THE ASSOCIATION BETWEEN STRUCTURED PROFESSIONAL JUDGMENT MEASURE TOTAL SCORES AND SUMMARY RISK RATINGS: IMPLICATIONS FOR PREDICTIVE VALIDITY

A Dissertation

Presented to

The Faculty of the Department of Psychology and Philosophy

Sam Houston State University

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

---

by

Caroline S. Chevalier

August, 2017

THE ASSOCIATION BETWEEN STRUCTURED PROFESSIONAL JUDGMENT MEASURE TOTAL SCORES AND SUMMARY RISK RATINGS: IMPLICATIONS FOR PREDICTIVE VALIDITY

by

Caroline S. Chevalier

_____

APPROVED:

Marcus Boccaccini, PhD
Dissertation Director

James Crosby, PhD
Committee Member

Samuel Hawes, PhD
Committee Member

Jorge Varela, PhD
Committee Member

Abbey Zink, PhD
Dean, College of Humanities and Social Sciences

DEDICATION

I would like to dedicate the completion of my dissertation to my family. To Mom and Dad: your support and encouragement allowed me to pursue higher education and believe that I could be successful. And to my husband, Adam: Okay. What's next?

ABSTRACT

Chevalier, Caroline S., *The association between structured professional judgment measure total scores and summary risk ratings: Implications for predictive validity*. Doctor of Philosophy (Clinical Psychology), August, 2017, Sam Houston State University, Huntsville, Texas.

Structured professional judgment (SPJ) instruments are used by mental health professionals to assess risk for future violence and assess treatment needs. Current literature tends to examine SPJ instruments in a way that is not congruent with how the instruments are designated to be used in the field. Specifically, studies often leave out analyses of the structured professional judgment piece of the instrument (summary risk rating, SRR) and, when the analysis is included, authors rarely compare the SRR to the actuarially derived total score. This study sought to provide practitioners with a comparison of the SPJ measure total scores and SRRs. I conducted a review of the literature to find studies that measured the predictive validity of SPJ measure total scores and SRRs. I requested additional data from corresponding authors in order to compare the two scores using varying statistical methods. In total I included 69 samples (n = 10,871). I performed several meta-analyses to determine if a) the predictive validity of the total score and SRR were similar, and b) if the SRR adds any additional predictive power to the total score. Findings suggest that the total score and SRR have similar predictive abilities. The small difference between the mean weighted SRR (AUC = .701) and total score (AUC = .698) effect sizes was not significant. I also calculated a z-score to test the difference between the SRR and total score effect size in each sample, and found a statistically significant difference in only 8 of the 69 samples. However, a meta-analysis of odds ratio values from logistic regression models including effects for both total scores and SRRS revealed a consistent incremental validity effect for SRRs (OR =

1.96, $p < .001$) over total scores. Overall, this review provides evidence to suggest that the total score and SRR provide similar predictive effects, but also reveals that using the SRR is worthwhile for practicing clinicians. Implications for both research and practice are discussed.

KEY WORDS: Structured professional judgment, Risk assessment, Predictive validity

ACKNOWLEDGEMENTS

I would likely to acknowledge Dr. James Hanley and Dr. David Wilson for their assistance in applying the appropriate statistical analyses used in this study.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

Introduction

Courts and other agencies tasked with treating individuals with mental illness or guarding public safety often ask psychologists, psychiatrists, and other mental health professionals to provide opinions about how likely a patient or offender is to engage in specific behaviors in (e.g., violence, sexual deviance, and treatment adherence) in the future. Mental health professionals perform various types of risk assessments in both hospital and correctional settings, where the goal is to protect patients, staff, and the public. Because some of the factors associated with risk are thought to be changing, or dynamic, treatment providers can use information from risk assessments to inform an individual's course of treatment and mitigate or decrease future risk. For these reasons, risk assessment is a large and growing area of research and development, with at least 400 different assessment instruments used in risk assessment evaluations across the world (Singh et al., 2014).

Because the relationship between any one risk factor (e.g., history of violence) and future behavior (e.g., committing future violence) is usually only moderate to small in size, evaluators often consider many risk factors before coming to a decision about an individual's level of risk (Hanson & Morton-Bourgon, 2009). The most often discussed and researched approaches for combining these factors into a final opinion on future risk are unstructured clinical judgment, actuarial judgment, and structured professional judgment (SPJ). The current study focused on measures evaluators use when following the SPJ approach.

Structured Professional Judgment

Structured Professional Judgment (SPJ) measures require evaluators to rate the individual being assessed on a set of factors (i.e., items) that the instrument developers places on the measure because of their empirical, or scientific, association with future behavior (e.g., future violence). The evaluator considers the item scores in addition to individual or contextual factors of the individual being assessed and produces a final or summary risk rating (SRR) that speaks to the potential for the individual to engage in the specific behavior. For most measures, the SRR options are high, moderate, or low risk.

Unlike actuarial instruments, which require evaluators to base their decision on the instrument's numerical score, SPJ instruments encourage evaluators to use discretion when making a final risk decision. The manual for the Historical, Clinical, Risk-20, Version three (HCR-20^{V3}; Douglas, Hart, Webster, & Belfrage, 2013) explains the formulation of a risk opinion as follows:

> The determination of individual relevance of risk factors… starts this process by having evaluators consider risk factors as they apply to the individual at hand. Formulation furthers this process by requiring evaluators to integrate separate risk factors into a conceptually meaningful framework that explains a person's violence. 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 best to intervene and manage such risk (Douglas et al., 2013, p. 53-54).

The developers of the Structured Assessment of Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2006) argue that the SPJ approach is best suited for

risk assessment "because it (a) is anchored in the empirical and professional literature, (b) allows for the appropriate consideration of developmental factors, and (c) emphasizes the dynamic, and often contextual, nature of risk" (Borum et al. 2006, p. 4). The authors note that while the individual risk factors are grounded in empirical literature, the clinician should be able to integrate contextual factors specific to the individual being assessed into their final risk estimate. Even though the summated numerical rating may have classified someone as high risk, the evaluator has the ability to increase or decrease the final risk rating in accordance with dynamic and contextual factors.

When clinicians use an SPJ approach, their final risk decision is commonly referred to as a summary risk rating (SRR). Although terms may vary slightly depending on the measure (e.g., overall risk estimate, etc.), this paper will use the SRR label globally to apply to any clinically derived final risk judgment applied to a total score. For most SRR measures, the evaluator makes a SRR of low, moderate, or high risk after considering the score assigned across the SPJ items and other potentially relevant information. In this way, the SPJ tools have been said to integrate the best parts of both clinical judgment and actuarial decision-making. SPJ measures encourage the clinician to focus on empirically supported risk factors, but allow the clinician to consider individual differences via clinical expertise (Doyle & Dolan, 2002).

Comparison of SPJ SRRs and Summated Scores

Because actuarial/mechanical measures (i.e., measures that use a score based ont e result of an equation to predict specific behaviors) generally outperform unstructured clinical judgment (Grove, Zald, Lebow, Snitz, & Nelson, 2000), one possible limitation of SPJ measures is that the incorporation of clinical judgment when making SRRs results

in attenuated predictive validity. In other words, the uninterpreted sum of the SPJ items may be a stronger predictor of future behavior than the judgment influenced SRR. If the addition of clinical judgment does lead to decreased predictive validity, we would expect actuarial measures—which do not allow for the alteration of final risk conclusions – to outperform SPJ measures in studies that measure predictive validity. If the addition of clinical judgment improves predictive validity, we would expect SPJ measures to outperform actuarial measures, as the inclusion of clinical judgment would add incremental validity to the actuarial or numerical total score. If the additional judgment neither increases nor decreases predictive validity, we would expect SPJ measures and actuarial measures to perform similarly.

Findings from meta-analyses comparing effects from SPJ and actuarial measures have come to somewhat different conclusions. A meta-analysis of predictive effects for risk assessment measures designed to predict sexual violence found that effects for scores from actuarial tools (d = .67) were similar to those for an SPJ measure (SVR-20) total score (d = .66), but stronger than those from the measure's SRR (d =.46; Hanson & Morton-Bourgon, 2009). However, an unpublished review of the broader violence risk assessment literature found that effects for SPJ tools (AUC = .71) were as strong as effects for actuarial tools (AUC = .68; Guy, 2008). However, the author averaged the SRR and total score AUC values for each sample to produce a single effect size, clouding any possible differences between SRR and total score performance. A third review also concluded that SPJ measures performed as well as actuarial measures, and even identified one specific SPJ tool, the SAVRY (Borum et al., 2006), as a stronger predictor than

related actuarial measures, but did not compare or report separate effects for total scores and SRRs (Singh, Grann, & Fazel, 2011).

An important limitation of some studies examining the predictive validity of SPJ measures that the authors provide predictive validity statistics for the total score, but not for the SRR. In these studies, researchers typically add the ratings on each item and use a summated score (i.e., total score) in their predictive validity analyses. In this way, the researchers are studying the SPJ measure as if it were going to be used as an actuarial measure, leaving out the integration of any final clinical judgment. In one meta-analysis, 18.5% (k = 5) of the 27 samples included reported effects using only the mechanical scoring method (Singh et al., 2011). Because the studies sometimes omit the inclusion of clinical judgment, it is difficult to ascertain if validity studies can properly translate into the field where clinicians, as directed by the SPJ instrument manuals, use clinical judgment when providing their SRR.

There are studies, however, that have reported predictive effects for both SPJ total scores and SRRs. Findings from these studies are mixed, with some suggesting comparable effects for SPJ total scores and SRRs and others suggesting that one tends to perform better than the other. For example, Arbach-Lucioni, Andrés-Pueyo, Pomarol-Clotet, and Gomar-Soñes (2011) reported a violent recidivism AUC of .77 for both the Historical Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1007) total score and SRR. Dolan and Rennie (2008) examined predictive effects for the SAVRY (Borum et al., 2006) and found the total score and SRR to have identical predictive abilities for both violent recidivism (AUC = .64) and general recidivism (AUC = .69).

Alternatively, de Vogel and de Ruiter (2005) reported a much higher AUC value for the SRR (.86) than the total score (.59) on the HCR-20.

Several measure specific meta-analyses have also reported separate effects for SPJ total scores and SRRs, but have failed to directly compare the predictive validity of total scores to the predictive validity of SRRs. O'Shea, Mitchell, Picchioni, and Dickens (2013) examined HCR-20 studies from 20 studies using inpatient samples and found a larger effect for SRRs predicting inaptient aggression (d = 1.16) than total scores (d = .65). O'Shea and Dickens (2014) performed a meta-analysis of nine studies examining the predictive validity of the START, and provided mean weighted AUC values for each total score and SRR. They found the effect for both the vulnerability (i.e., risk) total score and the SRR for physical aggression toward others to be large (.727, .760).

In a prior review of predictive validity findings for nine different SPJ measure (57 total samples), I found that the absolute value of the difference between SRR and total score AUC values ranged from -.14 (higher total score AUC) to .27 (higher SRR AUC; Chevalier & Boccaccini, 2015). The effect for the total score was larger in 48 comparisons, while the effect for the SRR was larger in 29 comparisons, but most differences (61%) were small (less than .05). Although these findings suggest the overall predictive abilities of the total score and SRR are similar, a meta-analysis would allow a direct comparison between the effects for the two values in addition to an examination of moderator variables that may help explain when the differences tend to be larger or smaller.

Incremental Validity

Another way that SPJ researchers study the utility of SRRs is to examine whether the SRR adds incremental validity to the total score in a regression model. For example, Neves, Gonçalves, and Palma-Oliveira (2011) used sequential binary logistic regression to compare the HCR-20 total score and SRR. They showed that the addition of the SRR increased the quality of the model. They framed their findings in this way:

> Regarding general and nonviolent recidivism, SPJ and actuarial scores showed similar positive predictive performance. Such close performance does not necessarily imply that using the HCR-20 as an actuarial or SPJ tool is indifferent… The HCR-20 was not designed to be used as an actuarial tool and the user may be confident the SPJ approach will lead to predictive results similar or better than the traditional actuarial assessment, while gaining in case management practical utility concerns (p. 146).

Several studies have found that the HCR-20 SRR significantly improved model fit when added to the total score (de Vogel, de Ruiter, Hildebrand, Bos, & de Van, 2004; de Vogel & de Ruiter, 2006; Douglas, Ogloff, & Hart, 2003' Ho et al., 2003; Neves, Conçalves, & Palma-Oliveira, 2011; Pederson, Rasmussen, & Elsass, 2010). Other studies examining the SAVRY (Borum et al., 2006) found that the SRR does not add significant incremental validity to the total score (Dolan & Rennie, 2008; Hilterman, Nicholls, & van Niewenhuizen, 2014; Schmidt, Campbell, & Houlding, 2011).

Although the studies reporting incremental validity provide some support for SRRs, it is difficult to come to firm conclusions about the value of SRRs on the basis of these studies because many other SPJ studies do not report incremental validity analyses.

One possibility is that researchers only report incremental validity findings when there is an incremental effect. There may be many other studies with no incremental effect of SRRs over total scores, leading to an overall pattern of similar performance for total scores and SRRs in most studies.

Current Study

Existing research seems to provide generally positive support for the utility of SPJ measure SRRs in risk assessment, as SRRs tend to be moderate-sized, statistically significant predictors of future violence and aggression. Nevertheless, it appears that the effects for SPJ total scores and SRRs are often similar, raising questions about the need for the inclusion of SRRs when the total score may be as useful for describing risk. In some studies, effects for SRRs appear to be significantly larger than those for total scores or SRRs add incremental validity to total scores in prediction models, but many SPJ measure studies have not examined whether SRRs are significantly stronger predictors than total scores or reported the results of incremental validity analyses. Because SPJ measures encourage evaluators to use SRRs instead of total scores for decision-making, it is important to understand if and when SRRs outperform total scores.

The goal of this study was to conduct a comprehensive review and meta-analysis of SPJ measure results, focusing on the question of whether SRRs are more useful than summated total scores for prediction. For this review, I conducted a thorough literature search in order to compile a list of SPJ studies examining the predictive validity of both SPJ measure total scores and SRRs. I contacted SPJ researchers and asked them to provide additional statistical information about predictive validity in their studies, including incremental validity analysis results and information needed to test the

difference between SRR and total score effects (see Hanley & McNeil, 1983). I used area under the receiver operating characteristic curve (AUC) values as the primary measure of effect size. In the context of risk assessment research, the AUC value indicates the likelihood that the risk score of a randomly selected recidivist will be higher than that of a randomly selected nonrecidivist. An AUC value of .50 is equivalent to chance-level prediction, and indicates that the risk assessment has no better ability to correctly identify a recidivist versus a nonrecidivist than chance alone. Rice and Harris (2005) provided the following benchmarks for AUC value interpretation: small (.556) medium (.639), and large (.714).

My goal was to use these data to answer a series of questions about SPJ measure SRRs and total scores. First, is there any evidence of statistically significant differences between AUC values for SRRs and total scores? To examine this question, I used the equations from Hanley and McNeil (1983) to find the z-score of the difference between the total score AUC and SRR AUC for each SPJ study. Although it is possible for researchers to test whether the AUC for the SRR is significantly different than the AUC for the total score in any individual study, few researchers conduct this type of comparison (see e.g., Chu et al., 2012; Lodewijks et al., 2008b). Indeed, researchers reported this comparison for only 2 of the 69 samples I was able to include in this review. Calculating these z scores allowed me to examine how common it was for the difference between AUCs to be statistically significant at the individuals study level. I also conducted a meta-analysis of AUC difference scores from these comparisons, which I calculated for each study by subtracting the smaller AUC value from the larger AUC value.

Second, is there any evidence that AUC values tend to be larger for SRRs than total scores? To examine this question, I first conducted a meta-analysis of AUC difference scores calculated by subtracting the total score AUC from the SRR AUC, with positive difference scores indicating a stronger effect for the SRR and negative difference scores indicating a stronger effect for the total score. If SRRs add something meaningful beyond total scores alone, the mean difference score across SPJ studies should be positive. If the clinical judgment inherent in making SRRs weakens predictive validity, the mean difference score should be negative. I also conducted separate meta-analyses of SRR AUC values and total score AUC values from the studies included in the difference score analyses. Although these meta-analyses do not allow for direct comparisons between SRRs and total scores, they provide information about the mean AUC values for SRRs and total scores among these studies and allow for an examination of factors that may explain variability in SRR and total score effects across samples.

Third, is there any evidence that SRRs consistently add incremental validity to the prediction of outcomes beyond total scores alone? Because SPJ researchers do not always report the results from incremental validity analyses, it is unclear whether the handful of reported incremental effects for SRRs reflect typical performance for SPJ measures or something unique about the studies that have reported incremental effects. To examine this question, I collected incremental validity results from research reports ($k = 6$), but also asked authors from other studies to conduct and provide the results of incremental validity analyses for this review. Authors provided additional data for 17 samples, all of which are new to the published research literature. My primary goal was to conduct a qualitative review of this literature, by providing information about how common it was

for there to be an incremental effect for SRRs over total scores. My secondary goal was, if possible (i.e., enough studies), to use odds ratio values from the incremental validity analyses to conduct a meta-analysis of the incremental effects.

CHAPTER II

Method

Search Strategy

Before searching for specific SPJ studies, I created a list of SPJ measures to guide my searches. The list of SPJ measures was based on previously reviews (Guy, 2008; Singh et al., 2011; Hanson & Morton-Bourgon, 2009) and a PsycInfo search using the search terms "structured professional judgment" and "SPJ*". I identified 22 measures SPJ measures (see Appendix C). I excluded one SPJ measure (SAPROF; De Vogel, De Ruiter, Bouman, & De Vries Robbé, 2007) from further literature searches because the measure focuses on protective factors rather than risk, and is designed to be used only in conjunction with another risk measure.

After I compiled a list of SPJ measures, I searched for studies using the measure's abbreviated title (e.g., START for the Short Term Assessment of Risk and Treatability) and the wildcard character (e.g., START*) in PsycInfo and ProQuest (Dissertation and Theses full text Global). I also examined reference sections from published studies, advanced online publication emails, and annotated bibliographies (available online, e.g., HCR-20 annotated bibliography found at https://kdouglas.wordpress.com/hcr-20/hcr-20-overview-and-annotated-bibliography/, Douglas, et al., 2014) to find additional studies.

I searched for studies until May 2017. The search methods yielded a total of 154,986 results. I screened the full-text article or report when the study abstract implied the examination of predictive validity. I screened 172 full text articles for eligibility.

Inclusion criteria. Because my goal was to compare effects for SRRs and total scores, I only included studies that examined the predictive validity of both SRRs and

total scores among the same sample of patients or offenders. I excluded 104 articles from further analyses because they did not examine the predictive validity of SRRs (e.g., Green et al., 2016, Vitacco et al., 2016). I excluded two articles because they appeared to treat multiple scores from the same patient as independent cases for the calculation of AUC values (Chu, Daffern, and Ogloff , 2013; Griffith, Daffern, & Godber, 2013). I excluded one study because I was unable to locate study information necessary to calculate the standard error for the AUC values (Gibas, 2008). I excluded five studies because they provided AUC values for samples that had been used in other studies (Dickens & Oshea, 2015; Vincent, Guy, Gershenson, & McCabe, 2012b; Wilson, Desmarais, Nicholls, & Brink, 2010; Worling, Bookalam, & Littljohn, 2012). In these instances, I included the most recent study or the study where the outcome variable most closely matched the intended purpose of the SPJ measure. For studies that included more than one eligible SPJ instrument, I used the AUC values for the instrument whose purpose most closely matched the measured outcome. This only happened in one instance, where I chose to use the predictive validity data for the START as opposed to the HCR-20 (Wilson et al., 2013). If a study reported effects for multiple samples, I coded each sample separately for the meta-analysis. For example, some studies provided separate predictive validity analyses for male versus female participants (Augimeri et al., 12012; Lodewijks et al., 2008; Oshea & Dickens, 2015; Penney et al., 2010). Each of these studies provided effects for two samples, resulting in four studies providing eight samples. In other cases, the author provided separate predictive validity analyses for samples categorized by diagnoses (Fitzgerald et al., 2013; Oshea et al., 2015; each providing two samples), legal status (Michel et al., 2013, provided two samples) or age

(Vincent et al, 2012, provided three samples). Together, these four studies provided nine samples. In total, 52 studies reported effects for one sample and eight studies provided multiple samples, yielding a total of 69 samples for analysis. For studies that reported effects for multiple outcome variables, I used the outcome variable that most closely matched the outcome the measure was designed to assess (i.e., violence for the HCR-20). For example, in studies where both sexual violence and any violence was measured using the SVR-20, I only used the sexual violence outcome because that outcome most closely matched the intended purpose of the instrument (e.g., see Dempster, 1998). After examining 172 studies, I excluded 112 studies because they failed to meet inclusion criteria. I included findings from 60 studies, with a total of 69 samples (in final analyses (see Table 1).

Table 1

Overview of Studies and Samples Included

Study	*N*	SPJ Measure	Outcome
Storey et al., 2014	249	B-SAFER	Intimate Partner Violence
Augimeri et al., 2012	573	EARL-20B	Criminal Offending
Augimeri et al., 2012	294	EARL-21G	Criminal Offending
Chu et al., 2012	104	ERASOR	Sexual Recidivism
Morton, 2003	78	ERASOR	Sexual Recidivism
Rajlic and Gretton, 2010	286	ERASOR	Sexual Recidivism
Skowron, 2004	220	ERASOR	Sexual Recidivism
Viljoen et al., 2009	193	ERASOR	Sexual Recidivism
Worling et al., 2015	191	ERASOR	Sexual Recidivism
Arbach-Lucioni et al., 2011	78	HCR-20	Physical Aggression
De Vogel et al., 2004	120	HCR-20	Violent Offending
De Vogel and de Ruiter, 2006	127	HCR-20	Physical Violence
Douglas et al., 2003	100	HCR-20	Any Violence

(continued)

Study	*N*	SPJ Measure	Outcome
Douglas et al., 2005	188	HCR-20	Violent Recidivism
Fitzgerald et al., 2013, ID Group	25	HCR-20	Physical Aggression
Fitzgerald et al., 2013, Control Group	45	HCR-20	Physical Aggression
Gunenc et al., 2015	613	HCR-20	Any Verbal Aggression
Hilterman et al., 2011	195	HCR-20	General Offending
Ho et al., 2013	110	HCR-20	Any Violence
Jovanovic et al., 2009	104	HCR-20	Any Violent Behavior
Langton et al., 2009	44	HCR-20	Physical Aggression
Michel et al., 2013, Forensic Group	150	HCR-20	Aggressive Behavior
Michel et al., 2013, General Group	98	HCR-20	Aggressive Behavior
Neal et al., 2015	230	HCR-20	Contact and Threat Violence
Neves et al., 2011	158	HCR-20	Violent Behavior
O'Shea et al., 2014	504	HCR-20	Self-Harm
O'Shea et al., 2015, ID Group	109	HCR-20	Physical Violence, Others
O'Shea et al., 2015, Control Group	504	HCR-20	Physical Violence, Others

(continued)

Study	*N*	SPJ Measure	Outcome
Pederson et al., 2010	107	HCR-20	Violent Crime
Pederson et al., 2012	81	HCR-20	Violent Recidivism
Sada et al., 2016	225	HCR-20	Violent Behavior
Schaap et al., 2009	45	HCR-20	Violent Recidivism
Strub et al., 2016	100	HCR-20	Any Violence
Verbrugge et al., 2013	59	HCR-20	Violent Recidivism
Wilson et al., 2013	30	HCR-20	Aggression
Hogan and Olver, 2016	90	HCR-20^{V3}	Imminent Violence
Strub et al., 2014	106	HCR-20^{V3}	Violence
Andres-Puueyo et al., 2008	102	SARA	Intimate Partner Violence
Belfrage et al., 2012	429	SARA	Further Contact with Police
Kropp and Hart, 2000	251	SARA	Intimate Partner Violence
Dolan and Rennie, 2008	99	SAVRY	Violent Recidivism
Hilterman et al., 2014	105	SAVRY	Violent Reoffending
Khanna et al., 2014	109	SAVRY	Violent Reconvictions

(continued)

Study	*N*	SPJ Measure	Outcome
Lodewijks et al., 2008, Boys	47	SAVRY	Violent Recidivism
Lodewijks et al., 2008, Girls	35	SAVRY	Violent Recidivism
Lodewijks et al., 2008b	66	SAVRY	Physical Violence against Persons
McEachran, 1999	108	SAVRY	Violent Recidivism
Penney et al., 2010, Boys	80	SAVRY	Violent Recidivism
Penney et al., 2010, Girls	64	SAVRY	Violent Recidivism
Schmidt et al., 2011	128	SAVRY	Violent Offenses
Shepherd et al., 2014	175	SAVRY	Violent Reoffenses
Viljoen et al., 2008	169	SAVRY	Any Nonsexual Violent Offense
Vincent et al., 2012, < 12 Years Old	38	SAVRY	New Petitions, Violence
Vincent et al., 2012, 13 to 15 Years Old	254	SAVRY	New Petitions, Violence
Vincent et al., 2012, 16 to 18 Years Old	382	SAVRY	New Petitions, Violence
Chu et al., 2011	50	START	Interpersonal Violence
Desmarais et al., 2012	119	START	Any Aggression
Gray et al., 2011	44	START	Violence to Others

(continued)

Study	*N*	SPJ Measure	Outcome
Lowder et al., 2017	550	START	Incarceration
Lowder et al., 2017b	95	START	General Offending
O'Shea and Dickens, 2015, Men	149	START	Physical Aggression
O'Shea and Dickens, 2015, Women	51	START	Physical Aggression
O'Shea et al., 2016	200	START	Physical Aggression
Quinn et al., 2013	80	START	Aversive Incidents
Troquete et al., 2015	163	START	Violence
Viljoen, 2012	90	START:AV	Any Offense
Dempster, 1998	95	SVR-20	Sexual Recidivism
De Vogel et al., 2004	122	SVR-20	Sexual Recidivism
Sjöstdedt and Långström, 2002	51	SVR-20	Sexual Recidivism

Data Requests

For each eligible study, I created a data request sheet with four sections (see Appendix A). The purpose of the data request sheet was to obtain data from the corresponding author that were not included in the original research report. I filled in sections on the data request sheets that were already provided in the research report, to streamline the data request process.

At the top of the form, I listed the complete reference for the study. This section also included spaces for listing the sample size, follow-up time, and outcome variable. The next section of the form asked authors to report AUC and SE values for both total score and SRR. For those authors whose studies examined an SPJ measure with more than one total score, I asked them to include AUC and SE values for both total scores. Once again, I listed these values if they were included in the original research report. Acquiring SE values from study authors provided more accuracy than attempting to calculate values myself without the raw data. These values were used in calculating the z-score difference between AUC values outlined in the proposed analyses section. The next section asked authors to calculate and provide a correlation (*r*) between the total score and SRR for the overall sample, and then the same statistic for both the recidivists and nonrecidivists. For those authors whose studies examined an SPJ measure with more than one total score, I asked the authors to calculate the correlations for both total scores (resulting in six correlations). I used the correlation values for the recidivists and nonrecidivists in the formula to find the z-score difference between AUC values for total score and SRR.

Finally, I asked authors who did not include analyses of incremental validity to run and report findings for such analyses. I asked authors to use hierarchical logistic regression for these analyses and to report the following values: B, SE, Wald statistic, p value, Odds ratio (OR), OR confidence interval, model chi-square, and chi-square change. I asked authors to run three models. For the first model, I asked the authors to use only the total score as a predictor. For the second model, I asked authors to use only the SRR as a predictor. For the third model, I asked authors to use both the total score and the SRR as predictors. For studies that examined SPJ measures with more than one total score, I asked authors to include four models: One model for the first total score, one model for the second total score, one model for the SRR, and one model including all three scores.

I contacted each corresponding author through electronic mail (e-mail) using the contact information provided on each published study. See Appendix B for an example of the e-mail template. I contacted each author, provided a two-month time period for a response and sent a second e-mail if I did not receive a response from the initial e-mail contact. Fourteen corresponding authors responded to my request with data for 20 samples. One sample was not used in final analyses due to sample overlap (de Vogel et al., 2005). Several authors who responded to my request chose to send raw data sets (de Vogel & de Ruiter, 2005; de Vogel & de Ruiter, 2006; de Vogel, de Ruiter, Hildebrand, Bos, & van de Ven, 2004; Gray et al., 2011; Sada, 2016). One raw data set was from a study examining the START (Gray et al., 2011), and the rest of the raw data sets examined the HCR-20. The final author response rate to my data request was 23%.

SPJ Measures

For 12 of the 22 SPJ measures, I could not locate any studies that reported effects for both the total score and SRR. The ten SPJ measures with studies examining their predictive validity in this meta-analysis are: Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp, Hart, & Belfrage, 2005), Early Assessment Risk List (EARL-20B) for boys (Augimeri, Koegl, Webster & Leveene, 2001) and EARL-21G for girls (Leveene et al., 2001), Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR; Worling & Curwin, 2001), Historical, Clinical, Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997), HCR-20, Version 3 (HCR-20^{V3}; Douglas, Hart, Webster, & Belfrage, 2013), Spousal Assault Risk Assessment Guide (SARA; Kropp, Hart, Webster, & Eaves, 1995), Structured Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2003), Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Middleton, 2004), Short-Term Assessment of Risk and Treatability, Adolescent Version (START:AV; Nicholls, Viljoen, Cruise, Desmarais, & Webster, 2010), and the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997). See Table 2 for an overview of the samples by SPJ measure.

Table 2

Sample Characteristics by SPJ Measure

SPJ Measure	*k*	Total *N*	Total Score AUC range	SRR AUC range
B-SAFER	1	249	.70	.65
EARL(20B/21G)	2	867	.64-.65	.59-.63
ERASOR	6	962	.59-.77	.53-.83
HCR-20	26	4,259	.43-.88	.44-.92
HCR-20^{V3}	2	205	.76-.77	.73-.75
SARA	3	633	.60-.77	.57-.87
SAVRY	15	1,836	.58-.84	.51-.89
START	10	1,502	.47-.79	.55-.85
START:AV	1	90	.70	.69
SVR-20	3	268	.49-.80	.56-.83

B-SAFER. The Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp, Hart, & Belfrage, 2005) was adapted from the SARA (Kropp, Hard Webster, & Eaves, 1994), another SPJ tool that measures the risk for intimate partner violence. It was designed specifically for use by police officers. The measure includes 10-items and asks police officers to rate offenders on two sections, each with five questions: Spousal Assault (related to the perpetrator's history of spousal violence), and Psychosocial Adjustment (related to the perpetrator's history of psychological and social functioning. Officers code B-SAFER items as either present, absent, or possible or partially present. After rating each item, the officer then provides final judgments of low, medium, or high on imminent risk (i.e., within two months), long-term risk (i.e., after two months), and risk of extremely severe or lethal violence (Kropp et al., 2005). Only one

B-SAFER article was included in final analyses, with a total score AUC value of .70 and an SRR AUC of .65 (Storey, Kropp, Hart, Belfrage, & Strand, 2014).

EARL. The Early Assessment Risk List (EARL-20B) for boys (Augimeri, Koegl, Webster, & Leveene, 2001) and EARL-21G for girls (Leveene et al., 2011) are SPJ instruments designed to aid in risk identification and clinical risk management. The EARL-20B is a 20-item measure, and the EARL-21G is a 21-tem measure that closely parallels the 20B, but has two unique items and is missing one item from the 20B. The tools are similar, but were created to be gender specific. Both were created to assess boys boys' and girls' risk for engaging in future antisocial or criminal behavior. The SRR (called the Overall Clinical Judgment rating for this measure; OCJ) has three levels: low, moderate, and high. One study with two samples was included in final analyses. AUC values for total score ranged from .64 (Augimeri et al., 2012, EARL-20B) to .65 (Augimeri et al., 2012, EARL-21G) and AUC values for SRRs ranged from .59 (Augimeri et al., 2012, EARL-21G) to .63 (Augimeri et al., 2012, EARL-20B).

ERASOR. The Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR; Worling & Curwen, 2001) was designed to measure risk of adolescent sexual reoffending. Specifically, it was created for use with adolescents aged 12 to 18 years. There are 25 items (nine static items and 16 dynamic items) that fall into one of five categories: sexual interests, attitudes, and behaviors, historical sexual assaults, psychosocial functioning, family/environmental functioning, and treatment. The ERASOR is the only SPJ instrument designed to assess juvenile sexual recidivism. Six ERASOR samples from six studies were included in final analyses, with total score AUC

values ranging from .59 (Morton, 2003) to .77 Worling et al., 2015), and SRR AUC values ranging from .53 (Morton, 2003) to .83 (Chu et al., 2012).

HCR-20. The Historical, Clinical, Risk Management-20 (HCR-20; Webster et al., 1997) is comprised of 20 items split between three scales; the Historical scale with ten items, the Clinical scale with five items, and the Risk Management scale with five items. These three scales are meant to specifically address past, present, and future correlates of violence demonstrated in the literature concerning violence risk. The HCR-20 was specifically developed for violence risk assessment and, according to Douglas and colleagues (2008) can be applied in a variety of settings. Twenty-three studies and 26 samples evaluating the predictive validity of the HCR-20 were included in final analyses. Total score AUC values range from .43 (Neal et al., 2015) to .88 (Wilson et al., 2013), and SRR AUC values ranged from .44 (Neal et al., 2015) to .92 (Sada et al., 2016).

HCR-20^{V3}. The Historical, Clinical, Risk Management-20, Version 3 (Douglas et al., 2013) is an updated version of the HCR-20 version two (Webster et al., 1997). In addition to rating whether or not specific risk factors are present or absent, the measure also asks evaluators to provide a relevance rating. The authors describe relevance ratings as, "… the relevance of the risk factors with respect to the development of future risk management strategies. By relevance, we mean the extent to which the factor is critical to the evaluator's formulation of what caused the evaluee to perpetrate violence and how best to prevent future violence" 9Douglas et al., p. 50). For the current study, I included only presence ratings in final analyses to prevent sample overlap. Two samples from two studies were included in final analyses. Total score AUC values ranged from .76 (Hogan

& Olver, 2016) to .77 (Strub et al., 2014), and SRR AUC values ranged from .73 Strub et al., 2014) to .75 (Hogan and Olver, 2016).

SARA. The Spousal Assault Risk Assessment Guide (SARA; Kropp, Hart, & Eaves, 1994) is a 20-item SPJ measure designed to assess risk for both general violence and, more specifically, intimate partner violence (IPV). The items are grouped into four sections: criminal history, psychosocial adjustment, spousal assault history, and alleged/current offense. Items from the first two groups are considered when estimating risk to others, and items from the second two groups are considered when estimating risk of IPV. While the outcome variables are described as imminent risk, the user manual does not define the word imminent or indicate the SARA can be used for long-term prediction, which some researchers have described as problematic (Helmus & Bourgon, 2011). Three samples from three studies were included in final analyses. Total score AUC values ranged from .60 (Kropp & Hart, 2000) to .77 (Andres-Pueyo et al., 2008), and SRR AUC values ranged from .57 (Belfrage et al., 2012) to .87 (Andres-Pueyo et al., 2008).

SAVRY. The Structured Assessment for Violence Risk in Youth (SAVRY; Borum et al., 2003) is a tool designed specifically to measure risk of violence in adolescents. The SAVRY contains historical items, social/contextual items, protective items, and individual risk factors. While the protective factors are not included in the total score (used for research purposes), they are considered when making structured professional judgment using the identified SAVRY risk variables. Even though the SRR for these measures provides an overall risk estimate based on both risk factors and protective factors, I was only able to use one total score from each sample in analyses in

order to prevent including samples multiple times. I chose to include risk total scores in these cases because it more closely matched the total scores included for other measures, which overwhelmingly consider risk as opposed to protective factors. Therefore, the SAVRY protective factor totals will not be considered in this meta-analysis. Fifteen samples from 11 studies were included in final analyses. Total score AUC values ranged from .58 (VIljoen et al., 2008) to .84 (Lodewijks et al, 2008; girl sample), and SRR AUC values ranged from .51 (Viljoen et al., 2008) to .89 (McEachran, 1999).

START. The Short-Term Assessment of Risk and Treatability (START; Webster et al., 2009) is comprised of 20 dynamic items related to risk (termed vulnerabilities) and protective factors (termed strengths). Strengths and vulnerabilities are scored for each item on two three-point scales. The measure is deemed short-term because each outcome estimate is only intended to be used for a maximum of three months post assessment. The START is meant to be given repeatedly in order to track change over time and estimate dynamic risk for violence. The current study utilizes only the vulnerability total score in analyses. Even though the SRR for these measures provides an overall risk estimate based on both strength and vulnerability scores, I was only able to use one total score from each sample in analyses in order to prevent including samples multiple times. I chose to include vulnerability total scores in these cases because it more closely matched the total scores included for other measures, which overwhelmingly consider risk as opposed to protective factors. Ten samples from nine studies were included in final analyses. Vulnerability total score AUC values ranged from .47 (Lowder, 2017b) to .79 (Desmarais et al., 2012), and SRR AUC values ranged from .55 (Lowder, 2017b) to .85 (Oshea & Dickens, 2015; Women sample).

START:AV. The Short Term Assessment of Risk and Treatability: Adolescent Version (START:AV; Nicholls et al., 2010) is a SPJ tool designed to assess for the risk of adverse outcomes for adolescents. The START:AV was adapted from the START and adds variables related to adolescent recidivism, such as family and peer systems, and adapts items specifically geared toward adult functioning (such as occupational questions) with items that are more meaningful for youth (such as questions related to school). As stated previously, only *vulnerability* scores will be used when analyzing the predictive validity of the START:AV. One sample from one study was included in final analyses. The total score AUC value was .70 and the SRR AUC value was .69

SVR-20. The Sexual Violence Risk-20 (SVR-20; Boer et al., 1997) was designed to assess the risk of sexual violence of adult sex offenders. The measure is composed of 20 items and three sections: psycho-social adjustment, sexual offenses, and future plans. In 2003 a more updated version of the SVR-20 called the Risk for Sexual Violence Protocol (RSVP; Hart et al., 2003) was introduced, and has become the most commonly used structured professional judgment tool for assessing risk for sexual violence (Sutherland et al., 2012). Unfortunately, no articles meeting inclusion criteria could be located for this updated measure. Tree samples from three studies were included in final analyses. Total score AUC values ranged from .49 (Sjöstdedt & Långström, 2002) to .80 (deVogel et al., 2004), and SRR AUC values ranged from .56 (Sjöstdedt & Långström, 2002) to .83 (deVogel et al., 2004).

Moderators

I coded several study and sample characteristics that might help to explain variability in effect size values across samples. These moderator variables included

sample size, outcome variable (i.e., violence, sexual violence), recidivism base rate, mean age, sex (i.e., male, female, combined), country, study design (i.e., field, research), mean follow-up time, SPJ measure, correlation between total score and SRR, target sample (i.e., adult, adolescent), and measure authorship (i.e., authorship allegiance). See Table 3 for an overview of moderators used in analyses and the number of studies providing data for the moderator variable.

Table 3

Overview of Data Available for Moderator Analyses

	AUC Meta-Analysis		OR Meta-Analysis	
Moderator	*k*	*n*	*k*	*n*
Sex	69	10,871		
Female	5	489		
Male	34	4,966	11	2080
Combined	30	5,416	11	1360
Study Design	68	10,646		
Field	15	3,360	6	978
Research	53	7,286	16	1,958
Country	60	9,624	20	3,301
Australia	3	284		
Canada	20	2,919	9	1,768
Netherlands	9	920	3	410
Spain	3	285		
Sweden	3	729		
United Kingdom	14	2,576	5	601
United States	9	1,911	3	522

(continued)

Moderator	AUC Meta-Analysis		OR Meta-Analysis	
	k	*n*	*k*	*n*
Outcome	62	8,580	21	2,867
Violence	42	5,759	13	1684
Sexual	10	1,320	4	664
Aggression	10	1,501	4	519
SPJ Target Age	69	10,871	23	3,684
Adult	45	7,116	14	1,886
Adolescent	24	3,755	9	1,798
SPJ Measure	63	9,460	15	1,998
ERASOR	6	962		
HCR-20	26	4,259	5	860
SARA	3	633		
SAVRY	15	1,836	4	412
START	10	1,502	6	726
SVR-20	3	268		
Incremental validity in original report	69	10,871		
Yes	6	692		
No	63	10,179		
Authorship Allegiance	69	10,871	23	3,684
Yes	17	3,321	5	1,176
No	52	7,550	18	2,558
Correlation between SRR and total score	25	4,422	19	3,292

(continued)

Moderator	AUC Meta-Analysis		OR Meta-Analysis	
	k	*n*	*k*	*n*
Age	60	9,531	23	3684
Recidivism Base Rate	62	9,962	23	3684
Follow-up Time	67	10,004	21	2867

Sample size. Sample size was the number of participants included in the predictive validity analysis used to calculate the AUC value.

Outcome variable. I defined outcome variable as the dependent variable each author was trying to predict with SPJ measure results (e.g., recidivism, violence; see Table 1). For the moderator analyses, I grouped studies into three mutually exclusive outcome variable groups: aggression ($k = 10$), violence ($k = 42$), and sexual ($k = 10$). Five samples were not included in the analysis for this moderator variable because the outcome variable did not fit into one of these three categories (e.g., criminal offending; Augimeri et al., 2012 20B and 21G; general offending, Hilterman et al., 2011, Lowder et al., 2017b; incarceration, Lowder et al., 2017; self-harm, Oshea et al., 2014, any aversive incident, Quinn et al., 2013). I classified studies as having aggression as the outcome variable when the authors used the term "aggression" (i.e., physical aggression, verbal aggression, overall aggression). I classified studies as having violence as the outcome variable when the authors used the term "violence" (i.e., physical violence, overall violence, violent recidivism), with the exception of sexual violence. I classified studies as having sexual offenses as the outcome variable when the authors used the term word "sexual" (i.e., sexual violence, sexual recidivism).

Recidivism base-rate. I defined the recidivism base rate as the percentage of the sample that was classified as having engaged in the behavior assessed by the outcome variable.

Mean age. I defined mean age as the mean age of the sample as reported by the study author(s). Some study authors reported mean ages for various time-points within the study. For example, some studies reported mean-age at time of admission, some reported mean age at the time of the SPJ assessment, and some did not identify at what time point the mean age was taken. I used whichever mean age the author chose to report, regardless of time point.

Sex. I defined sample sex as the sex of the participants that made up the sample. Samples were comprised of all males, all females, or both males and females. I coded this moderator as either "male" (k = 34) for all males, "female" (k = 5) for all females, or "combined" (k = 30) for samples that included both males and females.

Country. I defined the country moderator as the country where the study was conducted. Categories with more than three applicable samples included: Australia (k = 3), Canada (k = 20), Netherlands (k = 9), Spain (k = 3), Sweden (k = 3), United Kingdom (k = 14), and United States (k = 9). Countries that did not include enough samples to qualify as a separate category included Serbia (Jovanovic et al, 2009), Portugal (Neves et al., 2011), Mexico (Sada et al., 2016), China (Ho et al., 2013), Denmark (Pederson et al., 2010, Pederson et al., 2012), and Singapore (Chu et al., 2012). One study included samples from multiple countries (Michel et al., 2013) and therefor was not used.

Study design. I classified studies has either field studies (k = 15) or non-field studies (k = 53). The defining feature of field studies is that the SPJ measure was scored

for real-world use (e.g., sentencing, release, treatment planning). In non-field studies, the SPJ measure is scored for research purposes only. I was not able to code field study status for one study because the category it fell into was unclear (Sada et al., 2016).

Mean follow-up time. I coded the mean follow-up time for each sample. I defined the mean follow-up time as the number of months between release and the collection of outcome data. Studies provided either a standard follow-up time for every participant (i.e., all participants were followed for 12 months), or a mean follow-up time (i.e., participants were followed for an average of 24.6 months). I did not include two samples in mean follow-up time moderator analyses because the information was unavailable (Augimeri et al., 2012).

SPJ measure. I defined the SPJ measure as the measure examined by the study. If the study used multiple SPJ measures, I coded the moderator as the SPJ measure for which AUC value I used in AUC meta-analyses. I only examined this moderator in the AUC difference analyses, because the main purpose of this study was examining the comparison between the two scores (total score and SRR) rather than meta-analyzing the individual effects. Not all samples (k = 63) were included in SPJ moderator analyses. The samples that were not included did not have enough samples to form a separate category (EARL20B/21G, Augimeri et al., 2012; B-SAFER, Storey et al., 2014; HCR-20^{V3}, Hogan & Olver, 2016, Strub et al., 2016; START:AV, Viljoen et al., 2012).

Correlation between total score and SRR. I defined the correlation between the total score and SRR as the correlation value provided by study authors to describe the relationship between the total score and SRR ratings in the study sample. I used Pearson *r* correlation values, and coded the value to two decimal places. Only six research reports

provided this correlation. I obtained the correlation for 20 additional samples from study authors, yielding a total of 26 samples with total score/SRR correlations. Correlations ranged from .39 (Oshea et al., 2016) to .89 (Neves et al., 2011). The average correlation (*r* to *z* transformed) between total scores and SRRs was .72. I transformed the *r* values into *z* values for use in all moderator analyses.

Target sample. I defined target sample as the age range the measure was intended to be used for. Therefore, I defined "adolescent" measures as the EARL20B/21G, SAVRY, ERASOR, and START:AV. I defined "adult" measures as the B-SAFER, the HCR-20, the HCR-20^{V3}, the SARA, the START, and the SVR-20. All samples (k = 69) were included in these moderator analyses.

Incremental validity. I defined the incremental validity variable as whether or not the study author reported the results of an incremental validity analysis in their original publication. In order to be placed in the "yes" category, the study needed to include a logistic regression where total score and SRR were included in the same model, and had enough statistical information to allow for a calculation of an odds ratio for the incremental effect. Specifically, the article had to have either B, expB, and a corresponding *p* value. There were six studies that provided this information in the original research report.

Authorship allegiance. I coded each sample as to whether or not the study was authored or co-authored by one of the creators of the SPJ measure being examined. All of the samples were used in the moderator analysis, with 17 samples having a co-author that assisted in the created of the SPJ measure examined in the study.

Analysis Strategy

Measures of effect size. This meta-analysis used the area under the receiver operating characteristic curve (AUC) value as the primary measure of effect size. AUC values provide information about the probability that the total score or SRR will correctly rank a randomly chosen positive instance (i.e., recidivist) higher than a randomly chosen negative one (i.e., non-recidivist). For the purposes of risk assessment, the AUC value indicates the likelihood that the risk score of a randomly selected recidivist will be higher than that of a randomly selected nonrecidivist. An AUC value of .50 is equivalent to chance-level prediction, and indicates that the risk assessment has no better ability to correctly identify a recidivist versus a nonrecidivist than chance alone. Rice and Harris (2005) provided the following benchmarks for AUC value interpretation: small (.556) medium (.639), and large (.714). Because AUC analyses are common in risk assessment research, most research reports included AUC values. I had to transform *r* values to AUC values for one study (Skowon et al., 2004), and one study author supplied AUC values that were not included in the original research report (Augimeri et al., 2012). For the meta-analysis of incremental validity findings, I used odds ratio (i.e., expB) values from logistic regression as the measure of effect size. For each study, the odds ratio values came from a single logistic regression model that included both the total score and the SRR as predictors. In this two predictor model, the odds ratio (OR) value provides a measure of the association between the type of score (total score or SRR) and the outcome (e.g., violence) after controlling for the shared variance among the two types of scores. Specifically, the OR value provides information about the estimated increase in the odds of the outcome for a one unit increase in the value of the predictor. OR values

of 1.00 indicate no effect. OR values greater than 1.00 indicate that the likelihood of the outcome increases as the value of the predictor variable increases. OR values less than 1.00 indicate that the likelihood of the outcome decreases as the predictor increases.

Meta-analytic methods. I used 'metafor,'a Meta-Analysis Package for R (Viechtbauer, 2010) to conduct the AUC value, AUC difference score, and OR meta-analyses. The 'rma.uni' function fits the meta-analytic fixed- and random/mixed-effects models with or without moderators via linear models. The software package requires the effect size and standard error (SE) value for the effect size to calculate meta-analytic effects. The program uses the SE value to calculate the inverse variance weight for each effect size and weights each effect by the inverse variance. For the overall meta-analytic effects, the package automatically applies a random-effects model. A random-effects model is typically used when conducting a meta-analysis because it is assumed the researcher is testing a random selection of studies from a larger population. When moderators are included, the R package applies a mixed-effects model. In the mixed-effects model, the model assumes that the random effects follow a normal distribution. In a mixed-effects model, the moderator variables are scaled to the appropriate zero-value with the results following a normal distribution. The SE values for AUCs came from research reports, data request sheets, or were estimated using an online calculator (http://www.anaesthetist.com/mnm/stats/roc/Findex.htm) ($k = 20$). For AUC difference values, I used the following formula to calculate the SE of the difference between the AUC values, per James Hanley (personal communication, March 14, 2017):

$$SE_{AUCDif} = \sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}$$

SE^1 corresponds to the SE of the SRR AUC, and SE^2 corresponds to the SE of the total score AUC. The correlation *r* represents the correlation between the total score AUC and the SRR AUC, which is calculated using the procedures described by Hanley and McNeil (1983). To calculate the correlation between the two AUC values, you need two other correlations: a) the correlation between total scores and SRRs among those who were positive on the outcome variable (e.g., recidivated), and b) the correlation between total scores and SRRs among those who were negative on the outcome variable (e.g., did not recidivate). Because authors never report these correlations in their research reports, I had to contact each study author in an attempt to obtain this information. Authors provided these correlations for 20 samples. There were two reasons why I could not obtain an *r* value for some samples: Either the author did not respond to my data request, or the values provided by authors were outside the range of values in the table provided by Hanley and McNeil (1983). The table provides approximate *r* values for average AUC values ranging from .70 to .975 (found along the X axis of the column), and average correlation between recidivists and nonrecidivists (take the average between the *r* of total score and SRR for nonrecidivists and the *r* of total score and SRR for recidivists) from .02 to .9 (along the y axis). In some cases, studies had average AUC values that were too low to allow me to find an adequate representative *r* value. Without the *r* value, I would have been unable to complete the calculations for the AUC difference standard error. On the recommendation of James Hanley (personal communication March, 2017), I found a representative *r* value using the five raw datasets authors provided subsequent to my data requests. He provided the following method: Using the raw data sets, I ran ROC analyses in R using the pROC package (Robin et al., 2011). Once the AUC values

for the total score and SRR were calculated, I calculated the covariance of the two paired ROC curves using the 'cov.roc' function. I then calculated the SE of both AUC values by using the 'var.roc' function, which calculates the variance of an ROC curve. Using the variance of each ROC curve, I calculated the SE of each AUC by finding the square root of the variance of each AUC value. Finally, I calculated *r* using the following formula:

$$r = \frac{Covar(AUC_1, AUC_2)}{SE_1 \times SE_2}$$

Where AUC_1 and SE_1 correspond to the AUC and SE of the SRR, and AUC_2 and SE_2 correspond to the AUC and SE of the total score. I utilized the median *r* value from the raw data sets (.57) to calculate the z-score for each set of AUC scores from the 50 samples for which the study authors did not respond to my data request and the 11 where the average AUC value was outside of the value range in the Hanley and McNeil (1983) table. Once I had *r* values for all samples (both the representative *r* value, the *r* values from the raw data, and the *r* values obtained using the table provided by Hanley and McNeil, 1983), I was able to estimate SE values for each AUC difference. For the OR meta-analysis, I calculated SE values of for the log of the OR using OR confidence intervals and p values (when the CI was not available). First, I transformed the expB and corresponding 95% confidence intervals into the log of the expB and CI values. Next, I used the following formula to find the standard error (David Wilson, personal communication, June 6, 2017):

$$SE_{log} = \frac{(UbCI_{log} - LbCI_{log})/2}{1.96}$$

UbCIlog is the upper bound of the 95% confidence interval of the log of the OR and LbCIlog is the lower bound of the 95% confidence interval of the log of the OR. For studies that did not provide the confidence intervals ($k = 4$), I was able to calculate those values using the β and corresponding SE value. I used the following formula to calculate the expB confidence intervals.

$$expB\ 95\%CI = \beta \pm 1.96 \times SE_{\beta}$$

I conducted all analyses using the log transformed OR values (logits), which transformed back into ORs for effect size reporting. The software package reports a mean weighted effect size, standard error, confidence interval, and *p* value test of significance. The report also provides several indicators of study heterogeneity, or the extent to which the effect sizes vary across samples. Cochran's Q statistic tests the null hypothesis that there is no variability between the samples in the analysis other than that expected by sample error alone. The I^2 statistic provides an estimate of the overall heterogeneity across the samples. Low (25%) moderate (50%) and high (75%) I^2 values provide an estimation of how much variability across studies is not attributable to sampling error or chance (Higgins, Thompson, Deeks, & Douglas, 2003). Moderate to high levels of heterogeneity suggest that moderator variables may help to explain variability in the observed effect size estimates. I ran moderator analyses for AUC, AUC difference, and OR meta-analyses in R using 'metafor,' a meta-analysis package for R (Viechtbaur, 2010). The moderator analysis provides a QE value, which tests for model heterogeneity and has a corresponding *p* value. The statistic QM is the Q test for model fit, which also has a corresponding *p* value. If there is a significant moderator effect, QM will be

significant ($p < .05$). If the moderator effect is significant, it means the model accounts for a significant amount of heterogeneity across studies. The output also provides an unstandardized regression coefficient for the moderator variable, which represents the effect of a one unit change in the predictor variable on the outcome variable. For example, if the estimate for the moderator of the age moderator is .02, that implies that for every one year increase in age there is a .02 increase in the AUC value. If the age moderator is significant, this implies the .02 increase is a statistically significant increase. For categorical variables, the output provides an estimate that is equal to the distance from the intercept. The intercept is the AUC value for the reference group. For example, if the intercept weighted AUC is .67, and the categorical variables has a value of -.01, that implies that the weighted mean for the categorical variable is .66.

Z-score of the difference between AUC values. I used the z-score formula published by Hanley and McNeil (1983) to directly compare two AUC values from the same sample. Although there are other methods for testing the difference between two AUC values from the same sample (e.g., DeLong, DeLong, & Cleark-Pearson., 1988), those methods require raw data from each study. The Hanley and McNeil method can be applied using AUC and SE values included in research reports. The Hanley and McNeil formula provides a z test value to test the difference between two two AUC values from the same sample. Because it is a standard score, any z-score above the absolute value of 1.96 will be considered significant. The formula is as follows:

$$z = \frac{AUC_1 - AUC_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

For the purpose of my calculations, AUC_1 and SE_1 represent the AUC and standard error (SE) for the SRR, and AUC_2 and SE_2 represent the AUC and SE of the total score. The formula accounts for the correlation between the two AUC values taken from the same sample by including *r*, the correlation between the total score AUC and the SRR AUC. This is the same correlation used to calculate the standard error of the AUC difference scores (described above). Using the formula, I was able to find the z-score difference for each AUC pair (total score and SRR for each sample).

CHAPTER III

Results

Each AUC and AUC difference score analysis uses data from 69 samples, with 10,871 participants. The incremental validity analyses used data from 23 samples, with 3,684 participants.

AUC Comparisons

My first question was whether there was any evidence of statistically significant differences in the performance of SPJ measure total scores and SRRs across studies. Table 4 provides the z-score test result for the difference between the SRR AUC and the total score AUC for each of the 69 samples. There was a statistically significant difference in AUC values for only eight (11.6%) of the 69 samples. With 69 samples, I would have expected about four significant differences due to chance. Five of the significant differences indicated superior performance for the SRR (i.e., the SRR AUC was significantly larger than the total score AUC), and three differences indicated superior performance for the total score (i.e., the total score AUC was significantly larger than the SRR AUC).

Table 4

Z-Score of the AUC Difference

Study	SPJ Measure	Total Score AUC(SE)	SRR AUC(SE)	Z-Test
McEachran, 1999[a]	SAVRY	.70(.05)	.89(.04)	4.45*
Lowder et al., 2017[a]	START	.58(.03)	.65(.03)	3.32*
Andres-Pueyo et al., 2008[a]	SARA	.77(.05)	.87(.04)	2.34*
O'Shea and Dickens, 2015, Men[a]	START	.59(.05)	.68(.04)	2.15*
O'Shea and Dickens, 2015, Women[a]	START	.73(.07)	.85(.06)	2.00*
Kropp and Hart, 2000[a]	SARA	.60(.06)	.70(.06)	1.80
O'Shea et al., 2016[a]	START	.64(.04)	.70(.04)	1.59
Langton et al., 2009[a]	HCR-20	.68(.08)	.80(.09)	1.51
Lodewijks et al., 2008, Boys[a]	SAVRY	.76(.07)	.85(.07)	1.39
Chu et al., 2012[a]	ERASOR	.74(.07)	.83(.07)	1.39
Lowder et al., 2017b[a]	START	.48(.07)	.55(.06)	1.28
Fitzgerald et al., 2013, ID Group[a]	HCR-20	.77(.10)	.88(.09	1.24
Lodewijks et al., 2008b[d]	SAVRY	.80(.06)	.86(.05)	1.16
Schaap et al., 2009[a]	HCR-20	.54(.12)	.65(.14)	.90
Pederson et al., 2010[a]	HCR-20	.74(.05)	.78(.05)	.86
Sjöstedt and Långström, 2002[a]	SVR-20	.49(.09)	.56(.09)	.84

(continued)

Study	SPJ Measure	Total Score AUC(SE)	SRR AUC(SE)	Z-Test
Neves et al., 2011[c]	HCR-20	.81(.04)	.83(.04)	.77
Troquete et al., 2015[a]	START	.61(.09)	.66(.08)	.65
Dempster, 1998[a]	SVR-20	.74(.05)	.77(.06)	.58
Viljoen et al., 2009[a]	ERASOR	.60(.09)	.64(.07)	.52
Fitzgerald et al., 2013, Control Group[a]	HCR-20	.58(.08)	.62(.09)	.50
Wilson et al., 2013[a]	HCR-20	.88(.07)	.91(.06)	.49
Skowron, 2004[a]	ERASOR	.71(.05)	.73(.05)	.43
De Vogel et al., 2004[a]	SVR-20	.80(.04)	.83(.09)	.40
De Vogel and de Ruiter, 2006[b]	HCR-20	.85(.04)	.86(.04)	.27
Douglas et al., 2003[a]	HCR-20	.67(.08)	.69(.08)	.27
Desmarais et al., 2012[c]	START	.79(.04)	.80(.04)	.26
Sada et al., 2016[b]	HCR-20	.76(.04)	.76(.04)	.17
Neal et al., 2015[a]	HCR-20	.43(.06)	.44(.07)	.16
Verbrugge et al., 2013[a]	HCR-20	.80(.09)	.81(.07)	.13
Lodewijks et al., 2008, Girls[a]	SAVRY	.84(.09)	.85(.07)	.13
Khanna et al., 2014[a]	SAVRY	.63(.06)	.63(.06)	.05
Arbach-Lucioni et al., 2011[a]	HCR-20	.77(.06)	.77(.06)	<.01
Dolan and Rennie, 2008[a]	SAVRY	.64(.05)	.64(.05)	<.01
Penney et al., 2010, Girls[a]	SAVRY	.72(.08)	.72(.07)	<.01

(continued)

Study	SPJ Measure	Total Score AUC(SE)	SRR AUC(SE)	Z-Test
O'Shea et al., 2015, ID Group[a]	HCR-20	.61(.06)	.61(.06)	-.02
Viljoen et al., 2012[c]	START:AV	.70(.06)	.69(.08)	-.15
Michel et al., 2013, General[a]	HCR-20	.72(.07)	.71(.07)	-.15
Hogan and Olver, 2016[a]	HCR-20^{V3}	.76(.05)	.75(.06)	-.20
O'Shea et al., 2014[a]	HCR-20	.64(.04)	.63(.04)	-.23
Ho et al., 2013[a]	HCR-20	.68(.04)	.67(.04)	-.27
Augimeri et al., 2012	EARL-20B	.64(.04	.63(.04)	-.30
Schmidt et al., 2011[a]	SAVRY	.68(.05)	.66(.05)	-.39
Shepherd et al., 2014[a]	SAVRY	.66(.05)	.64(.05)	-.43
Michel et al., 2013, Forensic[a]	HCR-20	.72(.06)	.69(.08)	-.45
Gray et al., 2011[b]	START	.68(.08)	.63(.10)	-.45
Morton, 2003[a]	ERASOR	.59(.08)	.54(.09)	-.63
Augimeri 2012	EARL-21G	.65(.06)	.59(.10)	-.73
Chu et al., 2011[a]	START	.76(.09)	.69(.10)	-.79
Penney et al., 2010, Boys[a]	SAVRY	.69(.06)	.64(.07)	-.82
Strub et al., 2014[a]	HCR-20^{V3}	.77(.05)	.73(.05)	-.86
Vincent et al., 2012, 16 to 18 Years Old [a]	SAVRY	.68(.05)	.64(.05)	-.86
Hilterman et al., 2014[a]	SAVRY	.75(.05)	.68(.06)	-.90
De Vogel et al., 2004[b]	HCR-20	.82(.04)	.79(.04)	-1.00

(continued)

Study	SPJ Measure	Total Score AUC(SE)	SRR AUC(SE)	Z-Test
Vincent et al., 2012, < 12 Years Old [a]	SAVRY	.71(.13)	.57(.14)	-1.12
Strub et al., 2016[a]	HCR-20	.71(.06)	.64(.07)	-1.15
Rajlic and Gretton, 2010[c]	ERASOR	.71(.05)	.67(.05)	-1.21
O'Shea et al., 2015, Control Group[a]	HCR-20	.62(.06)	.56(.06)	-1.23
Quinn et al., 2013[a]	START	.67(.08)	.58(.08)	-1.28
Storey et al., 2014[a]	B-SAFER	.70(.04)	.65(.04)	-1.35
Pederson et al., 2012[a]	HCR-20	.66(.08)	.56(.08)	-1.35
Viljoen et al., 2008[a]	SAVRY	.58(.05)	.51(.06)	-1.17
Hilterman et al., 2011[a]	HCR-20	.70(.06)	.65(.06)	-1.35
Douglas et al., 2005[a]	HCR-20	.82(.03)	.78(.03)	-1.44
Jovanovic et al., 2009[a]	HCR-20	.85(.04)	.79(.04)	-1.62
Vincent et al., 2012[a], 13 to 15 Years Old	SAVRY	.69(.05)	.61(.05)	-1.73
Belfrage et al., 2012[a]	SARA	.63(.03)	.57(.03)	-1.99*
Worling et al., 2015[a]	ERASOR	.77(.07)	.63(.07)	-2.16*
Gunenc et al., 2015[a]	HCR-20	.68(.02)	.58(.02)	-4.88*

Note. A negative z-score indicates a higher SRR AUC value than total score AUC value, and a positive z-score indicates a higher total score AUC value than an SRR AUC value.

Although the z-test findings suggest that the difference between AUCs is rarely large enough to reach statistical significance at the level of the individual study, there may be still be a pattern of significant differences when I combine effects across studies. A meta-analysis of AUC difference scores from these studies reveled that, on average, the difference between AUC values was about .05, which was large enough to reach statistical significance (see Table 5). Thus, across studies, there was evidence of a difference in the performance of SRRs and total scores. The non-significant Q test and small I^2 value suggest there was not a significant amount of unexplained variability in these effects across samples. Thus, it is not surprising that none of variables were significant moderators of the effect for this AUC difference score (i.e., absolute value).

Table 5

Overall Meta-Analytic Effects

	AUC_w	SE	95% CI	*k*	*Q*	I^2
AUC Difference, Absolute Value	.0482***	.01	.04-.06	69	49.71	.08
AUC Difference	-.0002	.01	.02-.02	69	126.89***	.46
Meta-Analysis						
Total Score AUC	.6981***	.01	.68-.72	69	205.08	.67
SRR AUC	.7005***	.01	.68-.73	69	271.79	.75

*Note. *p < .05. **p < .01. ***p < .001.*

My second question was whether there was any evidence that AUC values tend to be larger for SRRs than total scores. The z-test findings in Table 4 and the meta-analysis of signed AUC difference scores (Table 5) suggest similar performance for SRRs and

total scores across studies. A stem-and-leaf plot of the signed difference scores (Table 6) also shows no clear evidence of superior performance for SRRs or total scores, with 31 (53.62%) differences favoring the SRR, and 32 (46.38%) differences favoring the total score.

Table 6

Stem and Leaf Plot of AUC Difference Values

Stem	Leaf
-0.1	44000
-0.0	877776666555554444332211111
0.0	000000111112223333444566788999
0.1	0011229

Note. Negative values indicate a larger total score of AUC values, positive values indicate a larger SRR AUC values.

Because I subtracted total score AUC values from SRR AUC values, the mean weighted AUC difference score of -.0002 (see Table 5) indicates that the total score AUC values were, on average, slightly larger than the SRR AUC values, although the difference was very small did not approach statistical significance. Although there was a statistically significant amount of variability in these signed difference scores [Q = 126.89, $p < .001$, $I^2 = .46$] only sample size was a statistically significant amount of variability in effects (QM = 4.12, p =.04), with effect size decreasing by .01 with every addition of 100 participants (estimate = -.01, SE = .005).

One additional way to test the difference between SRR and total score effects is to include both effects in the same meta-analysis, and examine score type (total vs. SRR) as a moderator (David Wilson, personal communication, April 5, 2017). While this

technically violates the rule of including multiple effect size values from one sample in a single analysis, the violation is between the two means (the total score AUC mean and the SRR AUC mean) and not for the effects contributing to each mean. The comparison between the two means is biased, but the direction of the bias is to underestimate the significance (the opposite of what happens when you include dependent effect sizes in a single mean). Therefore, the bias would not result in a significant difference between the two means when there really is none (David Wilson, personal communication April 5, 2017). Because each sample contributed two scores (one SRR, one total score) there were 138 effect size values in this analysis. Score type was not a statistically significant moderator of these effect size values (QM = 0.04, p = .85).

SRR and Total Score Meta-Analayses

The mean weighted AUC value for the total score and SRR were both .70 (see Table 5). There was a statistically significant amount of variability for both effects, with I^2 values of .67 for total scores and .75 for SRRs. There were only a few statistically moderator effects for the total score and SRR AUC analyses. Sample size was a significant moderator for both total score AUC (QM = 7.09, p = .008) and SRR AUC (QM = 15.01, p < .001), with the AUC decreasing by .02 as the sample size increased by 100 participants for the total score, and decreasing by .03 for every addition of 100 participants for the SRR. Base rate was a significant moderator for total score AUC (QM = 5.84, p = .02) with the AUC increasing by .0013 for every one percent increase in base rate. The correlation between total score and SRR (r to z transformed) was also a significant moderator for the total score AUC meta-analysis (QM = 4.06, p = .04), with the AUC for total scores increasing as the correlation between total scores and SRRS

increased (estimate = .14). In other words, the total score was more effective in studies in which the SRR and total score were more strongly correlated.

There were statistically significant effects for two categorical moderator variables: country and field study status (see Table 7). For the United States, United Kingdom, and Sweden, the weighted AUC values for both SRR and total score were notably smaller than those for Australia, Canada, and the Netherlands. For the total score AUC meta-analysis, the range of weighted AUCs for country was from .59 (United States) to .78 (Netherlands). For the SRR AUC meta-analysis, the weighted AUC values ranged from .58 (United States) to .79 (Netherlands). For both the SRR and total score AUC meta-analyses studies that were classified as research studies showed higher mean weighted AUC values than those classified as field studies (see Table 7).

Table 7

Moderator Analyses

	Total Score AUC Meta-Analysis					SRR AUC Meta-Analysis				
Moderator	AUC_w	SE	*k*	*Q*	I^2	AUC_w	SE	*k*	*Q*	I^2
Country										
Australia	.72	.05	3	2.30	.24	.71	.06	3	3.91	.50
Canada	.72	.02	20	35.15*	.44	.72	.02	20	58.64***	.66
Netherlands	.78	.02	9	14.58	.38	.79	.03	9	15.20*	.50
Spain	.76	.04	3	0.10	<.01	.78	.05	3	7.31	.71
Sweden	.64	.03	3	8.73	.64	.60	.05	3	2.59	.33
United Kingdom	.65	.02	14	8.73	.02	.66	.02	14	36.68***	.66
United States	.59	.02	8	18.56*	.68	.59	.03	8	13.22	.49
Research Design										
Field	.65	.02	15	33.27**	.59	.66	.01	15	53.84***	.78
Research	.71	.01	53	137.23***	.62	.71	.02	53	214.46***	.73

*Note. *p < .05. **p < .01. ***p < .001.*

Incremental Validity

I examined the incremental validity of the SRR and total score using odds ratios from 23 studies reporting the results of a logistic regression model with both the SRR and total score in the same model. Table 8 summarizes the findings from each study, including the OR values for total scores and SRRs. Because SRRs and total scores are scaled differently, larger OR values for SRRs than total scores do not necessarily indicate larger effects for SRRS than total score. SRRs have only three scoring options (0 = low, 1 = moderate, 2= high). Thus, OR values for SRRs indicate the change in odds that correspond with a change from low to moderate risk, or moderate to high risk. SPJ measures typically have a much larger range (e.g., 0-40 for the HCR-20). Thus, OR values for total scores indicate the change in odds for one point change (i.e., a small change) in the score. OR values for total scores ranged from 0.99 to 1.24, and were large enough to reach statistical significance in six of the 23 samples. OR values for SRRs ranged from 0.91 to 7.29, and were large enough to reach statistical significance in 11 of the 23 samples.

Table 8

Logistic Regression Findings

Study	Measure	SRR (OR)	Total (OR)	*r*
Augimeri 2012	EARL-20B	Yes(2.24)	No (1.04)	.74
Augimeri 2012	EARL-21G	No (1.55)	No (1.00)	.62
Rajlic and Gretton, 2010 [b]	ERASOR	No (1.41)	No (1.05)	.80
Viljoen et al., 2009 [b]	ERASOR	No (2.11)	No (1.00)	.78
Arboch-Lucioni et al., 2011[c]	HCR-20	No	Yes	.68
De Vogel et al., 2004 [a]	HCR-20	Yes (4.07)*	Yes (1.15)*	.82
De Vogel and de Ruiter, 2006 [a]	HCR-20	Yes (7.29)**	No (1.21)	.68
Neal et al., 2015 [b]	HCR-20	No (.91)	No (.99)	
Neves et al., 2011 [c]	HCR-20	Yes (5.77)*	No (1.04)	.89
Sada et al., 2016 [a]	HCR-20	Yes (3.27)**	No (1.11.)	.80
Hogan and Olver, 2016[b]	HCR-20^{V3}	Yes (2.77) *	Yes (1.14) *	.51
Strub et al., 2014[c]	HCR-20^{V3}	Yes (2.27)*	No (1.65)	
Dolan and Rennie, 2008[c]	SAVRY	No (2.11)	No (1.05)	
Khana et al., 2014[b]	SAVRY	No (1.74)	No (1.04)	.78

(continued)

Study	Measure	SRR (OR)	Total (OR)	*r*
Schmidt et al., 2011[b]	SAVRY	No (1.15)	Yes (1.09)***	.87
Viljoen et al., 2008 [b]	SAVRY	No (2.16)	No (1.01)	.74
Chu et al., 2011, risk, strength, SRR [c]	START	No (1.59)	No (1.08), No (1.24)	
Desmarais et al., 2012, vulnerability total score [b]	START	Yes (3.24)**	Yes (1.08)*	.69
Gray et al., 2011[a]	START	No (1.23)	No (1.12)	.42
O'Shea and Dickens, 2015, Men, Strength scores [c]	START	Yes (2.69)**	Yes (1.11) **	
O'Shea and Dickens, 2015, Women, vulnerability scores [c]	START	Yes (5.37) **	No (1.00)	
O'Shea et al., 2016, vulnerability score, strength score, SRE [c]	START	Yes (2.44) ***	No (1.00), No (1.05)	.39
Troquete et al., 2015, vulnerability and strength scores [b]	START	No (2.65)	No (1.08), No (1.08)	.37
Viljoen et al., 2012, vulnerability and strength scores [b]	START:AV	No (.948)	No (1.01), No (1.00)	.56
Dempster, 1998 [c]	SVR-20	Yes* (2.44)	No (1.07)	

*Note. [a]Study author responded to data request with raw data. [b]Study author responded to data request using the data request sheet or data output [c]Regression data pulled from original publication. ***p <.001, **p<.01, *p<.05.*

I ran separate meta-analyses for the SRR and total score OR values. Both OR values were large enough to reach statistical significance. The mean weighted OR for the total score analysis was 1.09 (95% CI 1.04 to 1.15), indicating a 1.09 increase in the odds of the outcome occurring for every one point increase in SPJ measure total score. Cochran's Q test for heterogeneity for the total score meta-analysis was statistically significant (Q(df=22)=210.74, p <.001), and the amount of unexplained heterogeneity was large (I^2 = .84) which indicates the differences between studies cannot be attributed to sampling error or chance alone.

The mean weighted OR for the SRR analysis was 1.96 (95% CI 1.54 to 2.51). Cochran's Q test for heterogeneity for the SRR meta-analysis was statistically significant (Q(df=22)=65.79, p < .001), and the I^2 value of .54 indicated that there was a moderate amount of unaccounted for variance across samples included in the analysis.

Due to the large amount of unexplained heterogeneity in each OR meta-analysis, I ran moderator analyses using R for all moderators defined previously. Only one moderator provided any significant explanation of study heterogeneity. For the SRR OR meta-analysis, only target demographic (QM(df = 1) = 4.12, p = .04 was a significant moderator. No moderators were significant for the total score OR meta-analysis. See table Table 9 for an overview of the significant moderator findings.

Table 9

OR Meta-Analysis Significant Moderators

	Total Score OR Meta-Analysis					SRR OR Meta-Anallysis				
Moderators	OR_w	SE	*k*	*Q*	I^2	OR_w	SE	*k*	*Q*	I^2
Target Demographic										
Adult	1.09	1.03	14	2.74	.64	1.61*	1.26	14	4.12*	.76
Adolescent	1.03	1.05	9	.90	.10	1.45*	1.19	9	4.16*	.76

*Note. *p < .05. ** p < .01. ***p < .001*

The moderator test of whether or not the logistic regression data was included in the original publication was not significant for either total score or SRR meta-analysis. In other words, there was no evidence of a reporting bias, with only especially large incremental validity findings being reported in published studies.

Representativeness of the Incremental Validity Effect

The finding of a statistically significant effect for the incremental validity of SRRs over total scores seems to be somewhat at odds with the findings from the AUC meta analyses. Those analyses suggested nearly identical performance for SRRs and total scores. Because fewer studies contributed to the incremental validity meta-analysis, it could be that these studies are in some way not representative of the SPJ literature as a whole. To examine this possibility, we re-ran each AUC meta-analysis (difference score, SRR, Total), and examined incremental validity study status (23 samples = yes, 46 samples = no) as a possible moderator effect. A finding that SRRs were especially strong predictors in the incremental validity studies, or that total scores were especially weak predictors in these studies, would suggest that the incremental effects may not generalize to all SPJ samples.

The moderator analysis was not significant for the total score AUC meta-analysis, the SRR AUC meta-analysis, or the signed AUC difference meta-analysis. The moderator was significant for the absolute value AUC difference meta-analysis ($QM(df = 1) = 8.20$, $p = .0042$. Samples that I used in the OR meta-analysis had a lower AUC absolute difference estimate than those samples that I did not use in the OR meta-analysis. See Table 10 for results.

Table 10

OR Meta-Analysis Status as a Moderator

	AUC Absolute Difference Value Meta-Analysis				
Moderator	AUC_{Dif}	SE	*k*	*Q*	I^2
Sample Included in OR meta-analysis					
Yes	.03**	.01	23	8.20**	.88
No	.06***	.01	46	75.37***	.99

*Note. *p < .05. **p < .01. ***p < .001.*

CHAPTER IV

Discussion

Meta-Analytic Findings

The current study was undertaken to provide a better understanding for practitioners in the field on the utility of clinical judgment in the use of SPJ instruments. In order to complete this study I acquired additional data from study authors. In total, I utilized 69 samples and acquired data that was not originally published for 17 samples. I meta-analyzed 69 pairs of AUC values and 23 pairs (17 unique) of OR values from logistic regressions examining the total score and SRR in the same model. Therefore, this study adds data not previously published in the field from several published manuscripts. Because so few authors report logistic regression analyses, this significantly adds to the knowledge in the field about incremental effects of SRR over total scores for SPJ measures.

Overall, findings demonstrate that the mean weighted total score AUC (.70) and SRR AUC (.70) were similar, showing moderate predictive abilities. A direct comparison of the mean weighted AUC values did not result in a significant difference. This finding is not wholly unexpected given previous research (Chevalier, 2015) which found similar effects for total score and SRR, but was unable to directly compare the two values. An unappreciable difference between the two weighted AUC values is also expected given the lack of significant differences at the study level. Less than 15% of samples showed a significant difference between the total score AUC and the SRR AUC as tested using Hanley and McNeil's (1983) z-score formula. The weighted OR value for

the SRR (1.96) was higher than the total score (1.09) indicating an incremental effect of the SRR even when the total score was controlled for.

Moderator Effects

Only a small number of moderators explained a significant amount of variability in AUC values across studies. Outcome, sex, age, follow-up time, SPJ measure, author allegiance, whether or not the author originally reported logistic regression data in their study, and whether or not the sample was included the OR meta-analysis were not significant moderators of total score AUC and SRR AUC meta-analyses .

A previous meta-analysis that examined actuarial measures found that instruments deigned to assess adults had better predictive validity than those designed to assess adolescents (Singh et al., 2011). The present findings suggest that the total score AUC, SRR AUC, and the difference between the two values for each sample do not vary systematically based on the average age of the sample, or whether the SPJ measure was intended for use with adults or adolescents. The total score OR meta-analysis, however, did find a moderating effect for target demographic on the OR value, with a larger OR value for SPJ measures meant for adults. The findings are consistent with a previous meta-analysis that examined the predictive validity of risk assessment instruments (Fazel, Singh, Doll, & Gann, 2012), which found that sex, age, type of instrument, and length of follow-up, were not significant moderators.

Unlike the aforementioned meta-analysis (Fazel et al., 2012) which found that sample size was not a significant moderator of predictive validity, my analyses revealed that the sample size was a significant moderator for both total score AUC and SRR AUC meta-analyses, with the AUC value decreasing by between .02 to .03 with every one unit

increase in sample size (one unit = 100 participants). This finding shows that predictive accuracy tended to decrease across studies as more participants were added.

Additionally, AUC values tended to increase across studies as the base rate of recidivism increased. Results indicate that SPJ measures did a better job at predicting recidivism if the rate of recidivism was higher. This finding makes sense given the increased likelihood a measure has of correctly identifying an individual as a recidivist if the likelihood of recidivating is high. Study design (i.e., research or field) was a significant moderator for the total score meta-analysis, with a higher weighted AUC value for research studies (.71) as opposed to field studies (.65). Possible explanations for this finding are unclear, but could be attributable to the individuals completing the SPJ measure. It could be that those individuals performing the risk assessments in field studies tend to be members of organizations such as police officers or nurses, whereas those completing the measure in research studies tend to be masters or PhD level clinicians who have undergone specific training on implanting the SPJ scheme. Also interesting is the pattern of a significant moderating effect for the total score meta-analysis but not the SRR meta-analysis. Given how similar the weighted AUC values are for both analyses, a similar pattern of moderation effects are expected to a certain extent.

The country the study was conducted in also explained a significant amount of study variability across both total score and SRR meta-analyses. Studies conducted in the Netherlands showed the most robust predictive effects (.78 for total score, .79 for SRR). Studies conducted in the United States evidenced the lowest AUC values for both total score (.59) and SRR (.59). This finding could be an indication of the effects of implementing SPJ instruments cross-culturally. Only one measure studied had a version

that was developed for a specific population: the Historische, Klinische, Toekmostige-30 (HKT-30), a version of the HCR-20 developed for use with a Dutch population. But perhaps these findings imply more attention should be paid to the application of these measures across cultures.

Interestingly, only one moderator (sample size) explained a significant amount of variability in the difference between AUC values across samples. The analysis package applied a random-effects model and the test for heterogeneity was significant, implying there is more variation than would be expected by chance. The source of the heterogeneity is unclear, and more moderator variables should be considered in future studies in order to provide further explanation of the variation in studies. Unlike the signed AUC difference value meta-analysis, the absolute difference value meta-analysis did not indicate a significant *Q* value, so it is not surprising that only one significant moderator, whether or not the sample was included in the logistic regression analysis, was found.

Implications for Research

This study highlights important implications for data reporting in empirical research. In order to better understand the interaction of clinical judgment and total scores, study authors should consider reporting not only the respective AUC values for each, but the direct comparison of the values. Authors should consider using the DeLong and colleagues (1988) test, available in software programs like STATA or MedCalc, to perform direct comparisons using their raw data. This way, the research consumer can better understand how the SRR compares to the ability of the total score to predict future behavior, and how much weight to give the SRR should be given in final risk estimates.

Without the raw data or, at a minimum, the *r* for the correlation between the two AUC values, research consumers cannot accurately compare the SRR and total score from the same sample and understand how the two methods perform with specific populations.

In addition to a direct comparison of the two AUC values, authors should also perform incremental validity analyses of the SRR over the total score. This way, practitioners can identify how much more predictive accuracy the SRR adds to the total score in SPJ risk assessment. In order for research consumers to completely understand the incremental validity findings, authors should err on the side of over-inclusion and provide the B, SE, expB, corresponding 95% CI values, and exact *p* values. By reporting those values, future researchers can accurately include those analyses in meta-analytic findings. Additionally, because of the low number of authors who originally reported these specific analyses in their published work ($k = 6$), if authors make reporting these values standard practice it will add significantly to the knowledge of SRR incremental effects within the field. Overall, researchers should endeavor to include as much data in their reports as possible to allow both fellow researchers and practitioners to accurately interpret and utilize the findings.

Implications for Practice

The main questions of this study were geared toward assisting practitioners in the field with determining the predictive utility of their clinical judgment when administering SPJ measures. Overall, the SRR and total score showed similar moderate predictive effects, with no appreciable difference between the two values. When examining the differences at the sample level, the difference in only a small percentage of samples reached significance. This seems to provide support for the idea that clinical judgment is

not harming, but also not besting, the predictive ability of using SPJ measures in an actuarial fashion (using a total score).

However, when examining the results from the logistic regression analyses there seems to be support for the idea that the SRR adds to the predictive effects already provided by the total score, and is therefore valuable to include when using SPJ measures. The finding remains that there is some aspect of the SRR that, even after controlling for total scores, accounts for variability in outcomes. In light of the findings from the AUC value meta-analyses, this finding sheds light on the potential utility of the SRR.

So, the question remains: Should practitioners simply use the mechanical total score, or should they also include their structured clinical judgment through the SRR? The findings suggest that there is utility in adding the SRR to mechanical scores on SPJ measures. However, the way these two scores interact, or what the SRR is capturing that the total score is not, is unclear. Practitioners should feel confident, however, that their SPJ SRRs are at least as predictive if not adding significantly to the predictive validity of the summated total scores they calculate on SPJ measures.

Limitations

One of the largest limitations for the present study was the inability to accurately account for each correlation between the total score AUC and the SRR AUC for each sample. Using the correct correlation for each sample would have been a more precise way to calculate the z-score difference values. For example, the z-score difference for Kropp and Hart's. (2010) sample using the representative *r* value of .57 was 1.80. By adjusting the *r* value up or down, the z-score difference value changes. Increasing the *r*

value to .7, indicating a higher correlation between the total score AUC and the SRR AUC, would have resulted in a z-score difference of 2.15 (i.e., a significant difference). By using the more conservative estimate of the correlation between the AUC values, the analysis was a more cautious approach. If I had access to more raw data, or had asked study authors to perform necessary calculations to find the correlation of the AUC values, more AUC differences may have achieved statistical significance. I perceived the burden of calculating the correlation between the two AUC values would significantly negatively impact response rate, but perhaps placing increased emphases on acquiring raw data sets would have allowed me to have a larger pool to calculate the representative *r* sample from.

Another limitation of this study is the lack of explanatory moderator variables for the various meta-analyses. Only one meta-analysis (AUC absolute difference value) indicated there was an insignificant amount of heterogeneity that could not be attributable to chance. Because few of the moderators analyzed were significantly related to systematic change in AUC scores across studies, coding and analyzing additional moderators could provide some explanation on factors that might help explain differences in findings.

Conclusion

Overall, the total score and SRR produced similar predictive effects. Practitioners in the field can confidently rely on either the total score or the SRR when assessing risk and expect to achieve moderate predictive ability. However, there is emerging support for the idea that the SRR adds to the predictive validity of the total score, and therefor yields some clinical utility.

REFERENCES

Andrés-Pueyo, A., López, S., and Álvarez, E. (2008). Aessessment of the risk of intimate partner violence and the SARA. *Papeles del Psicólogo, 29,* 107-122.

Arbach-Lucioni, K., Andrés-Pueyo, A., Pomarol-Clotet, E., & Gomar-Soñes, J. (2011). Predicting violence in psychiatric inpatients: A prospective study with the HCR-20 violence risk assessment scheme. *The Journal of Forensic Psychiatry and Psychology, 22,* 203-222. doi:10.1080/1478-9949

Augimeri, L. K., Koegl, C. K., Webster, C. D., & Leveene, K. (2001). *Early assessment risk list for boys. EARL-20B. Version 2.* Toranto, ON: Child Development Institute.

Augimeri, L. K., Walsh, M., Woods, S., & Jian, D. (2012). Risk assessment and clinical risk management for young antisocial children: The forgotten group. *Universitas Psychologica, 11,* 1147-1156.

Belfrage, H., Strand, S., Storey, J. E., Gibas, A. L., Kropp, P. R., & Hart, S. D. (2012). Assessment and management of risk for intimate partner violence by police officers using the Spousal Assault Risk Assessment Guide. *Law and Human Behavior, 36,* 60-67. doi: 10.1037/h0093948

Bengston, S., & Långström, N. (2006). Unguided clinical and actuarial assessment of re-offending risk: A direct comparison with sex offenders in Denmark. *Sex Abuse, 19,* 135-153. doi:10.1007/s11194-007-9044-5

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20. Professional guidelines for assessing risk of sexual violence.* Vancouver, British Columbia: Institute Against Family Violence.

Borum, R., Bartel, P., & Forth, A. (2003). *Manual for the Structured Assessment for Violence Risk in Youth (SAVRY).* Odessa, FL: Psychological Assessment Resources.

Borum, R., Lodewijks, H., Bartel, P. A., & Forth, A. (2010). Structured Assessment of Violence Risk in Youth (SAVRY). In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 63-80). New York, NY: Routledge.

Braithwaite, E., Charette, Y., Crocker, A. G., & Reyes, A. (2010). The predictive validity of clinical ratings of the Short-Term Assessment of Risk and Treatability (START). *International Journal of Forensic Mental Health, 9,* 271-281. doi:10.1080/14999013.2010.534378

Chevalier, C. S., Boccaccini, M. T. (2015, March). *A Comparison of Structured Professional Judgment Instrument Scores and Summary Risk Ratings.* A poster presented at the annual conference of the American Psychology-Law Society, San Diego, CA.

Chu, C. M., Thomas, S. D., Ogloff, J. R. P., & Daffern, M. (2011). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) in a secure forensic hospital: Risk factors and strengths. *International Journal of Forensic Mental Health, 10,* 337-345. doi:10.108014999013.2011.629715

Chu, M. C., Ng, K., Fong, J., & Teoh, J. (2012). Assessing youth who sexually offend: The predictive validity of the ERASOR, JSOAP-II, and YLS/CMI in a non-Western context. *Sexual Abuse: A Journal of Research and Treatment, 24,* 153-174. doi:10.1177/107906321140250

Chu, C. M., Daffern, M., & Ogloff, R. P. (2013). Predicting aggression in acute inpatient psychiatric setting using BVC, DASA, and HCR-20 Clinical scale. *The Journal of Forensic Psychiatry & Psychology, 24,* 269-285. doi:10.1080/147899492013.773456

DeLong, E. R., DeLong, D. M., & Clark-Pearson, D. L. Comparing areas under two or more correlated receiver operating characteristics curves: A nonparametric approach. *Biometrics, 44,* 837-845.

Dempster, J. R. (1998). *Prediction of sexually violent recidivism: A comparison of risk assessment instruments* (Unpublished master's thesis). Simon Fraser University.

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24,* 685-700. doi:10.1037/a0026668

de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical Psychology and Psychotherapy, 12,* 226-240. doi:10.1002/cpp452

de Vogel, V., & de Ruiter, C. (2006). Structured professional judgment of violence risk in forensic clinical practice: A prospective study into the predictive validity of the Dutch HCR-20. *Psychology, Crime & Law, 12,* 321-336. doi:10.1080/10683160600569029

de Vogel, V., de Ruiter, C., Bouman, Y., & de Vries Robbé, M. (2007). *Handleiding bij de SAPROF: Structured Assessment of Protective Factors for Violence Risk,*

Version 1. [SAPROF Manual. Structured Assessment of Protective Factors for Violence Risk. Version 1]. Utrecht, The Netherlands: Forum Educatief.

de Vogel, V., de Ruiter, C., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health, 3,* 149-165. doi:10.1080/14999013.2004.10471204

de Vogel, V., de Ruiter, C., van Beek, D., & Mead, G. (2004b). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior, 28,* 235-251.

Dickens, G. L., & Oshea, L. E. (2015). How short should short term risk assessment be? Determining the optimum interval for the START reassessment in a secure mental health service. *Journal of Psychiatric and Mental Health Nursing, 22,* 397-406.

Dolan, M. C., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorder. *Psychological Assessment, 20,* 35-46. doi:10.1037/1040-3590.20.1.35

Douglas, K. S., Shaffer, C., Blanchard, A., Guy, L. S., Reeves, K. A., & Weir, J. M. (2014). *HCR-20 Violence Risk Assessment Scheme: Overview and annotated bibliography.* Available from http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1362&context=psych_cmhsr {Accessed 20 March, 2017}.

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

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

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgments for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing, 9,* 649-657.

Falzer, P. R. (2013). Valuing structured professional judgment: Predictive validity, decision-making, and the clinical-actuarial conflict. *Behavioral Sciences and the Law, 31,* 40-54. doi:10.1002/bsl.2043

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 24827 people: Systematic review and meta-analysis. *British Medical Journal, 345,* 1-12.

Fitzgerald, S., Gray, N. S., Alexander, R. T., Bagshaw, R., Chesterman, P., Huckle, P., Jones, S. K., Taylor, J., Williams, T., & Snowden, R. J. (2013). Predicting institutional violence in offenders with intellectual disabilities: The predictive efficacy of the VRAG and the HCR-20. *Journal of Applied Research in Intellectual Disabilities, 26,* 384-393. doi:10.1111/jar.12032

Gray, N. S., Benson, R. Craig, R., Davies, H., Fitzgerald, S., Huckle, P., Maggs, R., Taylor, J., Trueman, M., Williams, T., & Snowden, R. J. (2011). The Short-Term

Assessment of Risk and Treatability (START): A prospective study of inpatient behavior. *International Journal of Forensic Mental Health, 10,* 305-313. doi:10.1080/14999013.2011.631692

Green, D., Schneider, M., Griswold, H., Belfi, B., Herrera, M., DeBlasi, A. (2016). A comparison of the HCR-20^{V3} among male and female insanity acquittees: A retrospective file study. *The International Journal of Forensic Mental Health, 15,* 48-64.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19-30. doi:10.1037//1040-3590.12.1.9

Guy, L. (2008). *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey.* Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1-21.

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 for assessing risk of sexual violence.* Burnaby, Canada: Mental Health, Law, & Policy Institute, Simon Fraser University.

Heilbrun, K., Yasuhara, K., & Shah, S. (2010). Violence Risk Assessment Tools: Overview and Critical Analysis. In Otto, R. K. & Douglas, K. S. (Eds.), *Handbook of Violence Risk Assessment* (1-17). New York, NY: Routledge.

Helmus, L., & Bourgon, G. (2011). Taking stock of 15 years of research on the Spousal Assault Risk Assessment Guide (SARA): A critical review. *International Journal of Forensic Mental Health, 10,* 64-75. doi:10.1080/14999013.2010.551709

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327,* 557-560.

Hilterman, E. B., Philipse, M. W. G., & de Graaf, N. D. (2011). Assessment of offending during leave: Development of the leave risk assessment in a sample of Dutch forensic psychiatric patients. *International Journal of Forensic Mental Health, 10,* 233-243. doi:10.108014999013.2011.598601

Hilterman, L. B., Nicholls, T. L., & van Nieuwenhuizen, C. (2014). Predictive validity of risk assessments in juvenile offenders: Comparing the SAVRY, PCL:YV, and YLS/CMI with unstructured clinical assessments. *Assessment, 21,* 324-339. doi:10.1177/10731911113498113

Ho., R. M. Y., Lau, J. S. F., Cheung, H. H. K., Lai, T. T. S., Tam, V. F. L., Chan, W. L., Yeun, K. K., & Yan, C. K. (2013). Evaluation of a model of violence risk assessment (HCR-20) among adult patients discharged from a gazette psychiatric hospital in Hong Kong. *The Journal of Forensic Psychiatry & Psychology, 24,* 479-495. doi:10.1080/14789949.2013.809467

Janus, E. S., & Prentky, R. A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, adminssibility, and accountability. *American Criminal Law Review, 40,* 1443-1499.

Khanna, D., Shaw, J., Dolan, M., Lennox, C. (2014). Does diagnosis affect predictive accuracy of risk assessment tools for juvenile offenders: Conduct disorder and Attention Deficit Hyperactivity Disorder. *Journal of Adolescence*, 37, 1171-1179.

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

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

Kropp, P., R., Hart, S., D., & Blefrage, H. (2005). *Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER); User manual.* Vancouver, British Columbia, Canada: ProActive ReSolutions.

Langton, C. M. (2003). *Personality traits and dynamic variables associated with types of aggression in high security forensic psychiatric inpatients* (Unpublished doctoral dissertation). University of Toronto.

Leveene, K. S., Augimeri, L. K., Pepler, D. J., Walsh, M. M., Koegl, C. J., & Webster, C. D. (2001). *Early Assessment Risk List for Girls: EARL-21G. Version 1.* [Consultation edition]. Toronto, ON: Child Development Institute.

Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2008). Gender differences in violent outcomes and risk assessment in adolescent offenders after residential treatment. *International Journal of Forensic Mental Health, 7,* 133-146. doi:10.1080/14999013.2008.9914410

Lodewijks, H. P. B., Doreleijers, T. A. H., de Ruiter, C., & Borum, R. (2008b). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. *International Journal of Law and Psychiatry, 31,* 263-271. doi:10.1016/j.jilp2008.04.009

Lowder, E. M., Desmarais, S. L., Rade, C. B., Johnson, K. L., & Can Dorn, R. A. (2017) Reliability and validity of START and LSI-R assessments in mental health jail diversion clients. *Assessment,* 1-15.

Lowder, E. M., et al. (2017b). Models of protection against recidivism in justice-involved adults with mental illness. *Criminal Justice and Behavior*. Manuscript accepted for publication.

McEachran, A. (1995). *The predictive validity of the PCL:YV and the SAVRY in a population of adolescent offenders* (Unpublished master's thesis). University of Western Ontario.

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

Messing, J. T., & Thaller, J. (2013). The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence, 29,* 1537-1558.

Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Müller-Isberner, R., Tiihonen, J., Repo-Tiihonen, E., Tuninger, E., & Hodgins, H. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and

dynamic risk factors. *International Journal of Forensic Mental Health, 12,* 1-13. doi:10.1080/149990132012.760182

Morton, K. E. (2003). *Psychometric properties of four risk assessment measures with male adolescent sexual offenders* (Unpublished master's thesis). Carleton University.

Neal, T. M. S., Miller, S. L., & Shealy, R. C. A field study of a comprehensive violence risk assessment battery. *Criminal Justice and Behavior, 42,* 952-968.

Neves, A. C., Gonçalves, R. A., & Palma-Oliveira, J. M. (2011). Assessing risk for violence and general recidivism: A study of the HCR-20 and the PCL-R with a non-clinical sample of Portuguese offenders. *International Journal of Forensic Mental Health, 10,* 137-149. doi:10.1080/14999013.2011.577290

Nicholls, T. L., Viljoen, J. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (2010). *Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV) (abbreviated manual).* Coquitlam, Canada: BC Mental Health and Addiction Services.

Ogloff, J., & Daffern, M. (2002). *Dynamic appraisal of situational aggression: Inpatient version.* Melbourne: Monash University and Forensicare.

O'Shea, L. E., & Dickens, G. L. (2014). Short-Term Assessment of Risk and Treatability (START): Systematic Review and Meta-Analysis. *Psychological Assessment, 26,* 990-1002. doi:10.1037/a0036794

O'Shea, L. E., & Dickens, G. L. (2015). Predictive validity of the Short-Term Assessment of Risk and Treatability (START) for aggression and self-harm in a

secure mental health service: Gender differences. *International Journal of Forensic Mental Health, 14,* 132-146.

O'Shea, L. E., Mitchell, A. E., Picchioni, M. P., & Dickens, G. L. (2013). Moderators of the predictive efficacy of the Historical, Clinical and Risk Management-20 for aggression in psychiatric facilities: Systematic review and meta-analysis. *Aggression and Violent Behavior, 18,* 255-270. doi:10.1016/j.avb2012.11.016

O'Shea, L. E., Picchioni, M. P., & Dickens, G. L. (2016). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. *Assessment, 23*, 150-162.

O'Shea, L., Picchioni, M. M., Mason, F. L., & Sugarman, P. A. (2014). Differential predictive validity of HCR-20 for inpatient aggression. *Psychiatry Research, 220,* 669-678.

O'Shea, L. E., Picchioni, M. M., McCarthy, J., Mason, F. L., & Dickens, G. L. (2015). Predictive validity of the HCR-20 for inpatient aggression: The effect of intellectual disability on accuracy. *Journal of Intellectual Disability Research, 59,* 983-1097.

Pederson, L., Rasmussen, K., & Elsass, P. (2010). Risk assessment: The value of Structured Professional Judgments. *International Journal of Forensic Mental Health, 9,* 74 – 81. doi:10.1080/14999013.2010.499556

Pederson, L., Rasmussen, K., & Elsass, P. (2012). HCR-20 violence risk assessments as a guide for treating and managing violence risk in a forensic psychiatric setting. *Psychology, Crime & Law, 18,* 733-743. doi:10.1080/1068316x.2010.548814

Penney, S. R., Lee, Z., & Moretti, M. M. (2010). Gender differences in risk factors for violence: An examination of the predictive validity of the Structured Assessment of Violence Risk in Youth. *Aggressive Behavior, 36,* 390-404. doi:10.1002/ab.20352

Quinn, R., Miles, H., & Kinane, C. (2013). The validity of the Short-Term Assessment of Risk and Treatability (START) in a UK medium secure forensic mental health service. *International Journal of Forensic Mental Health, 12,* 215-224.

Rajlic, G., & Gretton, H. M. (2010). Asn examination of two sexual recidivism risk measures in adolescent offenders: The moderating effect of offender type. *Criminal Justice and Behavior, 37,* 1066-1085. doi:10.1177/0093854810376354

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior, 29,* 615-620. doi:10.1007/s10979-005-6832-7

Robin, X., Turck, N., Hainard, A., Tiberti, N., Frédérique, L., Sanches, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC *Bioinformatics,* 12, 77. doi:10.1186/1471-2105-12-77

Sada, A., Robles-García, R., Martínez-López, N., Hernández-Ramírez, R., Rovilla-Zarate, CA., López-Minguía, F., Suárez-Alvarez, E., Ayala, X., Fresán, A. (2016). Assessing the reliability, predictive and construct validity of historical, clinical and risk management-20 (HCR-20) in Mexican psychiatric inpatients. *Nordic Journal of Psychiatry, 6,* 456-461.

Schaap, G., Lammers, S., & de Vogel, V. (2009). Risk assessment in female forensic psychiatric patients: A quasi-prospective study into the validity of the HCR-20

and PCL-R. *The Journal of Forensic Psychiatry and Psychology, 20,* 354-365. doi:10.1080/14789940802542873

Schmidt, F., Campbell, M. A., & Houlding, C. (2011). Comparative analyses of the YLS/CMI, SAVRY, and PCL:YV in adolescent offenders: A 10-year follow-up into adulthood. *Youth Violence and Juvenile Justice, 9,* 23-42. doi:10.1177/1541204010371793

Shepherd, S. M., Luebbers, S., Ferguson, M., Ogloff, J. R. P., & Dolan, M. (2014). The utility of the SAVRY across ethnicity in Australian young offenders. *Psychology, Public Policy, and Law, 20,* 31-45. doi:10.1037a0033972

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K. … Ottot, 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,* 193-206.

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,* 499-513. doi:10.1016/j.cpr.2010.11.009

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime & Law, 8,* 25-40. doi:10.1080/10683160208401807

Skowron, C. (2004). *Differentiation and predictive factors in adolescent sexual offending* (Unpublished master's thesis). Carleton University.

Storey, J. E., Kropp, R., Hart, S. D., Belfrage, H., & Strand, S. (2013). Assessment and management of risk for intimate partner violence by police officers using the brief spousal assault form for the evaluation of risk. *Criminal Justice and Behavior, 41,*

Strub, D. S., Douglas, K. S., & Nicholls, T. L. (2014). The validity of Version 3 of the HCR-20 violence risk assessment scheme amongst offenders and civil psychiatric patients. *International Journal of Forensic Mental Health, 13,* 148-159.

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,* 119-133. doi:10.1080/14999013.2012.690020

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19,* 227-229.

Tully, R. J., Chou, S., & Browne, K. D. (2013). A systematic review on the effectiveness of sex offender risk assessment rools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review, 33,* 287-316.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software, 36,* 1-48.

Viljoen, J. L., Beneteau, J. L., Gulbransen, E., Brodersen, E., Desmarais, S. L., Nicholls, T. L., & Cruise, K. R. (2012). Assessment of multiple risk outcomes, strengths, and change with the START:AV: A short-term prospective study with adolescent offenders. *International Journal of Forensic Mental Health, 11,* doi:10.1080/14999013.2012.737407

Viljoen, J. L., Elovitch, N., Scalora, M. J., & Ullman, D. (2009). Assessment of re-offense risk in adolescents who have committed sexual offenses. *Criminal Justice and Behavior, 36,* 981-1000. doi:10.1177/0093854809340991

Viljoen, J. L., Mordell, S., & Beneteau, J. L. (2012). Prediction of adolescent sexual offending: A meta-analysis of the J-SOAP-II, ERASOR, J-SORRAT-II, and Static-99. *Law and Human Behavior, 36,* 423-438. doi:10.1037/h0093938

Viljoen, J. L., Scalora, M., Cuadra, L., Bader, S., Chávez, V., Ullman, D., & Lawrence, L. (2008). Assessing risk for violence in adolescents who have sexually offended. A comparison of the J-SOAP-II, J-SORRAT-II, and SAVRY. *Criminal Justice and Behavior, 35,* 5-23. doi:10.1177/0093854807307521

Vincent, G. M., Guy, L. S., Gershenson, B. G., & McCabe, P. (2012b). Does risk assessment make a difference? Results of implementing the SAVRY in juvenile probation. *Behavioral Sciences and the Law, 30,* 384 – 405. doi:10.1002/bsl.2014

Vincent, G. M., Perrault, R. T., Guy, L. S., & Gershenson, B. G. (2012). Developmental issues in risk assessment: Implications for juvenile justice. *Victims and Offenders, 7,* 364-384. doi:10.1080/15564886.2012.713900

Vitacco, M. J., Tabernik. H. E., Zavodny, D., Bailey, K., Waggoner, C. (2016). Projecting risk: The importance of the HCR-20 risk management scale in predicting outcomes with forensic patients. *Behavioral Sciences, 34,* 306-320.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20. Assessing the risk of violence. Version 2.* Burnaby, BC, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Middleton, C. (2004). *Manual for the Short Term Assessment of Risk and Treatability (START). Version 1.0, consultation edition.* Hamilton, St. Joseph's Healthcare Hamilton Centre for Mountain Health Services and Port Coquitlam: Forensic Psychiatric Services Commission.

Wilson, D. B. (2006). Meta-analysis macros for SAS, SPSS, and Stat. Retrieved June 3, 2017 from http://mason.gmu.edu/~dwilsonb/ma.html.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., & Brink, J. (2010). The role of client strengths in assessments of violence risk using the Short-term Assessment of Risk and Treatability (START). *International Journal of Forensic Metnal Health, 9,* 282-293.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior, 37,* 377-388. doi:10.1037/lhb0000025

Worling, J. R., & Curwen, T. (2001). *Estimate of Risk of Adolescent Sexual Offense Recidivism* (Version 2.0). Toronto, Ontario, Canada: Ontario Ministry of Community and Social Services.

Worling, J. R., & Langton, C. M. (2015). A prospective investigation of factors that predict desistance from recidivism for adolescents who have sexually offended. *Sexual Abuse: A Journal of Research and Treatment, 27,* 127-142. doi:10.1177/1079063214549260

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). Statistical methods in diagnostic medicine. *Biometrics, 59,* 203 – 204. doi:10.1111/1541-0420.00266

APPENDIX A

Study:

SPJ Measure:

Sample:

Outcome:

My requests: I need some information about 1) AUC analyses, 2) correlations between total scores and summary risk ratings (SRRs), and 3) incremental validity (logistic regression). The three tables below list information I found in your study and identify information that I still need. It would be very helpful of you could either a) fill in the missing information, or b) send analysis output that would allow me to fill in the missing information

1) AUC values and standard error.

	AUC		Standard Error
Total Score			
Summary Risk Rating			

2) Correlation between SRR and total scores, for the overall sample and separately for recidivists and non-recidivists.

	r				
	Recidivists		Non Recidivists		Overall
Total Score and Summary Risk Rating					

3) Results from three logistic regression models. First using the total score to predict recidivism. Second using the SRR to predict recidivism. Third using both the total score and SRR to predict recidivism.

	B	Standard Error	Wald Statistic	P Value	Odds Ratio	Odds Ratio CI	Model Chi Square	Chi Square Change
Model 1								
Total score								
Model 2								
SRR								
Model 3								
Total Score								
SRR								

APPENDIX B

Dr. (insert name of first author),

My name is Caroline Chevalier and I am a doctoral candidate in clinical psychology at Sam Houston State University. For my doctoral dissertation, I am completing a systematic review of structured professional judgment (SPJ) studies under the guidance of my dissertation chair, Dr. Marcus Boccaccini. My primary goal is to compare the predictive validity of summary risk ratings and SPJ summated (total) scores. I have located both published and unpublished studies that report an AUC value for both the total score and summary risk rating for the same sample. Because you are listed as the corresponding author on one of these studies, I am reaching out to you in order to gather more information about your study. To provide the type of systematic review that would benefit the field, I need some information about AUC values, correlations between SPJ scores, and regression results. I have attached a document to this e-mail listing the information that I am using in my review. I have attempted to fill in all of the information that you provided in your original report and highlighted the information that I still need.

I am hoping that you will be able to add the missing information and send the form (or statistical output) back to me. I recognize that completing this form requires your time and greatly appreciate your willingness to consider my request. If you have any questions or would like further information about my study, please do not hesitate to ask. Thank you in advance for your time. Below you will find the citation for the article I am referencing.

Study citation

Caroline S. Chevalier, M.A.
Doctoral Candidate, Clinical Psychology
Sam Houston State University

APPENDIX C

Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER)

Dynamic Risk Assessment for Offender Re-entry (DRAOR)

Dynamic Appraisal of Situational Aggression (DASA)

DUNDRUM Risk Assessment Tool Kit

Early Assessment Risk List for Boys (EARL-20B)

Early Assessment Risk List for Girls (EARL-21G

Employee Risk Assessment-20 (ERA-20)

Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR)

Historical, Clinical, Risk Management-20 (HCR-20)

Historical, Clinical, Risk Management-20 Version 3 (HCR-20V3)

History, Current Behaviour & Future (HKT-30)

Risk for Sexual Violence Protocol (RSVP)

Guidelines for Stalking Assessment and Management (SAM)

Structured Assessment of Protective Factors (SAPROF)

Spousal Assault Risk Assessment Guide (SARA)

Structured Assessment of Violence Risk in Youth (SAVRY)

Structured Clinical Judgment: Risk (SCJ:Risk)

S-RAMM

Short-Term Assessment of Risk and Treatability (START)

Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV)

Sexual Violence Risk-20 (SVR-20)

Workplace Assessment of Violence Risk (WAVR-21)

Workplace Risk Assessment-20 (WRA-20)

VITA

Caroline S. Chevalier, M.A.

Education

In Progress **Doctor of Philosophy in Clinical Psychology**

Sam Houston State University
Huntsville, Texas
Dissertation: The Association between Structured Professional Judgment Measure Scores and Summary Risk Ratings: Implications for Predictive Validity (proposed September 2015)

December 2014 **Master of Arts in Clinical Psychology**
Sam Houston State University
Huntsville, Texas
Thesis: Evaluators' Use of PCL-R and Static-99R Scores in Forensic Reports

May 2009 **Bachelor of Science in Psychology (Criminal Justice Minor)**
Magna Cum Laude
James Madison University
Harrisonburg, Virginia
Thesis: Change Blindness and the Weapon Focus Effect

Publications

Boccaccini, M. T., Chevalier, C. S., Murry, D. C., & Varela, J. G. (in press). Psychopathy Checklist-Revised Use and Reporting Practices in Sexually Violent Predator Evaluations. *Sexual Abuse: A Journal of Research and Treatment*. doi: 10.1177/1079063215612443

Cramer, R.J., Chevalier, C.S., Gemberling, T.M., Stroud, C. H., & Graham, J. (in press). A Confirmatory Factor Analytic Evaluation of the Klein Sexual Orientation Grid. *Psychology of Sexual Orientation and Gender Diversity*.

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., Varela, J. G. (2014). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior*, *39*, 209-218. doi:10.1037/lhb0000114

Boccaccini, M. T., Turner, D. B., Murrie, D. C., Henderson, C. E., & Chevalier, C. (2013). Do scores from risk measures matter to jurors? *Psychology, Public Policy, and Law*, *19*, 259-269. doi:10.1037/a0031354

Conference Presentations

Chevalier, C. S., Boccaccini, M. T. (2015, March). *A Comparison of Structured Professional Judgment Instrument Scores and Summary Risk Ratings.* A poster presented at the annual conference of the American Psychology-Law Society, San Diego, CA.

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C. (2014, August). *Static-99R Reporting Practices in Sexually Violent Predator Cases.* Poster presented at the annual conference of the American Psychological Association, Washington, DC.

Chevalier, C. S., Boccaccini, M. T., & Murrie, D. C. (2014, August). *PCL-R Score Reporting and Interpretation Practices in Sexually Violent Predator Cases.* Poster presented at the annual conference of the American Psychological Association (Washington, D. C.).

Chevalier, C. S., McCallum, K., Bryson, C., & Boccaccini, M. T. (2014, March). *Risk Instrument Use and Integration in Sexually Violent Predator Evaluations.* Poster presented at the annual conference of the American Psychology-Law Society (New Orleans, Louisiana).

Chevalier, C. S., Gemberling, T., Cramer, R.J., Stroud, C.H., & Graham, J. (2013, August). *Victimization and Double Minority Status: Moderating Effects of Sexual Identity*. Poster presented at the annual conference of the American Psychological Association (Honolulu, Hawaii).

Turner, D. B., Chevalier, C. S., Boccaccini, M. T., & Murrie, D. (2012, March) *Do SVP Jurors Believe that Offenders with Higher Risk Measure Scores are More Likely to Reoffend?* Poster presented at the annual conference of the American Psychology-Law Society, San Juan, PR.

Chevalier, C. S., Huddletson, J., & Andre, J. (2008, March). Poster presented at the annual conference of the South Eastern Psychological Association, Charlotte, NC.