 1.00][a]	.91*** [.78, 1.00][b]	.96*** [.87, 1.00][a]	.87*** [.65, 1.00][b]	.93*** [.74, 1.00][b]	.96*** [.85, 1.00][a]	.89*** [.72, .99][b]	.98*** [.90, 1.00][a]
P7. Personality Dis…	.43*** [.00, .87][e]	.89*** [.57, 1.00][c]	.88*** [.56, 1.00][c]	.90*** [.70, 1.00][b]	.67*** [.34, 1.00][d]	.71*** [.36, 1.00][b]	.64*** [.15, 1.00][e]	.70*** [.37, 1.00][d]
P8. Substance Use	.96*** [.85, 1.00][a]	--	.99*** [.95, 1.00][a]	.96*** [.85, 1.00][a]	.61 [-.05, 1.00][e]	.91*** [.82, 1.00][c]	.55*** [.14, .95][e]	.91*** [.81, 1.00][b]
P9. Violent/Suicidal…	.99*** [.93, 1.00][a]	.98*** [.91, 1.00][a]	.98*** [.90, 1.00][a]	.71*** [.32, 1.00][d]	.94*** [.72, 1.00][b]	.94*** [.72, 1.00][a]	.94*** [.72, 1.00][b]	.96*** [.81, 1.00][a]
P10. Distorted Think…	.84*** [.57, 1.00][b]	.88*** [.54, 1.00][c]	.96*** [.90, 1.00][a]	.95*** [.77, 1.00][b]	.94*** [.83, 1.00][a]	.95*** [.86, 1.00][a]	.88*** [.76, 1.00][b]	.98*** [.94, 1.00][a]

Notes. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. D1 = Negative Attitudes, D2 = Negative Self-Concept, D3 = Alienation, D4 = Nihilism, D5 = Lack of Insight, D6 = Lack of Guilt, D7 = Lack of Anxiety, D8 = Lack of Empathy. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.14 Reliability (AC_1) of Perpetrator Risk Factors Associated with Destabilizers – Expert Raters

	Destabilizing Mechanisms AC_1 [95% CI]					
P Factors	**De1**	**De2**	**De3**	**De4**	**De5**	**De6**
P1. Intimate Relatio…	--	.96*** [.90, 1.00][a]	--	.78*** [.63, .94][b]	.20 [-.12, .52][f]	.15 [-.17, .48][f]
P2. Non-Intimate Re…	--	--	--	--	.94*** [.86, 1.00][a]	--
P3. Employment/ Fin…	--	--	--	--	--	--
P4. Trauma/ Victimi…	--	--	--	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]
P5. General Antisoc…	--	--	--	.98*** [.94, 1.00][a]	.98*** [.94, 1.00][a]	.91*** [.82, 1.00][a]
P6. Major Mental Dis…	--	.96*** [.89, 1.00][a]	--	.87*** [.75, .99][b]	.88*** [.76, .99][b]	.75*** [.57, .93][c]
P7. Personality Dis…	--	.96*** [.90, 1.00][a]	--	.33* [.03, .63][e]	.69*** [.48, .90][c]	.61*** [.38, .84][c]
P8. Substance Use	--	.94*** [.86, 1.00][a]	--	.08 [-.23, .38][f]	.42** [.14, .70][e]	.71*** [.51, .91][c]
P9. Violent/Suicidal…	--	--	--	.89*** [.78, .99][b]	.91*** [.82, 1.00][a]	.86*** [.75, .98][b]
P10. Distorted Think…	--	.91*** [.82, 1.00][a]	--	-.32 [-.59, -.05][f]	.58*** [.35, .82][d]	.55*** [.31, .80][d]

Notes. N = 50. De1 = Disturbed Attention and Concentration, De2 = Disturbed Sensation and Perception, De3 = Impaired Memory, De4 = Impaired Reasoning, De5 = Obsessive, Perseverative Thoughts, De6 = Impulsive, Intrusive Thoughts. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table B.15 Reliability (AC_1) of Perpetrator Risk Factors Associated with Destabilizers – Novice Raters

	Destabilizing Mechanisms AC_1 [95% CI]					
P Factors	**De1**	**De2**	**De3**	**De4**	**De5**	**De6**
P1. Intimate Relatio…	--	--	--	--	.37 [-.26, 1.00][f]	.94*** [.84, 1.00][a]
P2. Non-Intimate Re…	--	--	--	--	.97*** [.91, 1.00][a]	--
P3. Employment/ Fin…	--	--	--	--	.97*** [.91, 1.00][a]	--
P4. Trauma/ Victimi…	--	--	--	.95*** [.87, 1.00][a]	.97*** [.92, 1.00][a]	--
P5. General Antisoc…	--	--	--	.99*** [.96, 1.00][a]	.99*** [.96, 1.00][a]	--
P6. Major Mental Dis…	.96*** [.85, 1.00][a]	.97*** [.86, 1.00][a]	--	.82*** [.63, 1.00][b]	.82*** [.65, 1.00][b]	.73*** [.48, .99][c]
P7. Personality Dis…	--	.98*** [.90, 1.00][a]	--	.49 [-.04, 1.00][d]	.61*** [.37, .85][c]	.54*** [.21, .87][d]
P8. Substance Use	.98*** [.90, 1.00][a]	.91*** [.81, 1.00][b]	.98*** [.93, 1.00][a]	.47** [.14, .79][e]	.36* [.06, .66][e]	.59** [.17, 1.00][e]
P9. Violent/Suicidal…	--	--	--	.92*** [.76, 1.00][a]	.94*** [.87, 1.00][a]	.94*** [.72, 1.00][b]
P10. Distorted Think…	--	.98*** [.90, 1.00][a]	--	.44 [-.43, 1.00][f]	.51* [.11, .91][e]	.01 [-68, .69][f]

Notes. N = 50. De1 = Disturbed Attention and Concentration, De2 = Disturbed Sensation and Perception, De3 = Impaired Memory, De4 = Impaired Reasoning, De5 = Obsessive, Perseverative Thoughts, De6 = Impulsive, Intrusive Thoughts. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Bipolar weights and unconditional standard errors applied to coefficient calculations. Unspecified values are a result of zero variance. Qualitative interpretations are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Appendix C. Distribution of SARA-V3 N, P, and Relevance Ratings

Table C.1 Presence of SARA-V3 Past and Recent N and P Ratings

	Distribution							
	Past %				Recent %			
SARA-V3 Factors	Y	P	N	O	Y	P	N	O
N1. Intimidation	62	10	28	0	78	4	18	0
N2. Threats	60	6	34	0	80	2	18	0
N3. Physical Harm	74	12	14	0	80	2	18	0
N4. Sexual Harm	10	0	90	0	10	10	80	0
N5. Severe IPV	26	12	62	0	32	4	64	0
N6. Chronic IPV	56	22	22	0	74	14	12	0
N7. Escalating IPV	32	34	34	0	68	12	20	0
N8. IPV-Related…	36	0	64	0	58	0	42	0
P1. Intimate Rela…	88	8	4	0	98	0	2	0
P2. Non-Intimate…	48	14	38	0	48	20	32	0
P3. Employment…	74	24	2	0	76	16	8	0
P4. Trauma…	32	22	46	0	10	24	66	0
P5. General Antisoc…	34	26	40	0	26	34	40	0
P6. Major Mental…	18	24	58	0	16	28	56	0
P7. Personality Dis…	36	20	44	0	38	20	42	0
P8. Substance Use	78	8	14	0	80	6	14	0
P9. Violent/Suicidal…	28	22	50	0	14	42	44	0
P10. Distorted Think…	70	6	24	0	96	4	0	0

Notes. Y = Yes, Present; P = Possibly Present, N = Not present, O = Omit

Table C.2 Presence of SARA-V3 Relevance Ratings

SARA-V3 Factors	Distribution Relevance %			
	Y	P	N	O
P1. Intimate Relationships	100	0	0	0
P2. Non-Intimate Relationships	46	26	28	0
P3. Employment/ Finances	76	16	8	0
P4. Trauma/ Victimization	12	22	66	0
P5. General Antisocial Conduct	38	26	36	0
P6. Major Mental Disorder	14	28	58	0
P7. Personality Disorder	48	10	42	0
P8. Substance Use	82	4	14	0
P9. Violent/Suicidal Ideation	24	34	42	0
P10. Distorted Thinking about IPV	94	6	0	0

Notes. Y = Yes, Present; P = Possibly Present, N = Not Present, O = Omit.

Appendix D. Interrater Reliability of SARA-V3 N and P Factors and Relevance Ratings

Past Ratings

Table D.1 Reliability (AC_2) of SARA-V3 Past Factors Across Domains – All Raters

SARA-V3 Factors Past	All Raters AC_2	Qualitative Standard	*p*	[95% CI]
N1. Intimidation	.84	Substantial	≤.001	[.72, .97]
N2. Threats	.78	Substantial	≤.001	[.66, .90]
N3. Physical Harm	.84	Substantial	≤.001	[.71, .96]
N4. Sexual Harm	.79	Substantial	≤.001	[.68, .90]
N5. Severe IPV	.70	Moderate	≤.001	[.49, .92]
N6. Chronic IPV	.45	Slight	.053	[-.01, .90]
N7. Escalating IPV	.56	Fair	≤.001	[.24, .87]
N8. IPV-Related Supervision Violations	.77	Moderate	≤.001	[.57, .97]
P1. Intimate Relationships	.96	Almost Perfect	≤.001	[.91, 1.00]
P2. Non-Intimate Relationships	.68	Moderate	≤.001	[.42, .94]
P3. Employment/Finances	.68	Moderate	≤.001	[.47, .88]
P4. Trauma/Victimization	.66	Moderate	≤.001	[.47, .85]
P5. General Antisocial Conduct	.74	Moderate	≤.001	[.49, .99]
P6. Major Mental Disorder	.81	Substantial	≤.001	[.63, .98]
P7. Personality Disorder	.72	Moderate	≤.001	[.55, .88]
P8. Substance Use	.94	Almost Perfect	≤.001	[.87, 1.00]
P9. Violent/Suicidal Ideation	.73	Moderate	≤.001	[.53, .92]
P10. Distorted Thinking about IPV	.86	Substantial	≤.001	[.75, .99]

Notes. N = 50. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.2 Reliability (AC_2) of SARA-V3 Past Factors Across Domains – Expert and Novice Raters

SARA-V3 Factors Past	Expert Raters AC_2	[95% CI]	Novice Raters AC_2	[95% CI]
N1. Intimidation	.62***	[.42, .82][c]	.82***	[.70, .94][b]
N2. Threats	.69***	[.53, .84][c]	.62***	[.25, .99][d]
N3. Physical Harm	.91***	[.85, .98][a]	.82***	[.65, 1.00][b]
N4. Sexual Harm	.73***	[.57, .89][c]	.80***	[.67, .94][b]
N5. Severe IPV	.84***	[.74, .94][b]	.68***	[.30, 1.00][e]
N6. Chronic IPV	.77***	[.65, .88][b]	.25	[-.18, .68][f]
N7. Escalating IPV	.76***	[.63, .89][b]	.64***	[.38, .91][c]
N8. IPV-Related Supervision…	.80***	[.65, .95][b]	.77***	[.46, 1.00][c]
P1. Intimate Relationships	.89***	[.80, .98][a]	.95***	[.87, 1.00][a]
P2. Non-Intimate Relationships	.37**	[.14, .61][e]	.60***	[.32, .88][d]
P3. Employment/Finances	.68***	[.53, .84][c]	.74***	[.51, .97][c]
P4. Trauma/Victimization	.81***	[.70, .92][b]	.61***	[.33, .90][d]
P5. General Antisocial Conduct	.92***	[.84, .99][a]	.81***	[.67, .95][b]
P6. Major Mental Disorder	.84***	[.75, .94][b]	.77***	[.52, 1.00][c]
P7. Personality Disorder	.83***	[.72, .94][b]	.68***	[.47, .88][c]
P8. Substance Use	.92***	[.83, 1.00][a]	.93***	[.86, 1.00][a]
P9. Violent/Suicidal Ideation	.77***	[.63, .92][b]	.68***	[.28, 1.00][d]
P10. Distorted Thinking…	.66***	[.46, .86][c]	.94***	[.88, 1.00][a]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.3 Reliability (Various Coefficients) of SARA-V3 Past Nature of Violence Factors – All Raters

N Factors (Past)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.91*** [.86, .96]	.68*** [.51, .84]c	.41** [.16, .67]e	.40** [.13, .67]e	.41*** [.16, .65]d
N2	.87*** [.81, .93]	.59*** [.39, .79]c	.50*** [.27, .71]d	.49*** [.27, .72]d	.51*** [.28, .73]d
N3	.95*** [.91, .98]	.83*** [.73, .94]b	.63*** [.38, .88]c	.61*** [.36, .87]c	.62*** [.37, .88]c
N4	.90*** [.85, .94]	.66*** [.52, .81]c	.45*** [.22, .69]d	.46*** [.23, .69]d	.47*** [.24, .69]d
N5	.92*** [.88, .96]	.75*** [.62, .88]b	.60*** [.37, .82]c	.61*** [.40, .82]c	.61*** [.42, .81]c
N6	.88*** [.83, .93]	.61*** [.44, .79]c	.42*** [.17, .67]d	.39** [.15, .64]e	.41*** [.18, .64]e
N7	.88*** [.84, .93]	.63*** [.49, .77]c	.28* [.00, .56]e	.27* [.02, .52]e	.27* [.01, .52]e
N8	.91*** [.84, .98]	.72*** [.50, .95]c	.62*** [.31, .93]d	.63*** [.33, .92]d	.61*** [.34, .88]d

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.4 Reliability (Various Coefficients) of SARA-V3 Past Nature of Violence Factors – Expert Raters

N Factors (Past)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.81*** [.73, .89]	.39** [.12, .89][e]	.31** [.08, .54][e]	.25 [-.04, .55][e]	.26 [-.03, .56][e]
N2	.85*** [.78, .92]	.52*** [.30, .74][d]	.45*** [.22, .69][d]	.43** [.17, .69][d]	.43** [.17, .70][d]
N3	.95*** [.91, .98]	.83*** [.71, .95][b]	.60*** [.31, .88][d]	.59*** [.30, .87][d]	.60*** [.30, .89][d]
N4	.88*** [.81, .94]	.61*** [.40, .81][c]	.36** [.10, .63][e]	.36** [.09, .63][e]	.37** [.10, .63][e]
N5	.92*** [.88, .96]	.74*** [.60, .89][b]	.62*** [.43, .81][c]	.61*** [.42, .81][c]	.62*** [.42, .82][c]
N6	.88*** [.83, .93]	.66*** [.52, .80][c]	.44*** [.19, .68][d]	.44*** [.19, .68][d]	.44*** [.20, .69][d]
N7	.86*** [.80, .92]	.56*** [.37, .76][c]	.14 [-.13, .40][f]	.10 [-.20, .39][f]	.11 [-.19, .40][f]
N8	.89*** [.82, .97]	.66*** [.44, .89][c]	.57*** [.31, .82][d]	.56*** [.30, .82][d]	.57*** [.31, .83][d]

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.5 Reliability (Various Coefficients) of SARA-V3 Past Nature of Violence Factors – Novice Raters

N Factors (Past)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.91*** [.85, .97]	.72*** [.53, .91][c]	.64*** [.41, .87][c]	.63*** [.40, .87][c]	.62*** [.40, .84][c]
N2	.81*** [.63, 1.00]	.49 [-.01, 1.00][e]	.61** [.23, 1.00][d]	.62** [.24, 1.00][d]	.62*** [.26, .99][d]
N3	.89*** [.78, 1.00]	.69*** [.40, .99][c]	.62*** [.26, .98][d]	.61** [.22, 1.00][d]	.66*** [.36, .95][d]
N4	.89*** [.82, .96]	.64*** [.42, .87][c]	.41* [.08, .75][e]	.44* [.09, .78[e]	.44* [.09, .80][e]
N5	83.67*** [.67, 1.00]	.56* [.10, 1.00][e]	.57** [.18, .97][d]	.58** [.17, .99][e]	.58** [.22, .95][d]
N6	.69*** [.52, .86]	.17 [-.29, .63][f]	.28 [-.09, .65][f]	.24 [-.15, .64][f]	.30 [-.06, .65][e]
N7	.80*** [.67, .92]	.45* [.10, .79][e]	.39* [.03, .74][e]	.37 [-.05, .80][e]	.36 [-.35, 1.00][f]
N8	.87*** [.71, 1.00]	.66** [.23, 1.00][d]	.72*** [.39, 1.00][c]	.72*** [.35, 1.00][d]	.68*** [.38, .98][c]

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.6 Reliability (Various Coefficients) of SARA-V3 Past Perpetrator Risk Factors – All Raters

P Factors (Past)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.95*** [.91, 1.00]	.85*** [.72, .99][b]	.45* [.01, .89][e]	.41* [.01, .82][e]	.45* [.04, .85][e]
P2	.89*** [.84, .93]	.63*** [.48, .79][c]	.48*** [.30, .66][d]	.47*** [.28, .65][d]	.47*** [.27, .66][d]
P3	.85*** [.80, .90]	.60*** [.46, .73][c]	.53*** [.36, .71][d]	.53*** [.36, .71][d]	.54*** [.37, .72][d]
P4	.89*** [.83, .95]	.70*** [.54, .85][c]	.73*** [.58, .88][b]	.73*** [.57, .88][c]	.73*** [.58, .88][b]
P5	.91*** [.84, .97]	.75*** [.57, .92][c]	.78*** [.62, .94][b]	.78*** [.62, .94][b]	.78*** [.63, .92][b]
P6	.90*** [.84, .95]	.72*** [.56, .87][c]	.66*** [.48, .84][c]	.65*** [.48, .82][c]	.66*** [.47, .85][c]
P7	.90*** [.84, .95]	.72*** [.59, .86][b]	.71*** [.56, .86][c]	.71*** [.55, .87][c]	.70*** [.54, .86][c]
P8	.95*** [.92, .99]	.88*** [.78, .97][b]	.86*** [.73, .98][b]	.85*** [.72, .97][b]	.86*** [.73, .98][b]
P9	.89*** [.83, .95]	.71*** [.54, .88][c]	.72*** [.56, .89][c]	.72*** [.55, .89][c]	.74*** [.59, .88][b]
P10	.93*** [.86, .98]	.74*** [.55, .93][c]	.41* [.02, .81][e]	.39* [.06, .71][e]	.40* [.07, .73][e]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.7 Reliability (Various Coefficients) of SARA-V3 Past Perpetrator Risk Factors – Expert Raters

P Factors (Past)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.92*** [.87, .97]	.78*** [.64, .93][b]	.56*** [.32, .80][d]	.56*** [.30, .81][d]	.56*** [.31, .81][d]
P2	.77*** [.68, .85]	.37** [.14, .60][e]	.40*** [.18, .62][d]	.39*** [.16, .62][d]	405*** [.17, .63][d]
P3	.87*** [.81, .92]	.64*** [.49, .80][c]	.59*** [.40, .79][c]	.59*** [.40, .79][c]	.60*** [.40, .79][c]
P4	.93*** [.89, .97]	.80*** [.70, .91][b]	.83*** [.72, .93][b]	.83*** [.72, .93][b]	.83*** [.73, .93][b]
P5	.97*** [.94, 1.00]	.91*** [.83, .99][a]	.92*** [.85, .99][a]	.92*** [.85, .99][a]	.92*** [.85, .99][a]
P6	.91*** [.87, .96]	.77*** [.65, .88][b]	.68*** [.50, .86][c]	.68*** [.50, .86][c]	.68*** [.50, .86][c]
P7	.93*** [.89, .97]	.80*** [.70, .91][b]	.79*** [.68, .91][b]	.79*** [.68, .91][b]	.79*** [.68, .91][b]
P8	.95*** [.90, .99]	.86*** [.73, .99][b]	.82*** [.66, .99][b]	.82*** [.67, .99][b]	.83*** [.66, .99][b]
P9	.89*** [.84, .95]	.71*** [.56, .86][c]	.73*** [.58, .88][b]	.73*** [.58, .88][b]	.73*** [.58, .88][b]
P10	.82*** [.73, .91]	.51*** [.28, .75][d]	.47*** [.24, .71][d]	.46*** [.20, .72][d]	.47*** [.21, .73][d]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.8 Reliability (Various Coefficients) of SARA-V3 Past Perpetrator Risk Factors – Novice Raters

P Factors Past	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.96*** [.90, 1.00]	.88*** [.67, 1.00][b]	.51 [-.18, 1.00][f]	.40 [-.67, 1.00][f]	.44 [-.12, 1.00][f]
P2	.85*** [.75, .95]	.51** [.19, .83][d]	.35 [-.03, .73][e]	.34 [-.05, .73][e]	.36 [-.03, .74][e]
P3	.86*** [.74, .97]	.61*** [.30, .93][d]	.54** [.15, .93][e]	.52* [.18, .92][e]	.56** [.16, .96][e]
P4	.85*** [.75, .95]	.60*** [.32, .88][c]	.64*** [.35, .93][c]	.64*** [.35, .93][c]	.64*** [.38, .90][c]
P5	.92*** [.87, .98]	.79*** [.64, .95][b]	.82*** [.68, .95][b]	.81*** [.67, .95][b]	.83*** [.70, .95][b]
P6	.88*** [.77, .99]	.67*** [.37, .97][d]	.62*** [.30, .95][d]	.62*** [.32, .92][d]	.65*** [.32, .98][d]
P7	.87*** [.80, .95]	.66*** [.46, .86][c]	.65*** [.43, .86][c]	.64*** [.44, .84][c]	.63*** [.37, .89][c]
P8	.96*** [.92, 1.00]	.89*** [.78, 1,00][b]	.88*** [.73, 1.00][b]	.87*** [.71, 1.00][b]	.89*** [.74, 1.00][b]
P9	.85*** [.68. 1.00]	.60* [.14, 1.00][d]	.63** [.22, 1.00][d]	.62** [.21, 1.00][d]	.66*** [.26, 1.00][d]
P10	.95*** [.92, .99]	.85*** [.75, .96][b]	.60*** [.33, .87][d]	.49*** [.22, .77][d]	.45 [-.07, .97][e]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Recent Ratings

Table D.9 Reliability (AC_2) of SARA-V3 Recent Factors Across Domains – All Raters

SARA-V3 Factors Recent	All Raters AC_2	Qualitative Standard	*p*	[95% CI]
N1. Intimidation	.83	Substantial	≤.001	[.62, 1.00]
N2. Threats	.82	Substantial	≤.001	[.67, .97]
N3. Physical Harm	.87	Substantial	≤.001	[.77, .97]
N4. Sexual Harm	.45	Fair	≤.001	[.20, .70]
N5. Severe IPV	.73	Moderate	≤.001	[.57, .90]
N6. Chronic IPV	.70	Moderate	≤.001	[.46, .94]
N7. Escalating IPV	.26	Poor	≤.001	[-.09, .61]
N8. IPV-Related Supervis…	.87	Substantial	≤.001	[.72, 1.00]
P1. Intimate Relationships	.98	Almost Perfect	≤.001	[.96, 1.00]
P2. Non-Intimate Relationships	.79	Moderate	≤.001	[.54, 1.00]
P3. Employment/Finances	.67	Moderate	≤.001	[.45, .88]
P4. Trauma/Victimization	.70	Moderate	≤.001	[.51, .89]
P5. General Antisocial Conduct	.66	Fair	≤.001	[.33, .99]
P6. Major Mental Disorder	.78	Moderate	≤.001	[.57, .98]
P7. Personality Disorder	.78	Substantial	≤.001	[.64, .93]
P8. Substance Use	.92	Almost Perfect	≤.001	[.80, 1.00]
P9. Violent/Suicidal Ideation	.76	Substantial	≤.001	[.48, .88]
P10. Distorted Thinking…	.97	Almost Perfect	≤.001	[.94, 1.00]

Notes. N = 50. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.10 Reliability (AC_2) of SARA-V3 Recent Factors Across Domains – Expert and Novice Raters

SARA-V3 Factors Recent Ratings	Expert Raters AC_2	[95% CI]	Novice Raters AC_2	[95% CI]
N1. Intimidation	.85***	[.72, .98][b]	.76***	[.47, 1.00][d]
N2. Threats	.90***	[.83, .98][a]	.83***	[.57, 1.00][b]
N3. Physical Harm	.72***	[.53, .91][c]	.84***	[.70, .97][b]
N4. Sexual Harm	.53***	[.34, .72][d]	.50*	[.09, .91][e]
N5. Severe IPV	.76***	[.61, .92][b]	.73***	[.45, 1.00][c]
N6. Chronic IPV	.68***	[.48, .87][c]	.70***	[.39, 1.00][d]
N7. Escalating IPV	.42***	[.18, .65][d]	.13	[-.28, .55][f]
N8. IPV-Related Supervision…	.74***	[.54, .93][c]	.91***	[.74, 1.00][b]
P1. Intimate Relationships	.98***	[.95, 1.00][a]	.98***	[.93, 1.00][a]
P2. Non-Intimate Relationships	.70***	[.54, .85][c]	.69***	[.32, 1.00][e]
P3. Employment/Finances	.68***	[.54, .83][c]	.74***	[.55, .92][c]
P4. Trauma/Victimization	.85***	[.75, .95][b]	.66***	[.33, .99][d]
P5. General Antisocial Conduct	.83***	[.69, .96][b]	.56**	[.20, .92][d]
P6. Major Mental Disorder	.85***	[.75, .94][b]	.80***	[.59, 1.00][c]
P7. Personality Disorder	.82***	[.71, .92][b]	.73***	[.52, .95][c]
P8. Substance Use	.93***	[.87, .99][a]	.89***	[.71, 1.00][b]
P9. Violent/Suicidal Ideation	.73***	[.58, .88][b]	.76***	[.56, .96][c]
P10. Distorted Thinking…	.96***	[.90, 1.00][a]	.97***	[.93, 1.00][a]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.11 Reliability (Various Coefficients) of SARA-V3 Recent Nature of Violence Factors – All Raters

N Factors (Recent)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.88*** [.80, .96]	.68*** [.47, .90][c]	.63*** [.40, .85][c]	.55*** [.27, .83][d]	.58*** [.32, .85][d]
N2	.94*** [.90, .98]	.81*** [.69, .93][b]	.67*** [.45, .88][c]	.65*** [.42, .87][c]	.67*** [.45, .90][c]
N3	.93*** [.89, .97]	.77*** [.65, .89][b]	.66*** [.49, .83][c]	.66*** [.48, .83][c]	.67*** [.50, .84][c]
N4	.79*** [.72, .86]	.43*** [.24, .62][d]	.50*** [.32, .67][d]	.48*** [.30, .67][d]	.48*** [.30, .66][d]
N5	.87*** [.80, .93]	.64*** [.47, .81][c]	.69*** [.55, .84][c]	.69*** [.54, .84][c]	.70*** [.55, .86][c]
N6	.82*** [.72, .92]	.52*** [.24, .79][d]	.46*** [.81, .74][d]	.40** [.16, .69][e]	.46*** [.19, .74][d]
N7	.76*** [.65, .87]	.36* [.05, .66][e]	.44*** [.18, .71][d]	.42** [.11, .72][e]	.43*** [.17, .70][d]
N8	.92*** [.85, .99]	.78*** [.59, .97][b]	.83*** [.68, .98][b]	.82*** [.67, .97][b]	.82*** [.67, .97][b]

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.12 Reliability (Various Coefficients) of SARA-V3 Recent Nature of Violence Factors – Expert Raters

N Factors (Recent)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.89*** [.80, .98]	.71*** [.48, .93][c]	.53** [.20, .93][d]	.52** [.19, .85][d]	.53** [.21, .86][d]
N2	.94*** [.90, .98]	.81*** [.68, .94][b]	.67*** [.46, .88][c]	.67*** [.45, .88][c]	.67*** [.46, .89][c]
N3	.85*** [.76, .95]	.60*** [.35, .86][d]	.65*** [.43, .88][c]	.65*** [.43, .88][c]	.66*** [.43, .88][c]
N4	.82*** [.75, .89]	.51*** [.33, .69][d]	.52*** [.33, .70][d]	.51*** [.32, .70][d]	.52*** [.33, .71][d]
N5	.89*** [.82, .95]	.69*** [.52, .87][c]	.72*** [.55, .89][c]	.72*** [.55, .89][c]	.72*** [.55, .89][c]
N6	.84*** [.76, .92]	.57*** [.35, .79][d]	.50*** [.25, .76][d]	.50*** [.25, .76][d]	.51*** [.25, .76][d]
N7	.78*** [.69, .87]	.41*** [.17, .64][d]	.44*** [.22, .66][d]	.44*** [21, .66][d]	.44*** [.22, .67][d]
N8	.86*** [.76, .96]	.72*** [.52, .92][c]	.70*** [.49, .91][c]	.70*** [.49, .91][c]	.70*** [.49, .92][c]

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.13 Reliability (Various Coefficients) of SARA-V3 Recent Nature of Violence Factors – Novice Raters

N Factors Recent	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
N1	.84*** [.67, 1.00]	.57* [.12, 1.00][e]	.54* [.09, .98][e]	.43* [-.06, .93][e]	.53* [.11, .94][e]
N2	.89*** [.74, 1.00]	.70*** [.29, 1.00][d]	.68*** [.26, 1.00][d]	.63*** [.16, 1.00][d]	.69*** [.32, 1.00][d]
N3	.91*** [.84, .98]	.71*** [.48, .94][c]	.57*** [.24, .90][d]	.55*** [.22, .89][d]	.59*** [.30, .88][d]
N4	.78*** [.60, .96]	.41 [-.17, .88][e]	.50* [.09, .92][e]	.50* [.08, .92][e]	.50* [.09, .92][e]
N5	.88*** [.75, 1.00]	.66*** [.32, 1.00][d]	.72*** [.43, 1.00][c]	.72*** [.43, 1.00][c]	.74*** [.48, 1.00][c]
N6	.84*** [.70, .98]	.56** [.18, .94][d]	.52** [.14, .90][e]	.40 [-.03, .83][e]	.54** [.17, .90][d]
N7	.66*** [.49, .82]	.07 [-.37, .51][f]	.24 [.09, .57][f]	.20 [-.19, .58][f]	.25 [.06, .57][f]
N8	.95*** [.86, 1.00]	.87*** [.63, 1.00][b]	.90*** [.71, 1.00][b]	.90*** [.69, 1.00][b]	.89*** [.70, 1.00][b]

Notes. N = 50. N1 = Intimidation, N2 = Threats, N3 = Physical Harm, N4 = Sexual Harm, N5 = Severe IPV, N6 = Chronic IPV, N7 = Escalating IPV, N8 = IPV-Related Supervision Violations. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.14 Reliability (Various Coefficients) of SARA-V3 Recent Perpetrator Risk Factors – All Raters

P Factors (Recent)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.98*** [.96, 1.00]	.95*** [.90, 1.00][a]	.86*** [.66, 1.00][b]	.79*** [.51, 1.00][c]	.82*** [.58, 1.00][b]
P2	.92*** [.86, .98]	.73*** [.54, .93][c]	.59*** [.32, .85][d]	.58*** [.31, .84][d]	.58*** [.31, .84][d]
P3	.86*** [.80, .92]	.62*** [.46, .79][c]	.56*** [.37, .74][c]	.53*** [.34, .72][d]	.54*** [.36, .72][d]
P4	.84*** [.76, .92]	.57*** [.36, .78][d]	.33*** [.15, .51][e]	.27* [.05, .48][e]	.26* [.05, .48][e]
P5	.84*** [.74, .94]	.57*** [.30, .84][d]	.55*** [.33, .77][d]	.52*** [.30, .74][d]	.54*** [.31, .76][d]
P6	.90*** [.83, .97]	.73*** [.55, .91][c]	.71*** [.50, .92][c]	.70*** [.48, .92][c]	.71*** [.52, .91][c]
P7	.92*** [.88, .95]	.78*** [.69, .88][b]	.76*** [.66, .87][b]	.76*** [.65, .87][b]	.75*** [.64, .87][b]
P8	.95*** [.90, 1.00]	.86*** [.73, 1.00][b]	.82*** [.63, 1.00][b]	.92*** [.83, 1.00][b]	.82*** [.63, 1.00][b]
P9	.88*** [.83, .94]	.68*** [.53, .83][c]	.66*** [.47, .85][c]	.66*** [.49, .82][c]	.67*** [.51, .84][c]
P10	.97*** [.95, 1.00]	.93*** [.86, 1.00][a]	.63*** [.39, .86][c]	.47*** [.19, .75][d]	.48*** [.20, .76][d]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.15 Reliability (Various Coefficients) of SARA-V3 Recent Perpetrator Risk Factors – Expert Raters

P Factors (Recent)	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.98*** [.96, 1.00]	.95*** [.88, 1.00][a]	.74*** [.42, 1.00][c]	.74*** [.41, 1.00][c]	.74*** [.42, 1.00][c]
P2	.89*** [.83, .94]	.69*** [.54, .85][c]	.69*** [.53, .85][c]	.69*** [.53, .85][c]	.69*** [.53, .86][c]
P3	.86*** [.80, .92]	.62*** [.47, .78][c]	.52*** [.29, .74][d]	.51*** [.29, .74][d]	.52*** [.29, .74][d]
P4	.90*** [.85, .96]	.73*** [.58, .88][b]	.28 [-.01, .57][e]	.27 [-.02, .57][e]	.39 [-.01, .58][e]
P5	.90*** [.84, .96]	.73** [.56, .90][c]	.65*** [.44, .87][c]	.65*** [.43, .87][c]	.65*** [.43, .87][c]
P6	.92*** [.88, .96]	.78*** [.67, .89][b]	.74*** [.59, .89][b]	.74*** [.59, .89][b]	.74*** [.59, .89][b]
P7	.93*** [.89, .97]	.80*** [.70, .91][b]	.79*** [.67, .90][b]	.78*** [.66, .91][b]	.79*** [.67, .91][b]
P8	.95*** [.92, .99]	.87*** [.68, 1.00][b]	.83*** [.59, 1.00][b]	.82*** [.58, 1.00][b]	.83*** [.59, 1.00][b]
P9	.88*** [.81, .95]	.67*** [.51, .83][c]	.65*** [.47, .83][c]	.65*** [.46, .83][c]	.65*** [.47, .83][c]
P10	.96*** [.91, 1.00]	.89*** [.78, 1.00][b]	.23 [-.19, .64][f]	.21 [-.24, .66][f]	.22 [-.23, .66][f]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.16 Reliability (Various Coefficients) of SARA-V3 Recent Perpetrator Risk Factors – Novice Raters

P Factors Recent	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.98*** [.94, 1.00]	.95*** [.85, 1.00][a]	.87*** [.61, 1.00][b]	.79*** [.36, 1.00][c]	.84*** [.58, 1.00][b]
P2	.88*** [.74, 1.00]	.62** [.18, 1.00][d]	.43 [-.16, 1.00][f]	.43 [-018, 1.00][f]	.44 [-.11, .99][f]
P3	.87*** [.79, .95]	.65*** [.44, .86][c]	.61*** [.36, .85][d]	.57*** [.31, .82][d]	.57*** [.32, .83][d]
P4	.82*** [.68, .96]	.51** [.13, .89][e]	.37 [-.06, .79][e]	.36 [-.10, .82][f]	.35 [-.10, .80][f]
P5	.79*** [.66, .92]	.44* [.09, .79][e]	.46** [.15, .77][d]	.45** [.13, .77][e]	.47* [.09, .85][e]
P6	.90*** [.81, .99]	.73*** [.49, .98][c]	.72*** [.45, 1.00][c]	.72*** [.47, .97][c]	.75*** [.42, 1.00][c]
P7	.90*** [.82, .98]	.72*** [.51, .94][c]	.71*** [.48, .93][c]	.70*** [.49, .92][c]	.69*** [.44, .94][c]
P8	.93*** [.83, 1.00]	.82*** [.54, 1.00][c]	.78*** [.39, 1.00][c]	.77*** [.40, 1.00][c]	.78*** [.34, 1.00][d]
P9	.88*** [.78, .98]	.68*** [.41, .94][c]	.66*** [.41, .91][c]	.64*** [.36, .93][d]	.68*** [.37, .99][d]
P10	.97*** [.93, 1.00]	.93*** [.82, 1.00][a]	.68*** [.32, 1.00][d]	.47 [-.22, 1.00][f]	.49* [.08, .90][e]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Relevance Ratings

Table D.17 Reliability (AC_2) of SARA-V3 Relevance Ratings Across Domains – All Raters

SARA-V3 Factors Relevance Ratings	All Raters AC_2	Qualitative Standard	*p*	[95% CI]
P1. Intimate Relationships	.97	Almost Perfect	≤.001	[.95, 1.00]
P2. Non-Intimate Relationships	.71	Moderate	≤.001	[.50, .91]
P3. Employment/Finances	.55	Fair	≤.001	[.31, .79]
P4. Trauma/Victimization	.71	Moderate	≤.001	[.49, .94]
P5. General Antisocial Conduct	.62	Fair	≤.001	[.26, .98]
P6. Major Mental Disorder	.82	Substantial	≤.001	[.66, .98]
P7. Personality Disorder	.79	Substantial	≤.001	[.59, .99]
P8. Substance Use	.88	Substantial	≤.001	[.77, 1.00]
P9. Violent/Suicidal Ideation	.79	Substantial	≤.001	[.60, .99]
P10. Distorted Thinking about IPV	.98	Almost Perfect	≤.001	[.95, 1.00]

Notes. N = 50. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.18 Reliability (AC_2) of SARA-V3 Relevance Ratings Across Domains – Expert and Novice Raters

SARA-V3 Factors Relevance Ratings	Expert Raters AC_2	[95% CI]	Novice Raters AC_2	[95% CI]
P1. Intimate Relationships	.93***	[.85, 1.00][a]	.96***	[.93, 1.00][a]
P2. Non-Intimate Relationships	.65***	[.47, .82][c]	.66***	[.32, .99][d]
P3. Employment/Finances	.62***	[.41, .84][c]	.53**	[.14, .92][e]
P4. Trauma/Victimization	.84***	[.73, .95][b]	.65***	[.45, .86][c]
P5. General Antisocial Conduct	.79***	[.64, .94][b]	.51*	[.11, .91][e]
P6. Major Mental Disorder	.89***	[.81, .97][a]	.81***	[.66, .96][b]
P7. Personality Disorder	.82***	[.67, .96][b]	.74**	[.23, 1.00][d]
P8. Substance Use	.95***	[.90, 1.00][a]	.88***	[.74, 1.00][b]
P9. Violent/Suicidal Ideation	.81***	[.68, .94][b]	.75***	[.50, 1.00][c]
P10. Distorted Thinking…	.96***	[.91, 1.00][a]	.97***	[.92, 1.00][a]

Notes. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.19 Reliability (Various Coefficients) of SARA-V3 Relevant Perpetrator Risk Factors – All Raters

Relevance Ratings P Factors	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.97*** [.94, .99]	.91*** [.84, .98][a]	.58*** [.27, .89][d]	.36*** [.04, .69][e]	.38*** [.07, .69][e]
P2	.89*** [.82, .95]	.63*** [.43, .84][c]	.42*** [.17, .67][d]	.39** [.15, .64][e]	.40** [.15, .65][e]
P3	.79*** [.70, .89]	.44*** [.18, .69][d]	.45*** [.22, .68][d]	.45*** [.21, .69][d]	.46*** [.24, .68][d]
P4	.86*** [.78, .93]	.62*** [.42, .82][c]	.50*** [.30, .70][d]	.46*** [.24, .67][d]	.45*** [.28, .63][d]
P5	.84*** [.75, .93]	.56*** [.33, .80][d]	.56*** [.35, .78][d]	.55*** [.32, .78][d]	.54*** [.30, .77][d]
P6	.91*** [.85, .97]	.76*** [.61, .91][b]	.72*** [.56, .89][c]	.72*** [.54, .89][c]	.72*** [.55, .90][c]
P7	.89*** [.83, .96]	.71*** [.53, .89][c]	.78*** [.65, .91][b]	.78*** [.64, .91][b]	.78*** [.63, .92][b]
P8	.94*** [.88, .99]	.82*** [.68, .97][b]	.78*** [.59, .97][b]	.77*** [.58, .96][c]	.78*** [.57, .98][c]
P9	.87*** [.80, .94]	.65*** [.47, .82][c]	.48*** [.25, .72][d]	.47*** [.25, .68][d]	.48*** [.25, .70][d]
P10	.97*** [.94, 1.00]	.93*** [.85, 1.00][a]	.40** [.12, .69][e]	.20 [-.01, .42][e]	.21 [-.01, .44][e]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.20 Reliability (Various Coefficients) of SARA-V3 Relevant Perpetrator Risk Factors – Expert Raters

Relevance Ratings P Factors	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.94*** [.87, 1.00]	.88*** [.74, 1.00][b]	.38 [-.17, .93][f]	.37 [-.21, .95][f]	.37 [-.21, .96][f]
P2	.85*** [.79, .92]	.60*** [.42, .78][c]	.52*** [.30, .73][d]	.51*** [.29, .73][d]	.52*** [.30, .74][d]
P3	.85*** [.77, .93]	.59*** [.37, .81][c]	.58*** [.36, .80][d]	.58*** [.36, .81][d]	.59*** [.36, .81][d]
P4	.91*** [.85, .96]	.75*** [.60, .89][b]	.51*** [.29, .73][d]	.50** [.27, .73][d]	.50*** [.28, .73][d]
P5	.89*** [.83, .96]	.71*** [.54, .89][c]	.68*** [.50, .86][c]	.68*** [.49, .87][c]	.68*** [.49, .87][c]
P6	.93*** [.90, .97]	.82*** [.72, .92][b]	.76*** [.60, .91][b]	.75*** [.59, .91][b]	.76*** [.59, .92][b]
P7	.91*** [.85, .98]	.77*** [.60, .94][b]	.82*** [.69, .95][b]	.82*** [.68, .95][b]	.82*** [.68, .96][b]
P8	.97*** [.94, 1.00]	.91*** [.83, .99][a]	.87*** [.73, 1.00][b]	.87*** [.73, 1.00][b]	.87*** [.73, 1.00][b]
P9	.89*** [.82, .95]	.69*** [.52, .87][c]	.54*** [.31, .77][d]	.53*** [.28, .78][d]	.53*** [.28, .78][d]
P10	.97*** [.92, 1.00]	.91*** [.79, 1.00][a]	.27 [-.21, .75][f]	.25 [-.25, .76][f]	.26 [-.25, .77][f]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table D.21 Reliability (Various Coefficients) of SARA-V3 Relevant Perpetrator Risk Factors – Novice Raters

Relevance Ratings P Factors	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
P1	.97*** [.94, 1.00]	.92*** [.83, 1.00][a]	.68*** [.42, .95][c]	.48* [.07, .89][e]	.49* [.11, .87][e]
P2	.84*** [.70, .99]	.50* [.03, .96][e]	.29 [-.33, .92][f]	.30 [-.31, .91][f]	.30 [-.32, .93][f]
P3	.79*** [.63, .95]	.44* [.01, .87][e]	.46 [-.01, .93][e]	.47* [.02, .92][e]	.48 [-.03, .98][e]
P4	.82*** [.73, .91]	.51*** [.28, .75][d]	.44** [.14, .75][e]	.42* [.09, .75][e]	.42** [.13, .70][e]
P5	.78*** [.64, .92]	.40* [.03, .77][e]	.44** [.11, .76][e]	.42* [.08, .77][e]	.39* [.00, .77][e]
P6	.90*** [.83, .97]	.72*** [.53, .91][c]	.71*** [.50, .91][c]	.71*** [.50, .91][c]	.73*** [.53, .92][c]
P7	.86*** [.59, 1.00]	.62 [-.11, 1.00][f]	.71** [.19, 1.00][d]	.71* [.16, 1.00][d]	.71** [.23, 1.00][d]
P8	.93*** [.86, 1.00]	.81*** [.61, 1.00][c]	.77*** [.56, .99][c]	.77*** [.55, .98][c]	.76*** [.52, 1.00][c]
P9	.86*** [.72, .99]	.61*** [.25, .98][d]	.45 [-.04, .95][e]	.43 [-.15, 1.00][f]	.45 [-.33, 1.00][f]
P10	.98*** [.92, 1.00]	.93*** [.79, 1.00][a]	.42 [-.08, .91][e]	.06 [-1.00, 1.00][f]	.07 [-.25, .38][f]

Notes. N = 50. 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. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Appendix E. Interrater Reliability of SARA-V3 Scenario Planning Considerations

Table E.1 Reliability (AC_2) of Additional Repeat Scenario Planning Considerations – All Raters

Summary Rating	All Raters AC_2	Qualitative Standard	*p*	[95% CI]
Degree of Psychological Harm	.78	Substantial	≤.001	[.63, .93]
Degree of Physical Harm	.68	Moderate	≤.001	[.47, .90]
Potential Victim Acquaintanceship	.88	Almost Perfect	≤.001	[.80, .97]
Likelihood	.83	Substantial	≤.001	[.67, .97]
Imminence	.72	Moderate	≤.001	[.51, .92]

Note. N = 50. **p* ≤ .05, ***p* ≤ .01, ****p* ≤ .001. Ordinal weights and unconditional standard errors applied to coefficient calculations, with the exception of Victim Acquaintanceship, which is unweighted. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table E.2 Reliability (AC_2) of Additional Repeat Scenario Planning Considerations – Expert and Novice Raters

Summary Rating	Expert Raters AC_2	[95% CI]	Novice Raters AC_2	[95% CI]
Degree of Psychological Harm	-.34***	[-.63, -.06][f]	.74***	[.36, 1.00][c]
Degree of Physical Harm	.73***	[.59, .86][b]	.56*	[.08, 1.00][e]
Potential Victim Acquaintanceship	.86***	[.72, 1.00][b]	.88***	[.73, 1.00][b]
Likelihood	.52***	[.26, .78][d]	.80***	[.41, 1.00][c]
Imminence	.82***	[.73, .92][b]	.63***	[.41, .86][c]

Note. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations, with the exception of Victim Acquaintanceship, which is unweighted. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Appendix F. Distribution and Interrater Reliability of SARA-V3 Conclusory Opinions

Table F.1 Distribution (%) of SARA-V3 Conclusory Opinions

Summary Rating	Low/Routine	Moderate/Elevated	High/Urgent
Case Prioritization	0	48	52
Risk for Serious Harm	10	64	26
Imminent Violence	12	64	24
Other Risks Indicated	36	52	12

Table F.2 Reliability (AC_2) of SARA-V3 Conclusory Opinions – All Raters

Summary Rating	All Raters AC_2	Qualitative Standard	*p*	[95% CI]
Case Prioritization	.84	Substantial	≤.001	[.74, .93]
Risk for Serious Harm	.72	Moderate	≤.001	[.58, .87]
Imminent Violence	.58	Fair	≤.001	[.32, .84]
Other Risks Indicated	.78	Substantial	≤.001	[.72, .84]

Note. N = 50. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table F.3 Reliability (AC_2) of Qualitative SARA-V3 Conclusory Opinions – Expert and Novice Raters

Summary Rating	Expert Raters AC_2	[95% CI]	Novice Raters AC_2	[95% CI]
Case Prioritization	.49***	[.23, .75][d]	.82***	[.62, 1.00][c]
Risk for Serious Harm	.76***	[.65, .88][b]	.63***	[.29, .98][d]
Imminent Violence	.79***	[.67, .92][b]	.44*	[.10, .77][e]
Other Risks Indicated	.59***	[.43, .75][c]	.77***	[.55, .98][b]

Note. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table F.4 Reliability (Various Coefficients) of Qualitative SARA-V3 Conclusory Opinions – All Raters

Summary Rating	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
Case Prioritization	.92*** [.87, .96]	.77*** [.64, .90][b]	.52*** [.27, .78][d]	.51*** [.24, .78][d]	.51*** [.27, .75][d]
Risk for Serious Harm	.87*** [.82, .92]	.66*** [.52, .79][c]	.50*** [.33, .68][d]	.49*** [.32, .66][d]	.50*** [.30, .70][d]
Imminent Violence	.81*** [.73, .90]	.49*** [.27, .72][d]	.36*** [.14, .58][e]	.29* [.02, .55][e]	.32* [.06, .57][e]
Other Risks Indicated	.91*** [.88, .93]	.71*** [.63, .78][b]	.41*** [.23, .58][d]	.41*** [.24, .58][d]	.42*** [.27, .57][d]

Notes. N = 50. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table F.5 Reliability (Various Coefficients) of Qualitative SARA-V3 Conclusory Opinions – Expert Raters

Summary Rating	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
Case Prioritization	.74*** [.60, .88]	.49*** [.23, .75][d]	.49*** [.23, .75][d]	.49*** [.23, .75][d]	.49*** [.24, .75][d]
Risk for Serious Harm	.88*** [.83, .94]	.69*** [.56, .82][c]	.50*** [.29, .71][d]	.49*** [.27, .71][d]	.50*** [.29, .71][d]
Imminent Violence	.89*** [.83, .94]	.69*** [.54, .85][c]	.42** [.16, .69][e]	.41* [.13, .69][e]	.42** [.14, .70][e]
Other Risks Indicated	.84*** [.77, .91]	.56*** [.40, .72][b]	.47*** [.27, .67][d]	.48*** [.28, .68][d]	.47*** [.26, .67][d]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*

Table F.6 Reliability (Various Coefficients) of Qualitative SARA-V3 Conclusory Opinions – Novice Raters

Summary Rating	% Agreement [95% CI]	Brennan & Prediger [95% CI]	Cohen/ Conger κ [95% CI]	Scott/Fleiss κ [95% CI]	Krippendorff α [95% CI]
Case Prioritization	.91*** [.81, 1.00]	.75*** [.48, 1.00][c]	.49* [.00, .97][e]	.47 [-.06, 1.00][d]	.47 [-.03, .97][e]
Risk for Serious Harm	.84*** [.71, .98]	.58** [.22, .94][d]	.42* [.00, .83][e]	.42 [.00, .85][e]	.46* [.02, .89][e]
Imminent Violence	.78*** [.65, .91]	.41* [.06, .76][e]	.34 [-.06, .75][e]	.33 [-.13, .78][f]	.34 [-.00, .69][e]
Other Risks Indicated	.90*** [.81, .99]	.68*** [.40, .97][c]	.36 [-.16, .89][e]	.36 [-.26, .97][e]	.39*** [.04, .74][d]

Notes. $N = 50$. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Ordinal weights and unconditional standard errors applied to coefficient calculations. Qualitative classifications are probabilistic (see Gwet, 2014; Klein, 2018) and benchmarked with Landis and Koch (1977): a = *almost perfect,* b = *substantial,* c = *moderate,* d = *fair,* e = *slight,* f = *poor.*