EXPLORING THE RELATIONSHIP BETWEEN SCORING ACCURACY AND PREDICTIVE VALIDITY IN RISK ASSESSMENT USING THE SERVICE PLANNING INSTRUMENT (SPIn)

by

Julie Goodwin

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Arts

in

Psychology with Specialization in Data Science

Carleton University

Ottawa, Canada

Abstract

Accurate scoring is an essential component of a risk assessment's reliability, which in turn contributes to its predictive validity. This study tested a new method for evaluating the accuracy of completed risk assessments, described as intra-rater consistency. Pairs of cross-consistency items were identified in the Service Planning Instrument (SPIn), which were used to flag errors in completed assessments. Participants were 31,460 adults (20.5% female; 20.7% Indigenous) on community supervision in Alberta. The base rate of error was low; however, Indigenous participants had significantly higher rates of error. Contrary to my hypothesis, in the full sample assessments with errors had slightly higher predictive accuracy than assessments without errors. This pattern was repeated among non-Indigenous participants, but no significant differences in predictive accuracy based on errors were found among Indigenous participants. Differences in predictive accuracy based on sex and Indigenous status were observed in assessments with errors, but not assessments without errors.

Acknowledgements

First and foremost, I would like to express my gratitude for my phenomenal supervisor, Dr. Shelley Brown. It is not often that you come across someone who has such a clear, measurable impact on your development. I have felt myself grow, as both a researcher and a person, under your mentorship. Your unique ability to be caring and compassionate, while also encouraging your students to push themselves to produce the best work possible, is truly incredible. I feel unbelievably privileged to have been mentored by such a brilliant person through both my undergraduate and master's theses. Thank you for seeing my potential, even when I couldn't see it myself.

I would also like to thank my committee members, Dr. Ralph Serin and Dr. Craig Leth-Steensen. Your insight and feedback were absolutely essential to this research. I would also like to thank my external examiner, Dr. Steven Prus, and chair, Dr. Kevin Nunes for agreeing to participate in my graduate committee. I sincerely appreciate the time and effort you invested in my defence.

A very special thank you to Dr. David Robinson of Orbis Partners, who provided the original inspiration for this study. I am so thankful to you for sharing your ideas with me and for investing so much time and energy into providing your insight and feedback throughout this process. This study would most certainly not have been possible without your help. I would also like to extend my thanks to Christie Nicholson and the Alberta Solicitor General for supporting this research and providing the dataset. Needless to say, this study would not have been possible without you.

I would also like to express my gratitude for the incredibly supportive Gender and Crime lab, and specifically to my friend Colleen. Your friendship, support, and encouragement made

what could have been a very difficult and lonely experience into one that I will remember and cherish. Thank you for always being there to bounce ideas off of, commiserate when things didn't go as planned, and laugh about all these challenges along the way.

Finally, I would like to thank my family. To Nick, my wonderful partner, thank you for being by my side and cheering me on at every step—even when you didn't understand what I was doing. To my cats Gus and Cheeto, thank you for keeping me company and making me laugh. Above all, I must express my sincere and deep appreciation for my wonderful, sweet dog Sydney, who I lost just as I was finishing this thesis. I cannot begin to express the gratitude I feel for her unconditional love and support over so many years. I am who I am today because of Sydney's friendship.

Table of Contents

List of Tables

List of Figures

List of Appendices

The Relationship Between Scoring Accuracy and Predictive Validity in Risk Assessment

The justice system has long had the responsibility of evaluating those who have committed crimes to determine how likely they will be to engage in further criminal behaviour following release, or while incarcerated; this process is known as risk assessment. The manner in which risk is assessed has evolved considerably over the last 100 years, moving from the arbitrary judgment of justice system professionals to structured assessments based on empirically and theoretically derived factors. Modern risk assessments provide structured guidelines for raters who employ their professional judgment to assess justice-involved people on a range of factors that contribute to their overall likelihood of engaging in criminal behaviour (Bonta & Andrews, 2017). These assessments guide professionals in both estimating the risk that an individual will re-offend in the future, and in developing interventions for justice-involved individuals.

An extensive body of research exists evaluating risk assessments, most of which has focused on two key, interdependent elements which may determine the quality of a tool: reliability and validity. Reliability research has primarily depended upon assessments of inter-rater reliability (i.e., the amount of agreement between two raters' scores on independently conducted assessments), while studies assessing validity have primarily considered the predictive accuracy of risk assessment tools. While conducting tests of inter-rater reliability is entirely feasible in research settings, it is often impractical in field settings where these tools are being used. As such, limited research exists assessing the inter-rater reliability of many risk assessments, such as the Dynamic Factor and Identification Analysis (DFIA; Correctional Service of Canada). Similarly, the Service Planning Instrument (SPIn; Orbis Partners, 2003) is

one such assessment tool with no inter-rater reliability data despite being widely used in Canada and the United States by agencies that serve justice-involved people.

The inter-rater reliability of such tools must be assessed as poor inter-rater reliability will invariably translate into poor predictive validity. If the measurement of inter-rater reliability is not feasible, an alternative method to assess the general reliability of risk assessment tools has been proposed by Dr. Jones and Dr. Robinson of Orbis Partners, Inc. This method focuses on intra-rater consistency (or intra-rater reliability) between conceptually similar items within the same measure as one aspect of reliability, rather than inter-rater reliability.

The present study will examine if intra-rater consistency between conceptually similar SPIn items can enhance the predictive validity of the SPIn. Pairs of items will be identified which may be used to flag inconsistencies in scoring. For example, if an individual has a history of violent offences identified in the criminal history domain, corresponding evidence of aggressive behaviour should be found in the aggression domain. Using these paired cross-consistency items, inconsistencies in completed assessments may be identified. With this information, completed assessments may be categorized based on their degree of intra-rater consistency. This information can then be used to explore how intra-rater consistency may impact the predictive validity of the SPIn as a tool for predicting future justice-involvement. Enhancing the accuracy of scoring in assessments will benefit front-line professionals, and by extension the clients they serve, in improving the utility and predictive validity of completed assessments.

The literature review is organized as follows. First, a brief description of the SPIn is provided, along with an overview of the research to date. Then, an analysis of classical test theory will be presented, and the relationship between reliability and validity in psychological

assessments will be discussed. Next, potential sources of bias, such as the sex or Indigenous status of the individual being evaluated, will be explored. Lastly, the literature is summarized including gaps, and the current study research questions and hypotheses are described.

The Service Planning Instrument (SPIn)

The SPIn (Orbis Partners, 2003) is a risk, need, and strength assessment tool designed for use with justice-involved adults. The SPIn is comprised of 90 items surveying 11 unique content domains, which contribute to an individual's risk of future justice involvement and highlight criminogenic needs that may be useful targets in treatment. The SPIn Pre-Screen assessment is composed of a subset of 35 items and is used primarily for classification and risk prediction. Clients who are deemed to be medium- or high-risk based on the results Pre-Screen assessment must complete the Full Assessment, which offers additional information which is useful for case management purposes. At this time only four studies have examined the validity of the SPIn (Jones et al., 2015; Jones & Robinson, 2017a, 2017b, 2018); a summary of the populations, samples, and results of these studies may be seen in Appendix A. The results of these studies demonstrate support for both the classification and predictive validity of the SPIn.

SPIn Pre-Screen Classification Results

The first validation study available was published by Jones et al., (2015). A sample of Pre-Screen assessments completed between 2009 and 2011 for 3,656 individuals (19% female) serving community sentences in Alberta were included. The Pre-Screen classification of individuals into low, medium, and high risk and strength categories was found to be highly effective. As one would expect, individuals classified as low risk reoffended at lower rates than those classified as either medium or high risk (low risk = 8.6%, medium risk = 25.9%, high risk = 54.4%). Similarly, those classified as low strength reoffended at higher rates than those who

were classified as either medium or high strength (low strength = 33.7%, medium strength = 18.4%, high strength = 7.7%). The pattern of relationships between risk, strength, and recidivism were consistent across sex (male or female) and race (Indigenous or non-Indigenous).

A sample that partially overlaps with that used in Jones et al. (2015) was included in a subsequent study by Jones and Robinson (2017a). A total of 46,794 (21% female) SPIn Pre-Screen assessments were included in analyses, with recidivism follow-up measures at three time points: one year, three years, or six years after the initial assessment was completed. Classification results were comparable to those found by Jones et al. (2015); lower risk individuals reoffended at lower rates than higher risk individuals (low = 19.2%, moderate = 41.5%, high = 62.7%), and lower strength individuals reoffended at higher rates than lower risk individuals (low = 43.3%, moderate = 28.0%, high = 18.2%).

Jones and Robinson (2017b) conducted a nearly identical study evaluating the validity of the SPIn, this time in a sample of adult probationers in Clark County, Washington (N = 2248, 22.2% female). Classification results were very similar to those previously found, with higher risk individuals reoffending more often (low = 22.9%, moderate = 45.7%, high = 67.4%), and higher strength individuals reoffending less often (low = 58.2%, moderate = 39.4%, high = 28.1%).

Jones and Robinson (2018) made use of another sample from Alberta corrections, this time including 752 participants (16.5% female) assessed while in custody. Classification results followed the same pattern previously described, with higher risk individuals reoffending at higher rates than lower risk individuals (low = 18.3%, moderate = 49.3%, high = 78.8%). Taken together, the results of these studies indicate that the SPIn Pre-Screen is an excellent classification tool.

SPIn Pre-Screen Predictive Validity Results

The SPIn Pre-Screen risk score also demonstrated strong predictive validity in Jones and colleagues' 2015 study; an AUC value of .77 was found for the overall sample, which according to Rice and Harris (2005), is considered a large effect size. Only small variations in AUC values were observed among sub-samples based on sex and race (Indigenous vs. non-Indigenous), where values ranged from .77 for males, to .75 for Indigenous people. Finally, to assess the utility of including Pre-Screen strength scores in the prediction of recidivism, authors conducted a regression analysis. Results indicate significant main effects of both risk and strength scores, as well as a significant interaction between risk and strength scores. This suggests that not only are both risks and strengths useful in the prediction of recidivism, but that strengths may have a buffering effect on risk, where high strength scores have a greater impact on recidivism among those who are high risk than among those who are low risk. These effects were consistent among sub-groups, with the exception of Indigenous individuals, for whom the interaction between strengths and risks was non-significant.

Jones and Robinson (2017a) also found that SPIn Pre-Screen risk scores showed moderate to strong predictive validity (based on the criteria provided by Rice & Harris, 2005). AUC values ranged from .64 for any conviction at one, three, and six-year follow-ups; to .75 (one year) or .74 (three and six years) for any custody. A similar pattern of results emerged with Pre-Screen strength scores, though AUC values were small to moderate. At one-year follow-up, AUC values ranged from .61 (any conviction) to .68 (any custody), while at three-year follow-up AUC values ranged from .59 (any conviction) to .67 (any custody and technical violations). Finally, at the six-year follow-up AUC values ranged from .62 (violent offence and conviction resulting in custody) to .68 (any custody). These results remained stable across sex and

Indigenous status. Jones and Robinson (2017b) found similar results with the SPIn Pre-Screen in Clark County, Washington; the results of ROC analysis indicated moderate to strong predictive validity; AUC values ranged from .64 at the one-year follow-up to .70 at the five- and six-year follow-ups.

Jones and Robinson (2018) made use of another sample from Alberta corrections, this time including 752 participants (16.5% female) assessed while in custody. The results of ROC analysis with the SPIn Pre-Screen were similar to those previously discussed, with effect sizes for the overall sample ranging from moderate (technical violations AUC = .69, $p < .001$), to strong (new offences AUC = .77, $p < .001$). These results were comparable across sexes and in both Indigenous and non-Indigenous people.

The results of predictive validity analyses indicate that the SPIn offers moderate to strong predictive validity, and that these effects are consistent across sexes, Indigenous and non-Indigenous people, and different types and times to follow-up. The inclusion of strength demonstrates incremental validity beyond risks, further supporting the structured assessment of strengths to improve the predictive accuracy of risk assessments. Overall, the SPIn appears to have solid predictive validity. However, little is known about the SPIn's reliability or the consistency of raters' SPIn scoring. The following section will explore classical test theory (CTT) and the means CTT offers for evaluating the quality and accuracy of psychological assessments.

Psychometrics & Classical Test Theory

Classical test theory (CTT) provides the foundation upon which most psychological measures—including risk assessments—have been built (DeVellis, 2006). CTT expresses a set

of basic concepts that can inform the development of psychometric tests and provides analytic methods which allow us to interpret and make inferences based on the results of those tests.

In psychology, many of the variables and processes that are of interest to researchers cannot be measured directly. For example, variables such as an individual's social skills or their motivation to change behaviour are difficult to quantify. Instead, we must develop a means of evaluating such internal processes through observable behaviour. For example, while we cannot directly see or measure an individual's motivation, we can estimate it based on the way they describe feeling about behaviour change, steps towards change they may have made, or expressed self-efficacy to accomplish a change. These are called *proxy indicators* (DeVellis, 2006). With the methods provided by CTT, we can evaluate how well our proxy indicators perform.

Prior to assessing how effectively an item or scale performs, it is important to understand item properties under CTT. For any given item on a scale, such as an individual's motivation for change, there exists both a *true score* and an *observed score* (DeVellis, 2006). The true score is the most accurate version of an individual's score, or the version of a score that we would obtain if it were possible to see or measure the variable directly. It is impossible to be certain of a true score, however we can estimate it based on an observed score. The observed score is the score an individual actually receives on a given measure. CTT asserts that we must assume the observed score is dependent on the true score, however some measurement error will naturally occur. To illustrate, an individual's motivation for behavioural change as scored on the SPIn is related to their true motivation, but also includes some error, which may arise from the judgment of the interviewer. A well-constructed measure paired with training in the use of that measure may minimize the impact of an individual's biases, however we cannot expect to reduce error in

measurement to zero. CTT further assumes that error is random, and that ultimately errors throughout a scale will balance and leave the mean relatively unaffected.

Measures developed under CTT can be broadly grouped into two categories: norm-referenced scales, and criterion-referenced scales (Helmus & Babchishin, 2017). Norm-referenced scales are most common in psychology and describe those scales which are intended to measure the presence of a single underlying construct, such as intelligence or a personality type. Criterion-referenced scales are designed to predict a given outcome. Risk assessments offer one example of criterion-referenced scales, which are intended to predict whether or not someone will engage in criminal behaviour in the future. This distinction is important, as appropriate validation methods are different for each kind of scale. According to Helmus and Babchishin, norm-referenced scales are best assessed using measures of internal reliability and construct validity, while assessment of criterion-referenced scales should be primarily concerned with predictive validity.

While Helmus and Babchishin (2017) clearly describe the methods they assert are most appropriate for evaluating risk assessments (i.e., examination of predictive validity), a consensus has not been reached among academics who develop and evaluate risk assessments. Many academics continue to assert the necessity of evaluating the reliability – particularly inter-rater reliability – of risk assessments, as an essential contributing factor to a measure's predictive validity (e.g., Desmarais et al., 2016; Duwe & Rocque, 2017; Lowenkamp et al., 2004). Subsequent sections will describe various strategies used to assess the reliability and validity of measures such as risk assessments, followed by a discussion of the relationship between scale reliability and validity.

Reliability

Reliability is an essential component of a scale's overall validity under CTT. Broadly, reliability means that a tool can be used in a replicable manner, with consistent, predictable results (DeVellis, 2012). In the context of the SPIn, reliability means that the items within the SPIn can be scored in a predictable manner, based on observable traits and behaviours. In more technical terms, a scale or item's reliability can be quantified as the "proportion of variance attributable to the true score of a latent variable" (DeVellis, 2012, p. 31). Perfect reliability occurs when the observed score is identical to the true score, or when 100% of the variance in scores is attributable to the true score of the latent variable. Measurement without error is generally impossible in psychology where we are typically concerned with latent variables that can be difficult to quantify. However, it is possible to estimate the reliability of a scale or tool.

Reliability has been traditionally assessed in one of three ways. Internal consistency, meaning the amount of agreement or similarity of items within a scale measuring a given latent variable, is one such method of assessing reliability. Inter-rater reliability is commonly used to determine how consistently a given measure can be scored by evaluating the similarity between rater's scores. Alternate forms reliability is also used in some circumstances, where diverse items or measures that quantify the same or related latent variables can be compared. Intra-rater consistency, the method investigated in the present study, draws on the concepts of inter-rater and alternate forms reliability to create novel approach for estimating the accuracy of scoring in completed assessments

Internal consistency. A researcher may wish to measure the internal consistency of a scale that contains a series of items meant to evaluate the presence of a single underlying construct (Gravetter & Forzano, 2015). For example, an intelligence test contains a variety of

items, all of which will tap into the same underlying trait of intelligence. An effective scale that taps into a single latent variable should have a high level of internal consistency, whereas a measure intended to assess a diverse range of traits will have lower internal consistency. The threshold for what constitutes a good alpha is debated, however most authors agree that a minimum alpha of .65 is acceptable (Goforth, 2015).

Helmus and Babchishin (2017) contend that internal consistency and other related measures are illogical methods to evaluate a risk assessment's overall reliability. These methods for evaluating reliability are predicated on the assumption that a scale is assessing a single underlying construct (i.e., a norm-referenced scale), however this is not the case with risk assessments. Risk assessments follow a cumulative stochastic model, where a diverse range of items are assessed, and risk increases as a greater number of risk factors accumulate. A diverse range of items is indicative of a strength in a risk assessment, rather than the weakness it might be in a norm-referenced scale.

Results from studies assessing the SPIn's internal consistency provide an illustration of why this may not be an important measure of a risk assessment's reliability. Overall, the SPIn demonstrates good internal consistency (e.g., total risk $\alpha = .86$; total strength $\alpha = .93$; Jones & Robinson, 2018), however considerable variability was found in some domain scores (range from $\alpha = .21$ in the mental health domain to $\alpha = .90$ in the social/cognitive skills domain). These differences in alpha values provide an illustration of what measures of internal consistency represent. The total risk and strength scores likely demonstrated good internal consistency because they each tap into a single underlying construct: risk and strength respectively. However, individual domains contain a variety of contributing items. The stability domain, for example, likely produced low alpha values (range from $\alpha = .46$ to $\alpha = .67$) because this domain

contains questions pertaining to housing security, financial security, transportation, and life skills. It is entirely possible that a client may have, for example, stable housing, but lack transportation and life skills. Thus, we can see that a risk assessment's internal consistency is not a good indicator of its reliability.

Inter-rater reliability. Inter-rater reliability is one of the most commonly used methods for assessing the reliability of an instrument in psychology, particularly forensic risk assessments. Inter-rater reliability refers to a measure of the similarity of two rater's scores on a given scale (DeVellis, 2012; Gravetter & Forzano, 2015). This concept relies on the assumption that scoring is representative of observable factors and not confounding factors internal to the raters themselves. An adequate measure should provide enough guidance that raters can score the measure with a high degree of similarity.

A number of statistical and methodological strategies exist to allow for the estimation of inter-rater reliability, most of which depend on the Intraclass Correlation Coefficient (ICC), Crohnbach's alpha, or some variant thereof (DeVellis, 2012; McNeish, 2018). ICC values can range from 0 to 1, where 1 represents complete agreement. Perfect agreement between two raters' scores is unlikely, particularly when considering items that are rated on a continuum or require some discretion or interpretation on the part of the rater (e.g., an individual's motivation for change). As such, an ICC of .70, or 70% agreement between raters, is typically seen as acceptable (Multon & Coleman, 2018).

Inter-rater reliability is an excellent method for assessing the reliability of a risk assessment, when feasible. This is a practical option in many research settings; one member of a research team will typically conduct the majority of assessments and another team member will perform their own assessment of a subset of participants without seeing the first researcher's

scores. Then, an ICC value or the percent of agreement will be calculated to measure the similarity between the two raters' scores. Unfortunately, this is rarely possible in field settings where assessments like the SPIn are being used by justice system professionals. Professionals using tools like the SPIn, such as probation and parole officers, are typically burdened with large caseloads for whom they must conduct assessments and engage in case planning and management. As a result, these professionals often already have a considerable amount of work to complete within a limited timeframe. It would be unrealistic to expect these professionals to find time to conduct additional assessments of their colleagues' clients. For this reason, the SPIn and many other risk assessments in use have been limited in their ability to assess inter-rater reliability, particularly in the field.

Desmarais et al., (2016) conducted a systematic review of 53 studies in which 19 risk assessments used in U.S. adult correctional settings were evaluated. Of these 53 studies, only two reported on the inter-rater reliability of the measures under evaluation: Simourd (2006) assessed the Level of Service Inventory-Revised (LSI-R), while Walters (2011) evaluated the Level of Service Inventory-Revised: Screening Version (LSI-R: SV). While these two studies both reported good inter-rater reliability, other studies which have evaluated the inter-rater reliability of risk assessments have found less encouraging results. For example, Rocque and Plummer-Beale (2016) also evaluated the inter-rater reliability of the LSI-R, but found considerable variability in the measure's reliability, both overall and across sub-domains. The total LSI-R scores had an ICC value of .65, which is considered good reliability based on Hallgren (2012). However, sub-domain inter-rater ICC values ranged from .20 in the leisure/recreation domain to .80 in the accommodations domain. Seven of the fifteen domain inter-rater ICC values that were calculated fell below the "acceptable" cut-off of .40, as defined by Hallgren (2012). Chadwick

(2014) points out that those studies which reveal lower inter-rater reliability typically evaluate risk assessments and sub-domains of risk assessments which include items representing hard-to-define constructs that require a subjective evaluation on the part of the rater (e.g., Austin et al., 2003; Rocque & Plummer-Beale, 2016). Many risk assessments in use today—including the SPIn—contain many such subjective items, and as such it is of even greater concern that the inter-rater reliability of the SPIn and many other risk assessments has yet to be assessed.

For example, the Dynamic Factor and Identification Analysis (DFIA) component of the Offender Intake Assessment (OIA) used with federally sentenced people in Canada has only been subject to a limited investigation of reliability. Brown and Motiuk (2005) conducted a comprehensive meta-analysis and psychometric review of the DFIA, however only assessed reliability through internal consistency. While their results indicate that the DFIA shows acceptable to superior internal consistency, this is insufficient evidence to conclude that the DFIA is a reliable tool. In a similar review of the revised DFIA (DFIA-R), Stewart et al. (2017) also only reported on the DFIA-R's internal consistency; no mention of inter-rater reliability was made. As previously indicated, internal consistency is not a logical method for assessing the reliability of risk assessments, and thus the reliability of these tools—and many other risk assessments including the SPIn—remains unknown.

In addition to the limited research available on the inter-rater reliability of risk assessments, Chadwick (2020) describes a further limitation of those few studies which have assessed inter-rater reliability. Most studies which report inter-rater reliability typically involve assessments completed by researchers, either based on a file review, audio-taped interviews, or interview transcripts. While these researchers are trained in the use of the assessments, they are not the professionals who are actually using the tools in the field for their intended purpose, nor

are they conducting their assessments in field settings. As such, even those studies which have evaluated inter-rater reliability must have the ecological validity of their results questioned.

While no studies to date have evaluated the inter-rater reliability of the SPIn, Jones and Robinson (2018) suggest that the reliability of the SPIn is likely quite similar to that of the YASI, a highly similar tool also developed by Orbis Partners Inc. for use with justice-involved youth. Three studies have been published thus far that evaluated the inter-rater reliability of the YASI, two of which assessed reliability in scoring among front-line professionals. The results of these studies have been mixed. Geck (2012) found that ICC values measuring the overall agreement in YASI total risk and protective scores were good, ranging from .76 (static risk) to .93 (dynamic protective). However, ICC values for subdomains demonstrated more inconsistency. While some domains had extremely high levels of agreement (e.g., alcohol/drug use ICC = .99), others showed very low levels of agreement (e.g., peers ICC = .34). It is important to note that Geck (2012) employed academic raters only; no front-line professionals were included in her assessment of inter-rater reliability. Baird et al. (2013) also found that there was a considerable range in the levels of agreement attained by raters who were experts (i.e., clinical psychologists and academics) and front-line professionals. When assessing the level of agreement in rating the YASI overall, 84.7% agreement was attained among front-line professionals and 79.4% agreement was found between front-line professionals and experts. However, agreement in sub-domains was not as strong. Levels of agreement ranged from 46.3% to 100%, and only 10% of all domains exceeded 75% agreement. Finally, Kennealy et al. (2017) assessed the inter-rater reliability of scores produced by front-line staff on a modified version of the YASI, used by the California Department of Corrections. Comparatively low levels of agreement were found even in total YASI scores, with an ICC value of .63. Only 59% of staff scores exceeded the authors

definition of "good" scoring (ICC and Kappa values exceeding .60). Additionally, nearly half of all subscale scores (5 of 11) fell below the "good" cut off as defined by authors. These results are concerning given fact that this study assessed the agreement among professionals who actually use the YASI for its intended purpose, rather than expert or academic raters.

Alternate forms reliability. In its most limited definition, alternate forms reliability involves assessing the degree of agreement between two scales which measure the same underlying construct (DeVellis, 2012; DeVon et al., 2007). The two scales compared for alternate forms reliability must be presumed equivalent, and to be measuring the same construct in different ways. Using a statistical method such as the ICC, the degree of similarity in the results of the two scales can be measured. This method is not typically used in the field of forensic risk assessment.

Intra-rater consistency. The principles of inter-rater and alternate forms reliability constitutes the basis for the approach that will be tested by the present study. Rather than compare two scales intended to assess a single underlying construct, individual items (within the same assessment) which assess the same underlying construct will be compared, thus assessing the consistency within a single rater's scoring. Many items within the SPIn assess the same or related constructs in different ways. For example, if a history of violent offences has been indicated in the criminal history domain, evidence of violent behaviour should be found in the aggression/violence domain. By making such comparisons between individual items contained in the SPIn, we may be able to assess the consistency of individual assessments—what I have coined intra-rater consistency. The method employed in the present study will not represent a classic interpretation of alternate forms reliability, however the basic concept underlying

alternate forms reliability will be adapted to the challenge of assessing and improving the reliability of the SPIn.

In the present study, intra-rater consistency is measured as a function of detectable rater coding errors; the amount of error observed is used as a proxy indicator for scoring accuracy. The strategy of flagging errors and assessing resulting changes in predictive validity was originally conceived of by Dr. Natalie Jones and Dr. David Robinson of Orbis Partners Inc., for use with the YASI. While no published reports are currently available, Dr. Robinson has shared promising preliminary results, demonstrating that more consistently scored YASI assessments have greater predictive validity than less consistently scored assessments (D. Robinson, personal communication, August 5, 2020). For example, using the YASI Pre-Screen risk score, assessments with lower rates of error had an AUC of .76, while assessments with higher rates of error had an AUC of .70.

Few published studies have attempted to evaluate the accuracy of individual raters' scoring using alternative methods to infer scoring accuracy. Meade and Craig (2012) tested a variety of methods to identify careless responding in self-report data. While some of the methods evaluated in this study were only relevant to self-report survey data, they did evaluate several indices of scoring consistency which are similar to the concept of intra-rater consistency that is of interest in the present study. Results indicated that inconsistencies in scoring were useful in identifying completely random responding and were closely related to multivariate outliers based on Mahalanobis distances. As this study was primarily concerned with methods for identifying careless responding in self-report data from undergraduate research participants, it is unclear to what extent these results will translate to rater-based measures used by professionals, like the SPIn.

Scoring accuracy of risk assessments has also been estimated using alternative methods in several studies. Chadwick (2014) found that DRAOR assessments completed by correctional officers who had received formal training in the use of the tool had higher predictive accuracy than those completed by officers who had only informal training in the use of the tool. Hanson et al. (2015) attempted to evaluate the fidelity with which a sample of 139 Canadian correctional officers could apply a series of measures of risk for sexual recidivism following completion of training. Significant differences were observed in the predictive accuracy of assessments that were completed (i.e., all assessment items were scored) as compared to incomplete assessments (i.e., some items left blank). Complete assessments had significantly larger AUC values, particularly in predicting sexual recidivism. It was further observed that officers tended to be consistent in the assessments they submitted; some officers reliably submitted completed assessments, while others tended to submit incomplete assessment. Authors hypothesized that this pattern was indicative of the officers' conscientiousness; those who were more conscientious submitted more complete assessments. Relatedly, Miller and Maloney (2013) conducted a latent class analysis of American justice system professionals' scoring practices and found that only half of their sample completed assessments diligently and used these results to guide case planning decisions. The remaining half of participants were found to be either completing assessments accurately but not incorporating the results into their case planning practices, or else were neither completing assessments nor using these tools to guide service provision. The impact on predictive validity was not examined.

The results of these studies indicate that proxy indicators of scoring accuracy can be useful, and that differences in predictive accuracy are revealed when assessments can be sorted based on raters' scoring accuracy. Unfortunately, none of these methods of inferring rater

accuracy allow for the rate of error to be quantified, or the errors themselves to be evaluated. Intra-rater consistency may be a useful proxy indicator for rater accuracy, with the added benefit of being able to identify sources of error in completed assessments. Scoring accuracy is an essential component of reliability, which in turn forms the basis for test validity. The following section will explore the relationship between reliability and validity in greater detail.

From Reliability to Validity

According to CTT, risk assessments such as the SPIn are only useful and valid if they have been scored reliably (DeVellis, 2006). Inaccurate and inattentive scoring could have a significant impact on how clients are classified, and thus can have enormous impact on the validity of the assessment. Reliability is an essential component of validity (Gravetter & Forzano, 2015). Validity refers to the degree to which a given assessment actually measures what it is intended to measure. For example, an assessment such as the SPIn is only valid if it truly assesses risk and strength. In establishing a tool's reliability, a number of other essential properties of assessment may be attained. First, statistical power increases as reliability increases. Additionally, a tool's generalizability is improved with greater reliability. Finally, and most importantly, predictive validity is only achieved with sufficient reliability (DeVon et al., 2007).

The statistical power of a given test refers to the probability of finding a significant result if one exists (Gravetter & Wallnau, 2014). For example, if one is assessing the efficacy of risk classification of the SPIn, one may wish to conduct a test to determine whether there are significant differences between the recidivism rates of those classified as high risk and low risk. A test with more power will be capable of detecting smaller effects or differences between these groups than one that has less power. A variety of factors impact a test's power, including the

sample size, the size of the effect or difference between groups, the accuracy of scoring and reliability of the measure used. Increased accuracy and reliability improve statistical power because as the rate of error decreases, the ratio of explained variance to unexplained variance will increase, which is the base form of many statistical tests.

The generalizability of a measure will also be increased by improved reliability. Generalizability, a property of the external validity of a tool, is the extent to which a given result or measure can be used to the same effect in other studies (Gravetter & Forzano, 2015). A reliable measure can be more easily used, and the results gleaned from the measure will be more accurate, and thus more generalizable to other situations. Essentially, a reliable tool will be more likely to produce meaningful information regarding the underlying latent variables. In the case of a risk assessment, this is an essential component of predictive validity.

Predictive validity refers to the extent to which scores on a given measure are predictive of some future outcome (DeVon et al., 2007). In the case of risk assessments, we are concerned with the extent to which results on a given assessment predict whether or not an individual will recidivate in the future (Bonta & Andrews, 2017). A more accurately scored and reliable tool will be more likely to demonstrate predictive validity, as a result of both the increased statistical power and improved generalizability that arise from good reliability. While the reliability of the SPIn has not been properly assessed, several studies have evaluated its predictive validity using Receiver Operating Characteristic (ROC) analysis. These studies have demonstrated the predictive validity of total risk and strength scores (Jones & Robinson, 2017a, 2017b, 2018) and individual risk and strength domain scores (Brown et al., 2020).

Potential Sources of Bias

As previously mentioned, one of the core underlying assumptions of CTT is that errors in measurement are random, and not based on extraneous factors such as a rater's personal biases (DeVellis, 2012). A well-constructed measure accompanied by training, such as the training provided to those who use the SPIn, can reduce the amount of error that occurs, however it is unrealistic to assume that errors will be reduced to zero. The assumption that errors occur randomly is predicated on effective training that corrects a rater's personal biases. While this may be the case, it cannot be known whether errors are in fact occurring at random unless we investigate those errors, and the impact of potential sources of bias on error rate. Two potentially relevant sources of bias have been described in the literature and will be explored in subsequent sections: sex and Indigenous status.

Sex Bias in Risk Assessment. The differences in male and female justice involvement are well-established. We know that males commit more crime overall, and specifically more violent crime; females commit less crime overall, but more property crime and fraud as compared to other kinds of offences (Malakieh, 2019). Furthermore, the pathways to justice-involvement differ for males and females (e.g., Salisbury & Van Voorhis, 2009). As such, the legitimacy of using the same risk assessments with males and females has been called into question by many scholars (Belknap, 2015; Salisbury & Van Voorhis, 2009; Wattanaporn & Holtfreter, 2014). Research has demonstrated that some gender-neutral risk assessments over-classify women, meaning that women are classified as being higher risk than they truly are (e.g., Hamilton, 2019; Hannah-Moffat & Shaw, 2001). The SPIn has been found to be similarly predictive for males and females (e.g., Jones & Robinson, 2017a, 2017b, 2018), however misclassification is not the only kind of sexism that women experience in the justice system.

Arguably the most well-known form of sexism in the justice system is the manner in which female victims are treated in sexual assault cases. Female victims frequently have their character questioned, and those who do not conform to the patriarchal concept of an 'ideal victim' are blamed for their own victimization (Dick, 2020). Similarly sexist attitudes have been documented in juvenile probation officers: in one study clients who shared histories of victimization and abuse with their probation officers were seen as attention-seeking and manipulative (Gaader et al., 2004). Other forms of sexism arise in justice system, related to the perceived seriousness of crimes and the severity of sentences. So-called 'benevolent sexism'—which perpetuates traditional gender roles and the perception that women are in need of protection—has been found to result in more lenient sentences being given to females than males who have committed similar crimes (Connelly & Heesacker, 2012; Cutroni & Anderson, 2021; Stevens et al., 2021). Furthermore, in mock jury studies which ask raters to evaluate the seriousness of the identical crime scenarios, male perpetrators typically receive more serious ratings than female perpetrators (Herzog & Oreg, 2008; Stroh et al., 2016). However, when the female perpetrator is described as non-traditional and in violation of gender norms, their crimes were actually perceived as being more serious than those of males (Herzog & Oreg, 2008). No studies thus far have investigated whether these forms of sexism impact the accuracy of scoring on risk assessments, however, exploration of the potential impacts of sexism on the consistency of risk assessment scores is worthwhile.

Indigenous Status Bias in Risk Assessment. It is widely understood that Indigenous people are overrepresented in the justice-involved population of Canada. In 2017-2018 Indigenous adults accounted for 30% of provincial custody admissions and 29% of federal custody admissions, but only 4% of the general population (Malakieh, 2019). Despite national

efforts to ameliorate the treatment of Indigenous Canadians, such as the work of the Truth and Reconciliation Commission and the Inquiry into Missing and Murdered Indigenous women, this problematic over-representation of Indigenous people in Canada's justice system has continued to increase in recent years (Office of the Correctional Investigator, 2019). McGuire and Murdoch (2021) suggest that the over-representation of Indigenous people in carceral settings is indicative of continued settler-colonial practices that are intended to assimilate or exterminate Indigenous people. The House of Commons Standing Committee on Public Safety and National Security published a report in June 2021 which identified systemic racism in policing practices in Canada, and the need to address problematic practices such as racial profiling and discriminatory use of force (McKay, 2021). While no similar reports exist on racism in community corrections, it is reasonable to suspect similarly problematic practices may be found in these settings.

In the landmark case of Ewert v Canada (2018), the Supreme Court of Canada determined that Indigenous people have the right to be assessed with risk assessment tools that have been validated for use with Indigenous people. Since this time there has been a proliferation of research which has demonstrated that many risk assessment tools are not as effective predictors of future justice-involvement for Indigenous people as their non-Indigenous counterparts. For example, in their 2013 meta-analysis, Gutierrez and colleagues evaluated the relative predictive accuracy of the central eight risk factors described by Bonta and Andrews (2017). Their results indicate that while all factors were significant predictors of recidivism regardless of race, six of the eight factors had lower predictive accuracy for Indigenous people than their non-Indigenous counterparts. Numerous other studies have found similar results, across a range of different measures (e.g., Babchishin et al., 2012; Lee et al., 2020; Perley-Robertson et al., 2019). Those studies which have evaluated differences in the predictive

accuracy of the SPIn as a function of Indigenous status have demonstrated that while the SPIn is an effective predictor for Indigenous and non-Indigenous clients, predictive accuracy tends to be slightly higher among non-Indigenous participants (e.g., Jones & Robinson, 2017a). These lower levels of predictive accuracy have not yet been adequately explained in the literature, and differences in the rate of scoring errors as a function of Indigenous status have yet to be explored. One possible explanation is that a rater's personal racial biases may be influencing their attentiveness and the accuracy of their scoring.

Intersection of Sex and Indigenous Status. In 1989, Kimberlé Crenshaw introduced the concept of intersectionality, a now widely used framework for understanding the compounding effects of discrimination individuals experience when they belong to multiple marginalized groups. In other words, the interaction between sexism and racism (and other forms of systemic oppression) augments the impact of each. Evidence of the compounding effects of sexism and racism has been observed in the Canadian justice system. While the rate of incarceration has been declining for several years in Canada, the rate of incarceration for Indigenous people increased by 14.7% between 2013 and 2017 (Public Safety Canada, 2018). When sex is taken into consideration, the disparities are even larger: admissions to federal custody increased by 28% for Indigenous males, while an increase of 66% was observed in Indigenous females (Malakieh, 2019). Research has also demonstrated that race and sex together impact the way criminal justice system professionals approach their clients, influencing the harshness of sentences and the nature of the programming they are offered (Leiber et al., 2016). Experiences of discrimination have also been shown to impact women of colours' motivation and willingness to participate in rehabilitative programs (Wesely & Miller, 2018). If sex and Indigenous status

impact the accuracy of risk assessment scores uniquely, it follows that sex and Indigenous status may interact to further influence the consistency and reliability of risk assessments.

Summary

Modern risk assessments are multi-purpose tools, which allow both for the prediction of future justice involvement and the provision of guidance for professionals who match justice-involved individuals with appropriate treatments and programming to reduce their risk. To effectively serve both of these uses, risk assessments have become complex and include items assessing a diverse array of factors in an individual's life. With this increase in complexity, it has become more important than ever that the accuracy of scoring, reliability, and validity of risk assessments be evaluated.

Despite the importance of assessing the reliability of risk assessments, there exists little research which has done so effectively. While many would argue that inter-rater reliability is the most appropriate method for evaluating the reliability of risk assessments, relatively few tools have been evaluated in this way (Chadwick, 2020; Desmarais et al., 2016). Instead, much research has been conducted evaluating the internal consistency of risk assessments (e.g., Brown & Motiuk, 2005; Stewart et al., 2017 Yesberg & Polascheck, 2015). Regrettably, as Helmus and Babchishin (2017) indicate, internal consistency is a poor indicator of a risk assessment's reliability, as diversity in items is desirable in risk assessment.

Additionally, as Chadwick (2020) points out, the few studies that have evaluated the inter-rater reliability of risk assessments have yielded mixed results, with poorer reliability found for measures that require the subjective evaluation of dynamic factors. The SPIn contains many such items, and as such the fact that no studies have evaluated its inter-rater reliability is problematic. Further, studies which have evaluated the inter-rater reliability of other risk

assessments employ methods that limit the ecological validity of their results; the in-field reliability of many risk assessments has not been established.

Finally, no research to date has examined whether errors in risk assessments—assumed to occur at random by CTT—are truly occurring in a random fashion. Studies which have used proxy indicators of scoring accuracy have been unable to identify errors; rather, these studies have only demonstrated that differences in predictive accuracy may be observed based on assumed differences in scoring accuracy (e.g., Chadwick, 2014; Hanson et al., 2015). It has been well-documented that both sexism and racism impact the treatment of justice-involved people in Canada. It remains unknown if these biases impact the accuracy of risk assessments conducted with marginalized groups.

The gaps in the literature regarding the reliability of risk assessments are of great concern, as the predictive validity of such tools is inextricably linked to their reliability. It is possible that by improving the accuracy and hence the reliability of risk assessments, their predictive validity might also increase, as well as their utility in guiding treatment planning for justice-involved people. The present study seeks to address these gaps in our understanding of scoring accuracy, reliability, and validity in risk assessment.

Current Study

The field of risk assessment has been hindered by the simultaneous need for inter-rater reliability to legitimize measures, and limited capacity for conducting valid tests of inter-rater reliability. Additionally, even those studies that have investigated inter-rater reliability using more ecologically valid methods have produced mixed results. The methods to be used in the present study will evaluate a more practical alternative for determining assessment accuracy in field settings. Specifically, an alternative method to inter-rater reliability, an item-level intra-rater

consistency procedure will be assessed. This will be followed by an analysis of how the intra-rater consistency of the SPIn impacts predictive validity for the entire sample, and subgroups based on sex and Indigenous status. In view of evaluating the relationship between intra-rater consistency and predictive validity, the following questions were be addressed:

Research Question 1: What is the frequency and nature of SPIn coding errors, and do they vary as a function of sex or Indigenous status?

Hypothesis 1: Given the limited research on methods related to those that were employed in the current study, no specific hypotheses were made regarding differences in error rate for male and female participants, and Indigenous and non-Indigenous participants. Instead, I conducted exploratory analyses to determine whether there were any significant differences in error rate as a function of sex and/or Indigenous status.

Hypothesis 2: It is hypothesized that pairs of cross-consistency items that contain only static items will have the lowest error rate, pairs that include dynamic-checklist items will contain a moderate number of errors, and errors will occur at the highest rate for dynamic-continuum items, as these items require the rater to make a subjective evaluation of their client.

Hypothesis 3: It is hypothesized that certain errors in coding will occur more frequently than others, indicating which items are more difficult for practitioners to score accurately.

Research Question 2: Will changes in the intra-rater consistency of scored assessments impact their predictive validity based on ROC analysis? Will the relationship between the reliability of scored assessments and their predictive validity vary as a function of sex or race?

Hypothesis 4: It is hypothesized that more reliably scored assessments will demonstrate greater predictive validity in forecasting recidivism. Based on the known relationship between reliability and validity, and the preliminary results shared by David Robinson of Orbis Partners Inc. (D.

Robinson, personal communication, August 5 2020) with the YASI, it is believed that support for this relationship will be found in the present study.

Hypothesis 5: Given the limited research using methods related to those proposed in the current study, no specific hypotheses were made regarding differences in predictive validity based on sex or Indigenous status. Instead, exploratory analyses were conducted to examine any potential differences in the SPIn Pre-Screen's predictive accuracy as a function of sex and/or Indigenous status.

Method

Participants

The study used an archival dataset comprised of 31,460 participants, including 24,998 (79.5%) men and 6462 women (20.5%). Participants were provincially sentenced adults who began community supervision between 2008 and 2012 in Alberta. The Alberta Solicitor General provided SPIn Pre-Screen and Full Assessments for all participants, as well as demographic information, and information regarding failures on supervision for one- and three-year fixed follow-up periods. It should be noted that this sample overlaps with samples from other SPIn studies (i.e., Jones & Robinson, 2017a, 2018; Wanamaker, 2020).

Of the 31,460 participants in the sample, 6513 (20.7%) were Indigenous, and the remainder were non-Indigenous (24,947). Unfortunately, due to inconsistent coding practices[1], the dataset did not contain reliable information to ascertain the racial background of 22,161 of the non-Indigenous participants, thus impeding more detailed analysis of the impact of race in this sample. A greater proportion of Indigenous participants were female (n = 1966, 30.2%) than non-Indigenous participants (n = 4496, 18.0%), while a greater proportion of non-Indigenous participants were male (n = 20,451, 82.0%) as compared to Indigenous participants (n = 4547, 69.8%). A chi-square test of independence indicates that these differences were statistically significant with a small effect size (χ^2 = 468.18, φ = .12, p < .001), based on Cohen (1988). Participants ranged in age from 16 to 87, with a mean age of 33.15 (SD = 11.44); over 85% of the sample were aged 45 or younger.

[1] Four race-related variables appear in the dataset: two contain information written in by probation officers, and two were coded numerically. Only 2873 participants had clearly written information. In cross-checking between variables, it was discovered that there was some inconsistency in coding practices. For example, those identified as "Caucasian" in the write-in text were sometimes coded as 1 or 3, making those variables unreliable. Only the coding of the Indigenous/non-Indigenous variable appeared to be consistent with responses in the other three race variables.

The majority of participants were on probation or parole (*n* = 20,653; 65.6%) followed by a peace bond (*n* = 3674, 11.7%) or conditional sentence (*n* = 3136, 10.0%). A small portion of the sample were young offenders (*n* = 344, 1.1%) or on a temporary absence (*n* = 24, 0.0001%). Those on probation or parole included those who received a sole probation sentence, were sentenced to community service initially (plus a term of subsequent probation) or had been released on parole from custody to serve the remainder of their sentences in the community. Only minimal differences were observed in the proportions of participants on various kinds of supervision—all effect sizes were negligible ($\varphi < .10$). Indigenous males were most likely to be on probation or parole (75.4%), followed by Indigenous females (71.8%). Generally, non-Indigenous participants were less likely to be on probation or parole; 64% of males and 60.5% of females were on this kind of community supervision. Other types of community supervision occurred at such low rates across the sample that no statistically significant differences were observed as a function of sex or Indigenous status. A complete breakdown of community supervision orders by sex and Indigenous status can be viewed in Table 1.

Table 1

Community Supervision Orders by Indigenous Status and Sex

	Full Sample	Non-Indigenous			Indigenous		
	n/31,460 (%)	Female *n*/4496 (%)	Male *n*/20,451 (%)	χ^2 (φ)	Female *n*/1966 (%)	Male *n*/4547 (%)	χ^2 (φ)
Probation or Parole	20,653 (65.6)	2721 (60.5)	13,092 (64.0)	19.41*** (-.03)	1411 (71.8)	3429 (75.4)	9.54** (-.04)
Peace Bond	3674 (11.7)	665 (14.8)	2552 (12.5)	17.55*** (.03)	176 (9.0)	281 (6.2)	16.17*** (.05)
Conditional Sentence	3136 (10.0)	536 (11.9)	2212 (10.8)	4.60* (.01)	119 (6.1)	269 (5.9)	0.05 (.003)
Young Offender	344 (1.1)	54 (0.01)	207 (0.01)	1.18 (.01)	21 (0.01)	62 (0.01)	0.77 (-.01)
Temporary Absence	24 (0.0001)	3 (0.0007)	17 (0.0008)	0.12 (-.002)	0 (0.0)	4 (0.0009)	1.73 (-.02)
Information not available	3629 (11.5)	517 (11.5)	2371 (11.6)	0.03 (-.001)	239 (12.2)	500 (11.0)	1.70 (.02)

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

Several significant sex differences were observed in participants' index offences. However, while 12 of the chi-square tests of independence conducted were statistically significant, only five of those effects were large enough to be considered noteworthy based on Cohen's (1988) effect size thresholds. A complete summary of index offence types may be seen in Table 2. Violent index offences were the most common reason for participants to be on supervision. This is an unusual finding, as typically violent offences are less common than non-violent crimes. To put this in context, violent crimes accounted for 23.69% of all criminal charges laid in Alberta between 2009 and 2012 (Statistics Canada, 2021a), however 36.3% of participants in this sample had a violent index offence. This difference may not seem remarkable in the full sample given that a fair number of the charges laid are likely dropped for various reasons. However, when sex and Indigenous status are considered, disparities emerge. Non-Indigenous females had the lowest rate of violent index offences, at 13.3%, followed by non-Indigenous males at 36.3%; this difference was statistically significant with a small effect (χ^2 = 270.29, φ = -.10, $p < .001$). Thirty-seven percent of Indigenous females had violent index offences, and 48.4% of Indigenous males had violent index offences; males had significantly more violent index offences than females (χ^2 = 72.70, φ = -.11, $p < .001$). While both males and Indigenous people do typically commit more violent crimes than females or non-Indigenous people (e.g., Department of Justice Canada, 2019), it is remarkable to see that Indigenous females had approximately the same rate of violent index offences as non-Indigenous males.

Females were much more likely to have fraud as an index offence than males. Among non-Indigenous participants, 4.8% of females had an index offence of fraud, compared to only 2.5% of males; this difference was statistically significant with a small effect (χ^2 = 383.10, φ = .12, $p < .001$). Similarly, 4.1% of Indigenous females had committed fraud, while only 1.0% of

males had (χ^2 = 72.46, φ = .11, p < .001). The last notable difference was in the frequency of property crimes among non-Indigenous participants. Almost twice as many females had a property index offence than males (20.9% female, 11.6% male; χ^2 = 275.05, φ = .11, p < .001). A similar pattern of results was observed in Indigenous participants; however, the effect of the difference was too small to be considered important (φ = .08).

Table 2

Index Offences by Indigenous Status and Sex

	Full Sample	Non-Indigenous			Indigenous		
	n/31,460 (%)	Female *n*/4496 (%)	Male *n*/20,451 (%)	χ^2 (φ)	Female *n*/1966 (%)	Male *n*/4547 (%)	χ^2 (φ)
Administration of Justice	694 (2.2)	105 (2.3)	421 (2.1)	1.37 (.01)	68 (3.5)	100 (2.2)	8.67** (.04)
Drug	2022 (6.4)	274 (6.1)	1538 (7.5)	11.13*** (-.02)	73 (3.7)	137 (3.0)	2.16 (.02)
Driving while intoxicated	1200 (3.8)	160 (3.6)	821 (4.0)	2.03 (-.01)	59 (3.0)	160 (3.5)	1.13 (-.01)
Fraud	1009 (3.2)	379 (8.4)	505 (2.5)	383.10*** (.12)	81 (4.1)	44 (1.0)	72.46*** (.11)
Property	4143 (12.3)	941 (20.9)	2382 (11.6)	275.05*** (.11)	322 (16.4)	498 (11.0)	36.72*** (.08)
Sexual	855 (2.7)	43 (1.0)	664 (3.2)	70.21*** (-.05)	25 (1.3)	123 (2.7)	12.70*** (-.04)
Trafficking	191 (0.6)	33 (0.7)	140 (0.7)	0.13 (.00)	5 (0.3)	13 (0.3)	0.05 (.00)
Technical Violation	1212 (3.9)	160 (3.6)	714 (3.5)	0.05 (.00)	113 (5.7)	225 (4.9)	1.78 (.02)
Violent	11,418 (36.3)	1057 (23.5)	7432 (36.3)	270.29*** (-.10)	727 (37.0)	2202 (48.4)	72.70*** (-.11)
Other	3856 (12.3)	596 (13.3)	2713 (13.3)	0.00 (.00)	191 (9.7)	356 (7.8)	6.35* (.03)
Missing	4860 (15.4)	748 (16.6)	3121 (15.3)	128.63*** (.07)	302 (15.4)	689 (15.2)	10.92*** (.04)

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

Measures

The Service Planning Instrument (SPIn)—General Overview

The Service Planning Instrument (SPIn; Orbis Partners, 2003) is an interview-based, structured risk/needs/strengths assessment instrument designed to guide justice system professionals (e.g., probation, parole, and correctional officers) in making informed decisions regarding an individual's overall risk of recidivism and to direct case management decisions. The SPIn was developed following the success of the YASI, a similar tool developed by Orbis Partners Inc., for use with justice-involved youth (Orbis Partners Inc., 2000). The SPIn is considered a gender-informed tool, as it contains a combination of both gender-neutral and gender-responsive items and may be used with justice-involved men and women. Gender-responsive items within the SPIn, such as those exploring the individual's relationship with their children, are assumed to be more relevant in assessing and providing treatment to justice-involved women than they are in providing those services to justice-involved men. Professionals who use the SPIn must complete four days of training prior, two days learning to use the assessment itself, and two days dedicated to applying the results to case planning and management. Alberta Justice Department policy dictates that initial assessments must be completed within 45 days of the start of community supervision or following release from prison.

The SPIn is comprised of 90 items in total, dispersed across 11 content domains. Nine of the eleven domains capture a combination of static and dynamic risk factors, as well as static and dynamic strength factors. The two exceptions are the criminal history and response to supervision domains, which contain only static risk items. The SPIn's 11 domains include criminal history, response to supervision, aggression/violence, substance use, social influences,

family, employment, attitudes, social/cognitive skills, stability, and mental health. A subset of 35 questions contained in the full SPIn assessment are used as a Pre-Screen assessment, which allows professionals to triage or classify their clients and provide a more accurate estimation of risk. Orbis recommends that those who are found to be moderate to high risk have the Full Assessment completed, to provide more information to guide case management decisions. A comprehensive summary including all items found in the SPIn Pre-Screen and Full Assessment versions may be found in Appendix B.

SPIn Item Rating Formats. Items within the SPIn fit five different formats. Some static risk items relating to previous justice-involvement are scored based on the number of previous occurrences (e.g., previous adult adjudications: 0, 1, 2, 3+), while others include a checklist of options (e.g., variety of offences: assault/violence, robbery, fraud, etc.), or simply require a yes or no response (e.g., any failures to return from temporary releases). Some dynamic items are also scored based on a checklist (e.g., marital risk factors), which can include a combination of risk and strength sub-items. Finally, many dynamic items are scored on a five-point poled scale where responses can range from high risk to high strength. An example of the response format for dynamic-continuum items may be found in Figure 1.

Figure 1

Example Dynamic-Continuum Item with Response Range

	● -- --	● --	● 0	● +	● + +
1. Marital relationship:	High degree of instability and conflict, offender expresses high dissatisfaction	Some conflict and dissatisfaction evident in the relationship	Minimal satisfaction in relationship (or no current marital relationship)	Stability of relationship evident, offender expresses satisfaction	High degree of stability, satisfaction and commitment to the relationship

Note. Responses may indicate that the item represents a high risk ("- -"), a moderate risk ("-"), neutral (0), a moderate strength ("+"), or a high strength ("+ +").

Weighting and Scoring the Pre-Screen SPIn. The Pre-Screen SPIn is comprised of 35 items which includes a combination of static and dynamic factors. All 35 items contribute to a total pre-screen risk score, and a subset of 11 items are used to produce a total pre-screen protective score. It is important to note that many SPIn items can contribute to either the total risk or strength scores, depending on the response selected. For example, the item displayed in Figure 1 can increase the total risk score or the total protective score. These scores are then weighted using a version of the Nuffield method (Nuffield, 1982). The Pre-Screen risk score is most commonly used for classifying individuals as low, medium, or high risk, and low, medium, or high strength. Pre-Screen risk scores can range from 0 to 103, and Pre-Screen strength scores can range from 0 to 21. A comprehensive summary of score ranges by domain may be found in Table 3.

Table 3

SPIn Pre-Screen Scoring Summary

Domain	Content	Score range
Criminal History	6 static risk items	0-20
Response to Supervision	6 static risk items	0-9
Aggression/Violence	5 static risk items	0-7
Substance Use	1 static risk item	0-4
	1 dynamic risk item	0-23
Social Influences	2 dynamic risk items	0-7
	1 dynamic strength item	0-2
Family	3 dynamic risk items	0-9
	3 dynamic strength items	0-5
Employment	1 static risk item	0-5
	2 dynamic risk items	0-4
	1 dynamic strength item	0-2
Attitudes	2 dynamic risk items	0-4
	2 dynamic strength items	0-4
Social/Cognitive Skills	2 dynamic risk items	0-4
	2 dynamic strength items	0-4
Stability	2 dynamic risk items	0-7
	2 dynamic strength items	0-4
Mental Health	4 dynamic items*	N/A

Note. * Items in the mental health domain do not contribute to final risk or strength scores, but rather are included to aid case managers in the process of gathering information.

Weighting and Scoring the SPIn Full Assessment. The SPIn Full Assessment is comprised of 90 items, including all items found in the Pre-Screen SPIn. Full Assessment risk scores can range from 0 to 250 and strength scores range from 0 to 91. Scores in the Full Assessment may be further broken down by static or dynamic risk or strength in each domain. For a summary of domain risk and strength scores, see Table 4.

Table 4

SPIn Full Assessment Scoring Summary

Domain	Content	Score range
Criminal History	6 static risk items	0-20 (6 items)
Response to Supervision	10 static risk items	0-43 (10 items)
Aggression/Violence	10 static risk items	0-41 (10 items)
	4 dynamic risk items	0-8 (4 items)
	4 dynamic strength items	0-8 (4 items)
Substance Use	1 static risk item	0-9 (1 item)
	3 dynamic risk items	0-28 (3 items)
Social Influences	6 dynamic risk items	0-26 (6 items)
	5 dynamic strength items	0-15 (5 items)
Family	3 static risk items	0-9 (3 items)
	7 dynamic risk items	0-26 (7 items)
	2 static strength items	0-4 (2 items)
	7 dynamic strength items	0-14 (7 items)
Employment	2 static risk items	0-9 (2 items)
	6 dynamic risk items	0-14 (6 items)
	1 static strength item	0-3 (1 item)
	5 dynamic strength items	0-12 5 items)
Attitudes	9 dynamic risk items	0-14 (9 items)
	9 dynamic strength items	0-14 (9 items)
Social/Cognitive Skills	8 dynamic risk items	0-18 (8 items)
	8 dynamic strength items	0-18 (8 items)
Stability	1 static risk item	0-2 (1 item)
	4 dynamic risk items	0-13 (4 items)
	4 dynamic strength items	0-7 (4 items)
Mental Health	Flag	0-2 (5 items)

SPIn Classification Thresholds. These risk and strength scores can be used to classify individuals into low, medium, and high risk and strength categories. Classification thresholds vary based on sex and are adjusted in each jurisdiction based on norms of that area (Jones & Robinson, 2017a). Thresholds used in Alberta were based on an unpublished initial 6-month

follow-up recidivism study conducted by Jones and Robinson shortly after the province adopted the SPIn.

SPIn Empirical Evidence. As described in detail in the introduction, while only four studies thus far have evaluated the predictive accuracy of the SPIn (Jones et al., 2015; Jones & Robinson, 2017a, 2017b, 2018), the results of these studies indicate that the SPIn has moderate to strong predictive accuracy. AUC values range from .59 (Jones & Robinson, 2017a) to .77 (Jones et al., 2015). Additionally, these studies have demonstrated that the SPIn is an effective classification tool, with recidivism rates lowest among those classified as low-risk (range 8.6% to 22.9%), moderate among those classified as medium risk (range 25.9% to 49.3%) and highest among those classified as high risk (range 54.4% to 78.8%) (Jones et al., 2015; Jones & Robinson, 2017a; Jones & Robinson, 2017b; Jones & Robinson, 2018). These differences were found to be statistically significant in all studies. At this time, the SPIn's inter-rater reliability has not been assessed.

Failure Outcomes

The measurement of recidivism, or failure outcomes, was based on one- and three-year fixed follow-up periods, starting at the time the initial SPIn assessment was completed, which occurred within 45 days of the start of community supervision. Individuals began community supervision either following release from prison (i.e., parole), when starting probation, or when starting a conditional sentence. Six unique dichotomous measures of recidivism (yes/no) have been provided by the Alberta Solicitor General, each representing different reasons for renewed contact with correctional services. These outcomes are described below:

(1) **Any recidivism:** any contact with correctional services as a result of any infraction, including technical violations (e.g., violation of probation conditions), and any new charges (e.g., assault, theft).

(2) **Any technical violation(s):** includes failures to comply with conditions (e.g., failure to return home prior to curfew), and failures to appear in court.

(3) **Any new charge(s):** excludes technical violations but includes all other charges for new offences; the participant may not have been convicted of these charges.

(4) **Any return to custody:** includes returns to custody due to technical violations, new charges, and new convictions.

(5) **Any new conviction:** includes any convictions for new criminal charges.

(6) **Any new violent charge(s):** excludes technical violations, non-violent charges, and sexual charges (e.g., assault, robbery); the participant may not have been convicted of these charges.

Procedure

Acquisition of Data

The current study used an archival dataset provided by the Alberta Solicitor General via Orbis Partners Inc. Orbis is a Canadian consulting firm which develops and provides assessment tools, correctional programming, and training to government and community agencies working with justice-involved adults and youth. All adults included in the sample were assessed by probation and parole officers in Alberta, who had completed four days of training in the use of the SPIn. Assessments were coded based on a semi-structured interview, and additional information gathered from official sources (i.e., police and institutional records) and other

credible sources when appropriate (e.g., family members, social workers). Participants understood that their information was to be used to assess and manage their risk of recidivism.

An application was submitted to the Alberta Solicitor General on November 15, 2019, requesting access to the aforementioned dataset for the purposes of the present study. The Alberta Solicitor General granted tentative approval on February 26, 2019, pending clearance from the Carleton University Research Ethics Board (CUREB). CUREB clearance has since been granted; the ethics certificate may be found in Appendix C. Final approval from the Alberta Solicitor General was received along with the dataset on July 6, 2020; the decision letter from the Alberta Solicitor General may be found in Appendix D.

All data is now maintained on secure servers hosted by the Gender and Crime Lab at Carleton University and Orbis Partners. All personally identifying information was removed prior to receipt, and additional password protection on data files was used to ensure the safety and privacy of participants. Only the author, Dr. Brown, and Dr. Robinson (CEO Orbis Partners) have access to the data.

Identification of Cross-Consistency Items

A number of items exist within the SPIn which assess the same underlying construct; these items form the basis for the present study. For example, there are multiple items in the SPIn that assess related constructs, such as an individual's employment and life skills, history of criminal and anti-social behaviours, and experiences of trauma and abuse. The idea to use these related items (or cross-consistency items) to identify potential rater errors was developed by Dr. Jones and Dr. Robinson of Orbis Partners in the context of the YASI (D. Robinson, personal communication, November 15, 2019). Their preliminary results using cross-consistency items in the YASI to identify errors are promising (Robinson, 2019). Dr. Robinson provided a brief

explanation of the types of item relationships they identified within the YASI; with this understanding, I sought to identify similar relationships in the SPIn.

Following a thorough examination of the SPIn, including completing online training offered by Orbis Partners in the use of the SPIn, I identified logical relationships between individual items within the assessment. These relationships can be expressed in simple conditional statements; for example, if the participant has one or more delinquency adjudications (item A04), then their age at first arrest must be 17 or younger (item A01). Such conditional statements can be translated into Boolean expressions. Boolean expressions must be either true or false and form the basis of many programming languages (Genesereth et al., 2020; Steinbach & Posthoff, 2013).

After compiling a complete list of inter-item relationships, I compared these to the code used by Dr. Jones and Dr. Robinson to identify errors in the YASI. While the SPIn and the YASI are not identical, they share broad similarities and many analogous items. Thus, the YASI code was a useful source for comparison, and allowed me to make several adjustments to the list of inter-item SPIn relationships prior to proceeding. Next, Dr. Brown, Dr. Robinson, and I had several meetings in which each of these inter-item SPIn relationships were reviewed and discussed. Each pair of items was either retained, altered, or removed based on consensus. During this process, it was determined that there were two kinds of cross-consistency relationships between items. First, there were items which, in all conceivable situations, would necessitate a certain answer on another item. Using the same example as earlier, if a participant was indicated to have delinquency adjudications (item A04), then their age at first arrest must be 17 or younger (A01). These inter-item relationships may be considered “firm.” Other pairs of items had relationships that were less certain. In other words, certain responses on some items

should, in almost all situations, necessitate a specific answer on another item. However, it was determined that while unlikely, some pairings of answers were not completely impossible. For example, if an individual's employment plans (item G4) are rated as a strength, in almost all cases the same individual's employment motivation (G2) should also be rated as a strength. However, it is not impossible for an individual to have good employment plans while simultaneously lacking motivation to follow through on these plans. We described these pairs of items as *grey*, in that their responses couldn't be viewed with the same certainty as the *firm* items.

In total, 32 logical statements were retained: 13 were based on items in the Pre-Screen and 19 more found only in the Full Assessment. All 13 Pre-Screen errors were firm. Of the additional 19 Full Assessment items, 10 were classified as firm and the other 9 were grey. Thus, the plausible range of errors in the Pre-Screen was from 0 to 13. The plausible range of errors for the Full Assessment was from 0 to 32 when both firm and grey items were included, and from 0 to 23 when only firm items were included.

These 32 errors were further classified into three groups, based on the types of SPIn items included. Ten items were static errors, meaning that all SPIn items that they were based on were static SPIn items; these were all firm errors. Six were dynamic-checklist items, also all firm errors, and were scored based on at least one dynamic SPIn item with a checklist of options. Finally, 16 items were based on at least one dynamic-continuum SPIn item. Of these dynamic-continuum errors, seven were firm and nine were grey. A plain language explanation of each error item, as well as the type of error, may be viewed in Table 5. A more technical presentation of the logical statements used to develop the variable code may be found in Appendix E.

Table 5

Overview of SPIn Inter-Item Relationships

Error variable	Type	Explanation
Pre-Screen		
1	Static (firm)	If the participant has had one or more delinquency adjudications (A04), their age at first arrest must be 17 or younger (A01)
2	Static (firm)	If the participant has been indicated to have a history of violent convictions in one item, then equivalent evidence must be seen in the other item (A06 and C02)
3	Static (firm)	If the participant has a history of delinquency indicated in one item, equivalent evidence must be seen in the other item(s) (A01, A04, A05)
4	Static (firm)	If the participant has no delinquency adjudications (A04) they should have no incarcerations as a delinquent (A05)
5	Static (firm)	If the participant has been transferred to custody while on supervision (B04) then corresponding evidence of misbehaviour while on supervision must be indicated (B02, B03, and/or B05)
6	Dynamic-Checklist (firm)	If the participant associates with gang members or is a gang member (E01) then corresponding evidence of anti-social influences must be indicated (E02)
7	Dynamic-Continuum (firm)	If the participant has perpetrated domestic violence or expressed safety issues with their partner (F02) then their marital relationship must be rated as a risk (F01)
8	Dynamic-Checklist (firm)	If the participant has been indicated to be a perpetrator of domestic violence in one item, corresponding evidence should be indicated in the other item (F02 and C04)*
9	Static (firm)	If the participant has been indicated to be a perpetrator of domestic violence in one item, corresponding evidence should be indicated in the other item (F02 and C04)*
10	Dynamic-Checklist (firm)	If the participant must rely on social assistance (J01a) then they must also be indicated as having a low income for their household needs (J01b)
11	Static (firm)	If the participant has been indicated to have a history of sex offences (A06) then they must also be

Error variable	Type	Explanation
		indicated to have a history of sexual aggression (K04)
12	Dynamic-Continuum (firm)	If the participant belongs to a gang (E01) then their law-abiding attitudes must be rated as a risk (H01)
13	Dynamic-Continuum (firm)	If the participant expresses safety issues with their spouse (F02) their marital relationship must be rated as a risk (F01)
Full Assessment		
1	Static (firm)	If the participant has ever been placed in segregation for disciplinary reasons (B09) then they must have corresponding evidence of institutional misconduct (B07 and/or B08)
2	Dynamic-Checklist (firm)	If the participant belongs to a gang or associates with gang members (E01) then they must be indicated as having associates at risk for frequent or serious offending (E03 and/or E04)
3	Dynamic-Checklist (firm)	If the participant is involved with pro-social community organizations (E05) then they must also be involved in pro-social activities (E04)
4	Dynamic-Checklist (firm)	If the participant is involved with pro-social community organizations (E05) then they must associate with people who have a pro-social influence (E02)
5	Dynamic-Continuum (firm)	If the participant's parenting skills are rated as a risk (F04) then their attachment to their children must also be rated as a risk (F03)
6	Static (firm)	If the participant has experienced trauma or victimization as a child (K05) then they must have corresponding evidence of that trauma indicated with their family of origin (F06)
7	Static (firm)	If the participant is indicated to have a history of victimization or abuse in one item, corresponding evidence of that abuse must be seen in the other item (K05 and F06)
8	Dynamic-Continuum (firm)	If the participant's job search skills are rated as a risk (G07) then their employment plans must also be rated as a risk (G04)
9	Dynamic-Continuum (firm)	If the participant has post-secondary training or a higher level of education (G03) then their marketability should be rated as a strength (G06)

Error variable	Type	Explanation
10	Dynamic-Continuum (firm)	If the participant belongs to a gang (E01) then their attitudes towards the criminal justice system must be rated as a risk (H05)
11	Dynamic-Continuum (grey)	If the participant's parenting skills are rated as a strength (F04) their attachment to their children should also be rated as a strength (F03)
12	Dynamic-Continuum (grey)	If the participant's employment plans are rated as a strength (G04) their employment motivation should also be rated as a strength (G02)
13	Dynamic-Continuum (grey)	If the participant has literacy issues and/or low educational attainment (G03) their marketability should also be rated as a risk (G06)
14	Dynamic-Continuum (grey)	If the participant's willingness to accept responsibility is rated as a strength (H02) their ability to understand the impact of their behaviour should also be rated as a strength (H06)
15	Dynamic-Continuum (grey)	Participants should have matching scores (i.e., both risk or both strength) in items assessing the participant's impulsivity (I01) and consequential thinking (I03)
16	Dynamic-Continuum (grey)	Participants should have matching scores (i.e., both risk or both strength) in items assessing the participant's problem-solving skills (I05) and consequential thinking (I03)
17	Dynamic-Continuum (grey)	Participants should have matching scores (i.e., both risk or both strength) in items assessing the participant's hostile attributions (I02) and social perspective-taking (I04)
18	Dynamic-Continuum (grey)	Participants should have matching scores (i.e., both risk or both strength) in items assessing the participant's hostile attributions (I02) and interpersonal skills (I07)
19	Dynamic-Continuum (grey)	Participants should have matching scores (i.e., both risk or both strength) in items assessing the participant's social perspective-taking (I04) and interpersonal skills (I07)

Note. * Pre-Screen Errors 8 and 9 are based on the same SPIn items but were separated into two unique error items to capture errors made with static and dynamic sub-items separately.

Analytic Approach

Analyses were performed using Statistical Software Package for the Social Sciences (SPSS) version 27 (IBM Corp, 2020). Preliminary analyses included data cleaning to identify

potential errors in data entry and missing data. Descriptive statistics followed, which included demographic characteristics of the sample; and means, standard deviations, and ranges of SPIn Pre-Screen and Full Assessment scores.

Results

Data Cleaning

Data were screened for missing values, outliers, normality, and linearity. The dataset provided by the Alberta Solicitor General included a total of 31,477 completed SPIn Pre-Screen assessments, 20,532 of whom had a completed SPIn Full Assessment. A total of 17 cases were deleted due to missing data on key variables, including those which were missing recidivism data (n = 8) or information regarding Indigenous status (n = 9). Further screening was conducted for variables created as part of analyses; issues and transformations for these variables are discussed in subsequent sections.

Descriptive Statistics

Classification

The results of the SPIn Pre-Screen were used to classify participants as low risk, medium risk, or high risk, and low strength, medium strength, or high strength. Table 6 provides a complete overview of Pre-Screen classification results by both sex and Indigenous status. Most participants were classified as low risk (n = 18,201, 57.9%) and medium strength (n = 14,511, 46.1%), however differences emerged when classification results were disaggregated by sex and Indigenous status. Among non-Indigenous participants females were more likely to be classified as low risk, while males were more likely to be classified as high risk. The effect of this difference, however, was negligible and below the threshold to be considered important (χ^2 = 177.02, φ = .08, p < .001; Cohen, 1988). Similarly, non-Indigenous females were more likely to be classified as high strength while non-Indigenous males were more likely to be low strength, however the effect of this difference was again negligible (χ^2 = 61.56, φ = .05, p < .001). An almost identical pattern was observed among Indigenous participants; males were more likely to

be high risk and females were more likely to be low risk, with a small effect size (χ^2 = 132.64, φ = .14, p < .001). There were no significant differences in the strength classification of Indigenous participants (χ^2 = 3.25, φ = .02, p = .197). Among males, Indigenous participants were significantly more likely to be high risk, while non-Indigenous participants were more likely to be low risk, which represented a small effect size (χ^2 = 1411.38, φ = .24, p < .001). Indigenous males were also more likely to be classified as low strength, and non-Indigenous males were more likely to be classified as high strength, with a small effect size (χ^2 = 351.80, φ = .12, p < .001). Again, a similar pattern was repeated among female participants: Indigenous participants were more likely to be classified as high risk, and non-Indigenous participants were more likely to be classified as low risk, with a small effect (χ^2 = 394.18, φ = .25, p < .001). Indigenous females were also more likely to be classified as high risk, and non-Indigenous females were more likely to be classified as low risk, again with a small effect (χ^2 = 232.79, φ = .19, p < .001). These results are in keeping with previous SPIn research which has demonstrated that Indigenous people are typically classified as higher risk and lower strength than their non-Indigenous counterparts (e.g., Jones et al., 2015; Perley-Robertson et al., 2019). Women are also typically found to be lower risk (e.g., Coid et al., 2009; Jones et al., 2015). The few studies that have evaluated the relationship between sex, Indigenous status, and strength classification have inconsistent results, and thus it is less clear whether the results observed here can be considered normal (e.g., Jones et al., 2015; Viljoen et al., 2016).

Table 6

SPIn Pre-Screen Risk and Strength Classification by Sex and Indigenous Status.

	Total	Non-Indigenous		Indigenous	
		Female	Male	Female	Male
	n/31,460 (%)	*n*/4496 (%)	*n*/20,451 (%)	*n*/1966 (%)	*n*/4547 (%)
Pre-Screen Risk					
Low	18,201 (57.9)	3211 (71.4)	12,565 (61.4)	916 (46.6)	1509 (33.2)
Medium	10,702 (34.0)	1133 (25.2)	6527 (31.9)	848 (43.1)	2194 (48.3)
High	2557 (8.1)	152 (3.4)	1359 (6.6)	202 (10.3)	844 (18.6)
Pre-Screen Strength					
Low	9491 (30.2)	1117 (24.8)	5748 (28.1)	771 (39.2)	1855 (40.8)
Medium	14,511 (46.1)	2009 (44.7)	9600 (46.9)	909 (46.2)	1993 (43.8)
High	7458 (23.7)	1370 (30.5)	5103 (25.0)	286 (14.5)	699 (15.4)

Recidivism

Six different measures of recidivism were included in analyses for each of the one- and three-year follow-up periods: any recidivism, technical violations, any new offences, any return to custody, any new convictions, and any violent recidivism. Recidivism rates for the overall sample as well as by sex and Indigenous status may be viewed in Table 7.

Generally, recidivism results were as expected. Indigenous participants had higher rates of all kinds of recidivism, and significant sex-based differences in recidivism rate were observed among both Indigenous and non-Indigenous participants. Men had significantly higher rates of recidivism across all types of recidivism and follow-up periods, however in most cases the size of these effects were negligible, with a few exceptions for Indigenous participants. Small effect sizes were observed, where the rate of recidivism was higher for males than females, for any new offence at both one- and three-year follow-up, and for violent recidivism at both one- and three-year follow-up.

Table 7

Recidivism in One-Year Follow-Up Sample by Sex and Indigenous Status

		Non-Indigenous			Indigenous		
	Total (N = 31,460)	Female n/4496 (%)	Male n/20,451 (%)	χ^2 (φ)	Female n/1966 (%)	Male n/4547 (%)	χ^2 (φ)
Any Recidivism							
1 year	6346 (20.2)	664 (14.8)	3638 (17.8)	23.56*** (-.03)	515 (26.2)	1529 (33.6)	35.20*** (-.07)
3 year	9540 (30.3)	1014 (22.6)	5594 (27.4)	43.61*** (-.04)	764 (38.9)	2168 (47.7)	43.13*** (-.08)
Technical Violations							
1 year	4366 (13.9)	402 (8.9)	2418 (11.8)	30.54*** (-.04)	376 (19.1)	1170 (25.7)	33.09*** (-.07)
3 year	5981 (19.0)	562 (12.5)	3382 (16.5)	45.13*** (-.04)	520 (26.4)	1517 (33.4)	30.52*** (-.07)
Any New Offence							
1 year	4904 (15.6)	506 (11.3)	2881 (14.1)	25.21*** (-.03)	367 (18.7)	1150 (25.3)	33.71*** (-.07)
3 year	7518 (23.9)	776 (17.3)	4506 (22.0)	50.32*** (-.05)	554 (28.2)	1682 (37.0)	47.28*** (-.09)
Any Return to Custody							
1 year	4496 (14.3)	351 (7.8)	2535 (12.4)	75.86*** (-.06)	337 (17.1)	1273 (28.0)	86.91*** (-.12)
3 year	6384 (20.3)	500 (11.1)	3671 (18.0)	123.45*** (-.07)	479 (24.4)	1734 (38.1)	116.03*** (-.13)
Any New Conviction							
1 year	3525 (11.2)	419 (9.3)	2019 (9.9)	1.28 (-.01)	312 (15.9)	775 (17.0)	1.36 (-.01)
3 year	5876 (18.7)	677 (15.1)	3463 (16.9)	9.41** (-.02)	504 (25.6)	1231 (27.1)	1.45 (-.02)
Violent Recidivism							
1 year	2433 (7.7)	154 (3.4)	1419 (6.9)	77.01*** (-.06)	163 (8.3)	697 (15.2)	59.32*** (-.10)
3 year	3768 (12.0)	245 (5.4)	2253 (11.0)	126.78*** (-.07)	249 (12.7)	1021 (22.5)	83.79*** (-.11)

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

Main Analyses

Research Question 1: What is the frequency and nature of SPIn coding errors, and do they vary as a function of sex or race?

Differences in Error Totals by Sex and Race. To address the first research question and evaluate whether some errors occur more frequently than others, errors in SPIn scoring had to first be identified. Using SPSS Syntax, a code was developed to flag these potential errors. A new variable was created for each of the inter-item cross-consistency relationships, which indicated whether an error occurred in scoring that pair of items (1 = error, 0 = no error). In total 32 error variables were created: 13 firm Pre-Screen Assessment errors, and 19 additional errors from the Full Assessment, of which 10 were firm and 9 were grey. Of the 32 total errors, 10 were based on static items only (all of which were firm), 6 were based on dynamic-checklist items (all firm), and 16 were based on dynamic-continuum items, of which 7 were firm and 9 were grey.

Seven different error totals were calculated: Pre-Screen, Full Assessment (firm and grey), Full Assessment (firm only), Static (firm only, includes items from the Pre-Screen and Full Assessment), Dynamic-Checklist (firm only, includes items from the Pre-Screen and Full Assessment), Dynamic-Continuum (firm and grey, all from the Full Assessment), and Dynamic-Continuum (firm only, all from the Full Assessment). Raw mean total scores in the full sample, as well as by sex and Indigenous status, may be seen in Table 8. Overall, the base rate of errors was low. The mean error total for the Pre-Screen was consistently below 1, and the mean total for the Full Assessment was consistently below 2. In many cases, Indigenous participants appeared to have greater mean errors than non-Indigenous participants. Additionally, female participants appeared to have a greater number of mean errors than males. All error totals were positively skewed; most participants had either zero or one error. Histograms depicting the

distribution of each error total, both in the full sample and in each sub-group can be seen in Appendix F.

Table 8

SPIn Raw Mean Error Totals for the Full Sample and by Sex and Indigenous Status

Error Type	Plausible range	Full Sample $N = 31,460$	Non-Indigenous		Indigenous	
			Female $n = 4496$	Male $n = 20,451$	Female $n = 1966$	Male $n = 4547$
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Pre-Screen	0-13	0.63 (0.80)	0.57 (0.78)	0.60 (0.78)	0.76 (0.84)	0.75 (0.83)
		Full Sample* $N = 20,523$	Non-Indigenous		Indigenous	
			Female $n = 2478$	Male $n = 12,941$	Female $n = 1419$	Male $n = 3685$
Full Assessment (firm and grey)	0-32	1.49 (1.24)	1.49 (1.28)	1.40 (1.20)	1.78 (1.33)	1.69 (1.29)
Full Assessment (firm only)	0-23	1.20 (1.09)	1.27 (1.15)	1.10 (1.04)	1.55 (1.20)	1.38 (1.13)
Grey only	0-9	0.19 (0.49)	0.22 (0.52)	0.29 (0.59)	0.23 (0.53)	0.31 (0.59)
Static (firm only)	0-10	0.63 (0.77)	0.68 (0.83)	0.60 (0.75)	0.75 (0.85)	0.68 (0.75)
Dynamic-Checklist (firm only)	0-6	0.43 (0.61)	0.43 (0.61)	0.39 (0.57)	0.62 (0.69)	0.51 (0.67)
Dynamic-Continuum (firm and grey)	0-16	0.42 (0.69)	0.38 (0.66)	0.41 (0.68)	0.41 (0.69)	0.50 (0.74)
Dynamic-Continuum (firm only)	0-9	0.14 (0.37)	0.16 (0.40)	0.11 (0.34)	0.18 (0.42)	0.19 (0.43)

Note. * Sample size is reduced for Full Assessments and error totals that include Full Assessment items.

Prior to proceeding with analysis, the skew observed in error totals was addressed. Truncated versions of the error totals were calculated, wherein higher scores with low frequencies (< 2%) were collapsed. Error distributions remained skewed after totals were

transformed, however, Analysis of Variance (ANOVA) is known to be robust to violations of the assumption of normality, particularly when sample sizes are large (Gonzalez, 2009). These truncated versions of error totals were used as the dependent variables in subsequent analyses.

To determine whether the differences observed in error totals based on sex and Indigenous status were significant, a series of seven 2 (sex: male, female) by 2 (status: Indigenous, non-Indigenous) factorial ANOVAs were run, using each of the truncated total error variables as dependent variables. A summary of these results may be seen in Table 9. Virtually all omnibus and main effect results were statistically significant, though many effect sizes were below the threshold to be considered meaningful (i.e., $\eta^2 < .01$; Cohen, 1988). The omnibus results of the Pre-Screen error total, Full Assessment error total (firm and grey, and firm alone), dynamic-checklist error total, and the dynamic-continuum error total (firm only) were associated with small effects ($\eta^2 = .01$ or .02).

The main effects of sex were statistically significant for all error totals, with the exception of the Pre-Screen error total. However, the size of these effects were too small to be considered important ($\eta^2 < .01$). Indigenous status was also a statistically significant predictor for all error totals, though again, several of these effects were negligible. Small but meaningful effect sizes were observed for Pre-Screen errors, Full Assessment errors (firm and grey, and firm only), and dynamic-checklist errors. Finally, most of the interactions between sex and Indigenous status were not significant. Statistically significant results, albeit with negligible effect sizes, were only observed for dynamic-checklist errors, and dynamic-continuum errors (both with and without grey items).

Table 9

Analysis of Variance Results for Truncated SPIn Error Totals

Error Variable	Truncated range	*M* (*SD*)	*F* ratio	*df*	η^2
Pre-Screen	0-3+	0.63 (0.79)	66.20***	3	.01
Sex			0.63	1	.00
Indigenous status			175.16***	1	.01
Sex x Indigenous status			3.34	1	.00
Full Assessment (firm and grey)	0-5+	1.48 (1.22)	81.91***	3	.01
Sex			15.17***	1	.00
Indigenous status			151.17***	1	.01
Sex x Indigenous status			0.00	1	.00
Full Assessment (firm only)	0-4+	1.19 (1.06)	120.71***	3	.02
Sex			65.17***	1	.00
Indigenous status			183.55***	1	.01
Sex x Indigenous status			0.03	1	.00
Static	0-3+	0.63 (0.76)	26.73***	3	.00
Sex			25.56***	1	.00
Indigenous status			29.29***	1	.00
Sex x Indigenous status			0.01	1	.00
Dynamic-Checklist	0-2+	0.43 (0.58)	83.07***	3	.01
Sex			40.32***	1	.00
Indigenous status			166.50***	1	.01
Sex x Indigenous status			6.43*	1	.00
Dynamic-Continuum (firm and grey)	0-3+	0.42 (0.67)	19.40***	3	.00
Sex			18.09***	1	.00
Indigenous status			21.09***	1	.00
Sex x Indigenous status			4.75*	1	.00
Dynamic-Continuum (firm only)	0-2+	0.14 (0.37)	48.62***	3	.01
Sex			6.80**	1	.00
Indigenous status			44.77***	1	.00
Sex x Indigenous status			13.78***	1	.00

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

Next, I addressed Hypothesis 2, which stated that among pairs of cross-consistency items, static items would have the lowest rate of error, followed by dynamic-checklist items, and dynamic-continuum items would have the highest rate of error. As the plausible range of these total scores vary significantly, further transformation was required before the mean totals could be compared for different error types. The planned *t*-tests would not have provided valid results, had the raw scores been used; means can only be compared in this manner if they are based in the same range. Total scores for static errors, dynamic-checklist errors, and dynamic-continuum errors (with and without grey items) were transformed using the following formula to create a standardized range: $Y = \left(\frac{x - x_{min}}{x_{range}}\right) n$, where *x* is the raw error total score and *n* is the new maximum score. The new range for all mean scores was set to 0-10. These standardized mean scores may be seen in Table 10.

Table 10

Standardized Mean Error Total Scores

Error Type	Full Sample *N* = 20,523
	M (*SE*)
Static	1.27 (1.53)
Dynamic-Checklist	1.08 (1.52)
Dynamic-Continuum (firm and grey)	0.84 (1.38)
Dynamic-Continuum (firm only)	0.46 (1.23)

The standardized mean scores were then compared in the full sample with a series of within subjects *t* tests. Counter to what was hypothesized, the mean total of static errors (*M* = 1.27) was significantly greater than the mean total of dynamic-checklist errors (*M* = 1.08; $t(20{,}522) = 118.15$, Cohen's $d = .83$, $p < .001$) and dynamic-continuum errors (*M* = 0.84;

$t(20,522) = 87.35$, Cohen's $d = .60$, $p < .001$); both of these differences were medium effects (Cohen, 1988). Finally, dynamic-checklist errors ($M = 1.08$) were compared with dynamic-continuum errors ($M = 0.84$). The results indicate that the mean number of errors was significantly higher in dynamic-checklist errors, with a medium effect as well ($t(20,522) = 102.17$, Cohen's $d = .71$, $p < .001$). This pattern of results is the opposite of what was hypothesized.

Item-Level Differences in Error Rate. To follow-up the examination of overall error rates, Hypothesis 3, which stated that certain errors would occur more frequently than others was addressed. It should be noted that no specific hypotheses were made regarding which individual items would have elevated error rates.The rate of error for each of the 32 individual error variable was explored, as well as differences in error rate based on sex and Indigenous status.

Table 11 presents item-level results for the full sample. As previously mentioned, the base rate of error was quite low. The error rate was exceedingly low for the majority of individual error items (i.e., less than 2% in most cases). However, the following six errors occurred 10 to 23% of the time: (1) Pre-Screen Error 2 (static), which was based on the consistency of scoring on items regarding violent offences; (2) Pre-Screen Error 9 (static), based on the consistency of scoring between items regarding domestic violence perpetration; (3) Pre-Screen Error 10 (dynamic-checklist), based on the consistency of scoring between items regarding the participant's income; (4) Full Assessment Error 3 (dynamic-checklist), based on the consistency of scoring between items regarding pro-social community involvement; (5) Full Assessment Error 7 (static), based on the consistency of scoring in items about the participant's experiences of abuse; and (6) Grey Error 3 (dynamic-continuum), based on the consistency of scoring between items regarding the participant's education and employment marketability.

Table 11

The Rate of SPIn Rater Assessment Errors Rate in Individual Inter-Item Consistency Variables

Error Variable	Type	*n*/31,460 (%)
Pre-Screen		
1	Static (firm)	412 (1.3)
2	Static (firm)	5871 (18.7)
3	Static (firm)	370 (1.2)
4	Static (firm)	73 (0.2)
5	Static (firm)	67 (0.2)
6	Dynamic-Checklist (firm)	183 (0.6)
7	Dynamic-Continuum (firm)	2233 (7.1)
8	Dynamic-Checklist (firm)	353 (1.1)
9	Static (firm)	6425 (20.4)
10	Dynamic-Checklist (firm)	3174 (10.1)
11	Static (firm)	455 (1.4)
12	Dynamic-Continuum (firm)	107 (0.3)
13	Dynamic-Continuum (firm)	113 (0.4)
Full Assessment *	Type	*n*/20,523 (%)
1	Static (firm)	172 (0.8)
2	Dynamic-Checklist (firm)	616 (3.0)
3	Dynamic-Checklist (firm)	4846 (23.6)
4	Dynamic-Checklist (firm)	489 (2.4)
5	Dynamic-Continuum (firm)	954 (4.6)
6	Static (firm)	1830 (8.9)
7	Static (firm)	2111 (10.3)
8	Dynamic-Continuum (firm)	123 (0.6)
9	Dynamic-Continuum (firm)	197 (1.0)
10	Dynamic-Continuum (firm)	51 (0.2)
11	Dynamic-Continuum (grey)	119 (0.6)
12	Dynamic-Continuum (grey)	59 (0.3)
13	Dynamic-Continuum (grey)	2569 (12.5)
14	Dynamic-Continuum (grey)	341 (1.7)
15	Dynamic-Continuum (grey)	634 (3.1)
16	Dynamic-Continuum (grey)	400 (1.9)
17	Dynamic-Continuum (grey)	291 (1.4)
18	Dynamic-Continuum (grey)	696 (3.4)
19	Dynamic-Continuum (grey)	695 (3.4)

Note. * Sample size is reduced for items in the Full Assessment

Sex differences in error rate were only noteworthy in three error items, in which female assessments had a greater rate of error than male assessments: Pre-Screen Error 10 (consistency of scoring between items regarding the participant's income), Full Assessment Error 5

(consistency between items regarding parenting skills), and Full Assessment Error 7 (consistency between items about the participant's experiences of abuse. Most individual error items occurred at higher rates for Indigenous participants than non-Indigenous participants, however, most of these effects were negligible ($\varphi < .10$). Only with Pre-Screen Error 10 (consistency of scoring between items regarding the participant's income) was the effect of the difference noteworthy; this error occurred in 19.7% of Indigenous assessments and only 7.6% of non-Indigenous assessments ($\varphi = .16$). See Appendix G for detailed sex by Indigenous status results.

Post Hoc: Differences in Error Rate Based on Risk Level. Following the planned analyses for research question 1, a final factor with the potential to influence error rate was evaluated: risk classification level, based on the Pre-Screen assessment. A chi-square test of independence was run, the results of which may be seen in Table 12. Medium and high risk participants were 10 to 13% more likely to have errors in their assessments than low risk participants. The size of this effect was small.

Table 12

Chi-Square Results of Errors as a Function of Pre-Screen Risk Level

Risk Level	Error *n* (%)	χ^2	φ
Low Risk	7519 (41.31)	372.55***	.11
Medium Risk	5563 (51.98)		
High Risk	1367 (53.46)		

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

Summary of Research Question 1 Results. The overall base rate of error was quite low, with most assessments containing either zero errors or one error. Indigenous participants appeared to have a greater number of errors than non-Indigenous participants, both overall and at the item level. ANOVA results supported these findings, indicating that Indigenous participants had consistently higher rates of error, across all error totals. Sex-based differences in error rate

were also examined, and it appeared that in some cases females had higher rates of error than males. However, in most cases these differences were too small to be considered significant. Item-level results were much the same: all items occurred at higher rates for Indigenous participants, and several occurred at higher rates for female participants. The rate of error was also compared for different item types. Contrary to what was hypothesized, static errors were most common, and dynamic-continuum errors were least common. Finally, the relationship between risk level and the presence of errors was assessed. Generally, participants who were rated as higher risk had more errors in their assessments, however, these effects were quite small, and thus may not be important.

Research Question 2: Will changes in the consistency of scored assessments impact their predictive validity based on ROC analysis? Will the relationship between the consistency of scored assessments and their predictive validity vary as a function of sex or race?

The predictive accuracy of assessments based on error rate was evaluated using SPIn Pre-Screen assessment total risk scores. The Pre-Screen total risk score is typically used to estimate an individual's risk of recidivism and to classify individuals as low risk, medium risk, or high risk. Prior to assessing predictive accuracy, assessments had to be sorted into high and low error groups based on the total number of errors that occurred in Pre-Screen assessments. As the base rate of error was low, two different groupings were created to ensure analyses were as detailed as possible while minimizing the impact of observed skewness. Primary analyses were conducted comparing Pre-Screen assessments with zero errors to those with one or more errors. As a follow-up, some analyses were repeated using four error groups: zero errors, one error, two errors, and three or more errors. These results may be viewed in Appendix H. AUCs were calculated using Pre-Screen risk scores and compared based on the presence or absence of errors

for all six recidivism outcomes at one- and three-year follow-ups in the full sample. Comparisons were made using the method developed by Hanley and McNeil (1983). A pooled standard error that corrects for the assumed high correlation between ROC curves, defined as $SE_{(AUC1 - AUC2)} = \sqrt{SE^2(AUC1) + SE^2(AUC2) - 2rSE(AUC1)SE(AUC2)}$ was used to calculate a z-score to compare groups ($z = AUC1 - AUC2/SE_{(AUC1-AUC2)}$). Further analysis broke the sample into sub-groups based on sex and Indigenous status.

Differences in Predictive Validity of Assessments Based on Error Rate in the Full Sample. Area Under the Curve (AUC) values were computed for all six recidivism outcome measures at both one- and three-year follow-up in the full sample, split based on whether the assessment had zero, or one or more errors. These results, along with *z*-scores testing the significance of observed differences, may be seen in Table 13.

A consistent pattern emerged that directly contradicted Hypothesis 4, which stated that assessments with fewer errors would have greater predictive accuracy than assessments with more errors. Assessments with one or more errors had consistently and significantly greater predictive accuracy at both one- and three-year follow-up for any recidivism, any new offence, any return to custody, any new conviction, and violent recidivism. However, the differences in AUC values were generally small, with a maximum difference of .03 (any new conviction, one-year follow-up). The minimum difference in AUC values was .01 for technical violations, at both one- and three-year follow-ups, which was not statistically significant.

Table 13

SPIn Pre-Screen Predictive Validity Results Based on Presence of Errors in the Full Sample

	Error (n = 14,449)	No Error (n = 17,011)	
	AUC (*SE*)	AUC (*SE*)	*z*
Any Recidivism			
1 year	.70*** (.01)	.68*** (.01)	-3.06**
3 years	.71*** (.01)	.69*** (.01)	-3.91***
Technical Violations			
1 year	.73*** (.01)	.72*** (.01)	-0.79
3 years	.73*** (.01)	.72*** (.01)	-1.17
Any New Offence			
1 year	.69*** (.01)	.67*** (.01)	-3.01**
3 years	.69*** (.01)	.67*** (.01)	-3.14**
Any Return to Custody			
1 year	.75*** (.01)	.73*** (.01)	-1.95
3 years	.75*** (.01)	.73*** (.01)	-3.16**
Any New Conviction			
1 year	.65*** (.01)	.62*** (.01)	-3.89***
3 years	.65*** (.01)	.63*** (.01)	-3.34**
Violent Recidivism			
1 year	.69*** (.01)	.71*** (.01)	2.14*
3 years	.68*** (.01)	.70*** (.01)	3.04**

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

Differences in the Predictive Validity of SPIn Pre-Screen Assessments Based on Errors in Demographic Sub-Groups. In an attempt to clarify the root of these unusual results and conduct exploratory analysis of differences in the impact of errors based on sex and Indigenous status, the same analysis was repeated for each demographic sub-group. First, Table 14 provides the AUCs and subsequent *z*-tests for non-Indigenous and Indigenous females. Table 15 provides the same results for non-Indigenous and Indigenous males. The results for each group will be discussed in turn.

The impact of errors on the predictive validity of SPIn Pre-Screen assessments was first explored in female participants. Among non-Indigenous females, AUCs based on assessments

with one or more errors were consistently larger than AUCs based on assessments with zero errors. This difference was significant at both the one- and three-year follow-ups for any recidivism, any new offence, any return to custody, and any new conviction. The differences were too small to reach significance for technical violations and violent recidivism. The largest AUC differences between assessments with and without errors were for the one-year follow-ups of any new offences and any new convictions. For both kinds of recidivism, the assessments with one or more errors had AUCs that were .07 greater than the assessments with no errors.

Divergent results were found for Indigenous females. Smaller differences between AUCs based on assessments with and without errors were found, none of which were significant. Of the 12 AUC comparisons made, half indicated that assessments with errors had greater predictive accuracy, while the other half indicated that assessments without errors had greater predictive accuracy. The largest observed difference in AUC values was for any new offence at both one- and three-year follow-ups, where assessments with errors had AUCs that were .05 greater than assessments without errors. The lack of significant differences for Indigenous females is notable, given that this group had the highest overall rate of error.

Table 14

Differences in Predictive Accuracy by Error Rate Within Female Sub-Groups: Indigenous vs. non-Indigenous

	Non-Indigenous Females			Indigenous Females		
	Error (n = 1873)	No Error (n = 2623)		Error (n = 1070)	No Error (n = 896)	
	AUC (*SE*)	AUC (*SE*)	*z*	AUC (*SE*)	AUC (*SE*)	*z*
Any Recidivism						
1 year	.71*** (.02)	.65*** (.02)	-2.33*	.67*** (.02)	.66*** (.02)	-0.45
3 years	.71*** (.01)	.65*** (.01)	-2.84**	.66*** (.02)	.66*** (.02)	0.06
Technical Violations						
1 year	.74*** (.02)	.74*** (.02)	-0.01	.70*** (.02)	.71*** (.03)	0.24
3 years	.73*** (.02)	.73*** (.02)	0.11	.69*** (.02)	.70*** (.02)	0.65
Any New Offence						
1 year	.70*** (.02)	.63*** (.02)	-2.69**	.68*** (.02)	.63*** (.03)	-1.57
3 years	.69*** (.02)	.63***(.02)	-2.86**	.67*** (.02)	.62*** (.02)	-1.54
Any Return to Custody						
1 year	.78*** (.02)	.75*** (.02)	-1.43	.71*** (.02)	.70*** (.03)	-0.46
3 years	.79*** (.02)	.73*** (.02)	-2.70**	.71*** (.02)	.69*** (.02)	-0.63
Any New Conviction						
1 year	.68*** (.02)	.61*** (.02)	-2.54*	.61*** (.02)	.61*** (.03)	-0.22
3 years	.66*** (.02)	.61*** (.02)	-2.35*	.58*** (.02)	.61*** (.02)	1.10
Violent Recidivism						
1 year	.69*** (.03)	.66*** (.03)	-0.65	.69*** (.03)	.70*** (.03)	0.02
3 years	.67*** (.02)	.67*** (.03)	-0.03	.67*** (.02)	.68*** (.03)	0.40

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

Next, the impact of errors on the predictive validity of SPIn Pre-Screen assessments was explored in male participants. Table 15 illustrates these results for non-Indigenous and Indigenous male participants. The results for non-Indigenous males were similar to those of non-Indigenous females, albeit with smaller effects. Across all twelve recidivism outcomes, AUC values were larger for assessments with errors than to those without errors. However, the difference was only statistically significant in four comparisons: one- and three-year follow-ups for any new conviction, and one- and three-year follow-ups for violent recidivism. AUC values were .03 to .04 larger for groups with error.

Indigenous males had similar results to Indigenous females. Smaller and less consistent differences were observed between AUCs based on assessments with and without errors; for seven recidivism outcomes the assessments with errors had larger AUCs, while the AUCs for the other five outcome measures were larger for assessments without errors. None of these differences were statistically significant, and in all but one item the difference was .01; the AUC for any recidivism at three-year follow-up was .03 greater in assessments with errors.

Taken together, the results presented in Tables 14 and 15 suggest that the counter-intuitive pattern, wherein SPIn Pre-Screen assessments with errors appeared to have greater predictive accuracy than those with no errors, is only true for non-Indigenous participants. Similar to what was observed in the full sample, the assessments of non-Indigenous males and females with errors had greater predictive accuracy than those without errors. However, there were no differences in the predictive accuracy of Indigenous participants' assessments with and without errors, regardless of sex.

Table 15

Differences in Predictive Accuracy by Error Rate Within Male Sub-Groups: Indigenous vs. Non-Indigenous

	Non-Indigenous Males			Indigenous Males		
	Error (n = 9069)	No Error (n = 11,382)		Error (n = 2437)	No Error (n = 2110)	
	AUC (*SE*)	AUC (*SE*)	*z*	AUC (*SE*)	AUC (*SE*)	*z*
Any Recidivism						
1 year	.68*** (.01)	.67*** (.01)	-1.65	.67*** (.01)	.66*** (.01)	-0.37
3 years	.69*** (.01)	.67*** (.01)	-2.14*	.70*** (.01)	.67*** (.01)	-1.47
Technical Violations						
1 year	.71*** (.01)	.70*** (.01)	-1.07	.67*** (.01)	.68*** (.01)	0.32
3 years	.72*** (.01)	.70*** (.01)	-1.52	.68*** (.01)	.68*** (.01)	-0.06
Any New Offence						
1 year	.68*** (.01)	.66*** (.01)	-1.48	.64*** (.01)	.64*** (.01)	0.10
3 years	.68*** (.01)	.67*** (.01)	-1.34	.65*** (.01)	.64*** (.01)	-0.68
Any Return to Custody						
1 year	.73*** (.01)	.71*** (.01)	-1.68	.69*** (.01)	.69*** (.01)	0.12
3 years	.74*** (.01)	.72*** (.01)	-1.95	.71*** (.01)	.70*** (.01)	-1.12
Any New Conviction						
1 year	.64*** (.01)	.61*** (.01)	-2.35*	.61*** (.01)	.60*** (.02)	-0.29
3 years	.68*** (.01)	.62*** (.01)	-2.55*	.61*** (.01)	.60*** (.01)	-0.17
Violent Recidivism						
1 year	.67*** (.01)	.70*** (.01)	2.36*	.64*** (.02)	.65*** (.02)	0.37
3 years	.67*** (.01)	.70*** (.01)	3.12**	.63*** (.01)	.64*** (.02)	0.23

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

Sex Differences in the Predictive Accuracy of the SPIn Pre-Screen. To follow the comparison of AUC values based on error rate, differences in predictive accuracy as a function of sex were evaluated within each error and Indigenous status sub-group. Table 16 illustrates the sex differences among non-Indigenous participants with and without errors in their assessments, while Table 17 provides the same results for Indigenous participants.

The predictive validity of assessments with errors did not vary much as a function of sex among non-Indigenous participants. Overall, most AUC differences were small (i.e., .01 to .03), however, a consistent pattern emerged. The AUC values for females were larger for all six recidivism outcomes at both one- and three-year follow-ups. This difference only reached statistical significance for any return to custody at both follow-ups. Non-Indigenous females had an AUC of .78 at one-year follow-up, and .79 at three-year follow-up, compared to an AUC of .73 at one-year and AUC of .74 at three-year for non-Indigenous males. Interestingly, these sex differences were not present in the assessments without errors; none of the *z*-test results were statistically significant, and eight of the twelve AUC values were greater for males, while the remaining four were greater for females.

Table 16

Sex-Based Differences in Predictive Accuracy Within Non-Indigenous Error Sub-Groups

	Error			No Error		
	Non-Indigenous Females (n = 1873)	Non-Indigenous Males (n = 9069)		Non-Indigenous Females (n = 2623)	Non-Indigenous Males (n = 11,382)	
	AUC (*SE*)	AUC (*SE*)	z	AUC (*SE*)	AUC (*SE*)	z
Any Recidivism						
1 year	.71*** (.02)	.68*** (.01)	-1.36	.65*** (.02)	.67*** (.01)	0.71
3 years	.71*** (.01)	.69*** (.01)	-1.07	.65*** (.01)	.67*** (.01)	1.42
Technical Violations						
1 year	.74*** (.02)	.71*** (.01)	-1.37	.74*** (.02)	.70*** (.01)	-1.99
3 years	.73*** (.02)	.72*** (.01)	-0.65	.73*** (.02)	.70*** (.01)	-1.71
Any New Offence						
1 year	.70*** (.02)	.68*** (.01)	-1.13	.63*** (.02)	.66*** (.01)	1.61
3 years	.69*** (.02)	.68*** (.01)	-0.88	.63*** (.02)	.67*** (.01)	2.20
Any Return to Custody						
1 year	.78*** (.02)	.73*** (.01)	-2.83**	.75*** (.02)	.71*** (.01)	-1.66
3 years	.79*** (.02)	.74*** (.01)	-3.37***	.73*** (.02)	.72*** (.01)	-0.68
Any New Conviction						
1 year	.68*** (.02)	.64*** (.01)	-1.62	.61*** (.02)	.61*** (.01)	0.18
3 years	.66*** (.02)	.65*** (.01)	-0.82	.61*** (.02)	.62*** (.01)	0.77
Violent Recidivism						
1 year	.69*** (.02)	.67*** (.01)	-0.71	.66*** (.03)	.70*** (.01)	1.21
3 years	.67*** (.02)	.66*** (.01)	-0.23	.67*** (.03)	.70*** (.01)	1.15

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

Sex-based differences in predictive accuracy for Indigenous participants may be seen in Table 17. Overall, AUC values were notably smaller for Indigenous participants, and there was no clear pattern of sex-based differences. Within assessments that contained one or more errors AUC values varied by maximum of .04 (any new offence, one-year follow-up); none of the *z*-tests reached statistical significance. Females had larger AUC values for nine of the twelve recidivism follow-ups, while males had larger AUCs for the remaining three. Similarly, within assessments that contained zero errors AUC values varied by a maximum of .05 (any recidivism, three-year follow-up; violent recidivism, one-year follow-up), and none of the *z*-tests were statistically significant. AUC values were greater for females at seven of the twelve recidivism follow-ups, spread across different outcomes and follow-up measures.

Table 17

Sex-Based Differences in Predictive Accuracy Within Indigenous Error Sub-Groups

	Error			No Error		
	Indigenous Females (n = 1070)	Indigenous Males (n = 2437)		Indigenous Females (n = 896)	Indigenous Males (n = 2110)	
	AUC (*SE*)	AUC (*SE*)	z	AUC (*SE*)	AUC (*SE*)	z
Any Recidivism						
1 year	.67*** (.02)	.67*** (.01)	-0.20	.66*** (.02)	.66*** (.01)	0.09
3 years	.66*** (.02)	.70*** (.01)	1.67	.70*** (.04)	.65*** (.02)	-1.26
Technical Violations						
1 year	.70*** (.02)	.67*** (.01)	-1.22	.71*** (.03)	.68*** (.01)	-1.03
3 years	.69*** (.02)	.68*** (.01)	-0.29	.66*** (.02)	.67*** (.01)	0.38
Any New Offence						
1 year	.68*** (.02)	.64*** (.01)	-1.76	.63*** (.03)	.64*** (.01)	0.40
3 years	.67*** (.02)	.65*** (.01)	-0.61	.70*** (.02)	.68*** (.01)	-1.02
Any Return to Custody						
1 year	.71*** (.02)	.69*** (.01)	-0.99	.70*** (.03)	.69*** (.01)	-0.15
3 years	.71*** (.02)	.71*** (.01)	0.20	.62*** (.02)	.64*** (.01)	0.77
Any New Conviction						
1 year	.61*** (.02)	.61*** (.01)	-0.13	.60*** (.03)	.60*** (.01)	-0.07
3 years	.58*** (.02)	.61*** (.01)	1.04	.69*** (.02)	.70*** (.01)	0.18
Violent Recidivism						
1 year	.69*** (.03)	.64*** (.02)	-1.84	.70*** (.04)	.65*** (.02)	-1.26
3 years	.67*** (.02)	.63*** (.01)	-1.20	.61*** (.02)	.60*** (.01)	-0.43

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

Differences in the Predictive Accuracy of the SPIn Pre-Screen by Indigenous Status. Finally, differences in predictive accuracy based on Indigenous status were evaluated in error and sex sub-groups. First, differences in predictive accuracy among female participants were evaluated. These results may be seen in Table 18. Among females with one or more errors in their assessments, eleven of the twelve AUC values compared were larger for non-Indigenous participants, though the difference was only statistically significant at four of the follow-ups. For any return to custody, Indigenous females had an AUC of .71 at one-year follow-up, while non-Indigenous females had an AUC of .78. Similarly, Indigenous females had an AUC of .71 at the three-year follow-up for any return to custody, while non-Indigenous females had an AUC of .79. Significant differences were also observed in both follow-up periods for any new conviction. Indigenous women had AUCs of .61 at one-year follow-up and .66 at three-year follow-up, while non-Indigenous women had AUCs of .68 at one-year follow-up and .66 at three-year follow-up. AUC values for both one- and three-year follow-up with any new conviction and violent recidivism were the same, regardless of Indigenous status.

The pattern of results for female participants whose assessments had zero errors were less clear. In five of the twelve comparisons made, the AUCs were actually larger for Indigenous females than non-Indigenous females, while the rest were larger for non-Indigenous women. None of these differences were statistically significant. The largest difference in AUC values was for any return to custody at one-year follow-up, where non-Indigenous females had an AUC of .75, while non-Indigenous females had an AUC of .70.

Table 18

Indigenous Status-Based Differences in Predictive Accuracy Within Female Error Sub-Groups

	Error			No Error		
	Non-Indigenous Females (n = 1873)	Indigenous Females (n = 1070)		Non-Indigenous Females (n = 1873)	Indigenous Females (n =1070)	
	AUC (*SE*)	AUC (*SE*)	*z*	AUC (*SE*)	AUC (*SE*)	*z*
Any Recidivism						
1 year	.71*** (.02)	.67*** (.02)	1.44	.65*** (.02)	.66*** (.02)	-0.19
3 years	.71*** (.01)	.66*** (.02)	2.08*	.65*** (.01)	.66*** (.02)	-0.48
Technical Violations						
1 year	.74*** (.02)	.70*** (.02)	1.48	.71*** (.03)	.74*** (.02)	1.00
3 years	.73*** (.02)	.69*** (.02)	1.76	.73*** (.02)	.70*** (.02)	1.04
Any New Offence						
1 year	.70*** (.02)	.68*** (.02)	0.72	.63*** (.02)	.63*** (.03)	0.02
3 years	.69*** (.02)	.67*** (.02)	1.10	.63*** (.02)	.62*** (.02)	0.25
Any Return to Custody						
1 year	.78*** (.02)	.71*** (.02)	2.71*	.75*** (.02)	.70*** (.03)	1.42
3 years	.79*** (.02)	.71*** (.02)	3.49***	.73*** (.02)	.69*** (.02)	1.33
Any New Conviction						
1 year	.68*** (.02)	.61*** (.02)	2.17*	.61*** (.02)	.61*** (.03)	0.09
3 years	.66*** (.02)	.58*** (.02)	3.14**	.61*** (.02)	.61*** (.02)	-0.17
Violent Recidivism						
1 year	.69*** (.03)	.69*** (.03)	-0.10	.66*** (.03)	.70*** (.04)	-0.68
3 years	.67*** (.02)	.67*** (.02)	0.05	.67*** (.03)	.68*** (.03)	-0.37

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

Finally, differences in predictive accuracy based on Indigenous status were evaluated in male error sub-groups; these results may be seen in Table 19. For participants with one or more errors, AUCs were generally larger for non-Indigenous participants. These differences were statistically significant for four outcome measures. For technical violations, non-Indigenous males had AUCs of .71 at one-year follow-up and .72 at three-year follow-up, while Indigenous males had AUCs of .67 at one-year-follow-up and .68 at three-year follow-up. At one-year follow-up for any new offence, non-Indigenous males had an AUC of .68, while Indigenous males had an AUC of .64. The difference was not statistically significant at three-year follow-up. Similarly, at one-year follow-up for any return to custody, non-Indigenous males had an AUC of .73, while Indigenous males had an AUC of .69. Finally, the AUCs for any new conviction were significantly larger for non-Indigenous males than Indigenous males at both one- and three-year follow-ups. At one-year follow-up non-Indigenous males had an AUC of .64, while non-Indigenous males had an AUC of .61. At three-year follow-up non-Indigenous males had an AUC of .65, while Indigenous males had an AUC of .61.

Fewer statistically significant differences were observed among males with no errors in their assessments, though again non-Indigenous males had consistently larger AUCs. The only significant differences observed were for violent recidivism. At one-year follow-up non-Indigenous males had an AUC of .70, while Indigenous males had an AUC of .65. At three-year follow-up, non-Indigenous males had an AUC of .70, while Indigenous males had an AUC of .64.

Table 19

Indigenous Status-Based Differences in Predictive Accuracy Within Male Error Sub-Groups

	Error			No Error		
	Non-Indigenous Males (n = 9069)	Indigenous Males (n = 2437)		Non-Indigenous Males (n = 11,382)	Indigenous Males (n =2110)	
	AUC (*SE*)	AUC (*SE*)	*z*	AUC (*SE*)	AUC (*SE*)	*z*
Any Recidivism						
1 year	.68*** (.01)	.67*** (.01)	1.11	.67*** (.01)	.66*** (.01)	0.33
3 years	.69*** (.01)	.70*** (.01)	-0.32	.67*** (.01)	.67*** (.01)	0.08
Technical Violations						
1 year	.71*** (.01)	.67*** (.01)	2.62**	.70*** (.01)	.68*** (.01)	1.31
3 years	.72*** (.01)	.68*** (.01)	2.80**	.70*** (.01)	.68*** (.01)	1.64
Any New Offence						
1 year	.68*** (.01)	.64*** (.01)	2.56*	.66*** (.01)	.64*** (.01)	1.26
3 years	.68*** (.01)	.65*** (.01)	1.83	.67*** (.01)	.64*** (.01)	1.67
Any Return to Custody						
1 year	.73*** (.01)	.69*** (.01)	2.73**	.71*** (.01)	.69*** (.01)	1.20
3 years	.74*** (.01)	.71*** (.01)	1.72	.72*** (.01)	.70*** (.01)	1.57
Any New Conviction						
1 year	.64*** (.01)	.61*** (.01)	1.96*	.61*** (.01)	.60*** (.02)	0.48
3 years	.65*** (.01)	.61*** (.01)	2.98**	.62*** (.01)	.60*** (.01)	1.28
Violent Recidivism						
1 year	.67*** (.02)	.64*** (.02)	1.73	.70*** (.01)	.65*** (.02)	2.94**
3 years	.66*** (.01)	.63*** (.01)	1.82	.70*** (.01)	.64*** (.01)	3.60***

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

Summary of Research Question 2 Results. I hypothesized that overall, assessments without errors would demonstrate greater predictive accuracy than those which had errors. The results presented above directly contradicted my hypothesis. In the full sample, SPIn Pre-Screen assessments with one or more errors generally had higher predictive accuracy than those with zero errors, across all kinds of recidivism and follow-up periods. These analyses were repeated in demographic sub-groups. The same counter-intuitive pattern was observed among non-Indigenous participants—assessments with one or more errors had greater predictive accuracy than those with zero errors. However, no significant differences or clear patterns were observed among Indigenous participants. This suggests that the counter-intuitive results seen in the overall sample are being driven by non-Indigenous participants.

Only small differences in predictive accuracy were observed as a function of sex. Among non-Indigenous participants with errors in their assessments, the SPIn Pre-Screen risk score was a better predictor for females, however there were no clear differences in assessments with zero errors. No significant sex-based differences were observed among Indigenous participants, though within assessments which contained errors, the SPIn Pre-Screen risk score was a slightly better predictor for females than males. It appears that the minimal differences observed in the predictive accuracy of the SPIn Pre-Screen as a function of sex were driven by errors in scoring.

Differences in the predictive accuracy of the Pre-Screen risk score based on Indigenous status were more consistent. For both males and females with errors in their assessments, the SPIn was a better predictor of all kinds of recidivism for non-Indigenous participants than Indigenous participants. For females with no errors in their assessments, there was no clear pattern of differences in predictive validity based on Indigenous status. For males with no errors in their assessments, AUCs were slightly larger for non-Indigenous participants, however these

differences were too small to reach significance. Again, it appears that differences in the predictive accuracy of the SPIn Pre-Screen assessment based on Indigenous status were being driven by errors in scoring.

Discussion

Justice system professionals, such as probation and parole officers, rely on risk assessments to estimate their clients' risk of recidivism and to guide case planning that targets criminogenic needs (Bonta & Andrews, 2017). Risk assessments such as the SPIn are only useful if they may be considered both reliable and valid (DeVellis, 2012). Scoring accuracy is an essential component of the reliability and predictive validity of completed risk assessments, but very little research has been conducted to date evaluating the relationship between scoring accuracy and predictive validity. Previous research that has used proxy indicators for rater accuracy and conscientiousness has indicated that the predictive accuracy of risk assessments may be increased when raters are more accurate (Chadwick, 2014; Hanson et al., 2015), but unfortunately these studies were unable to detect specific sources of error in completed assessments. Additionally, no prior research has examined what kinds of items may be more prone to error, or whether traits of the individual under assessment, such as their sex or race, may impact the rate of errors in completed assessments.

The present study drew on the concepts of inter-rater reliability and alternate forms reliability to develop and test a method for estimating intra-rater consistency. Errors in scoring were identified using inter-item cross-consistency relationships between conceptually similar SPIn items, and two research questions were addressed. First, I asked what the frequency and nature of SPIn coding errors would be, and whether they varied as a function of the sex or Indigenous status of participants. Differences in error rate as a function of both the sex and Indigenous status of participants would be evidence for potential rater bias. Error rates were also compared for different error types to determine whether certain types of items were harder to score accurately than others. Types of inter-item cross-consistency relationships examined in this

study included errors based solely in static SPIn items, errors based on at least one dynamic-checklist SPIn item, and errors based on at least one dynamic-continuum item. Second, I asked whether differences in the error rate of completed assessments would impact their predictive validity, and whether the sex or Indigenous status of participants would impact predictive validity, again as a function of error rate. I hypothesized that assessments with fewer errors would have greater predictive accuracy than those with more errors. I also explored differences in the predictive accuracy of the SPIn Pre-Screen based on the sex and race of the participant. The results of these analyses will be discussed in turn. Limitations of the present study, as well as the implications of these results for practice and research, are also considered.

Error Rate

The rate of error in SPIn assessments was quite low, with most participants having either zero or one error in their assessments. Approximately 65% of the potential error items (21 of 32) generated errors in less than 2% of assessments, and only a few items were found to have occurred in more than 10% of assessments. The six items that occurred in more than 10% of assessments were found in a variety of SPIn domains; three involved only static items, two involved dynamic-checklist items, and one involved dynamic-continuum items. It is unclear why these items had a higher error rate, as there were no tangible differences between these high frequency error items and the other low frequency error items. The static errors were all based in the consistency of scoring between various historical items, such as past perpetration of violent crimes or domestic violence, or victimization the individual may have experienced in childhood; these items should be as easy to score accurately as any other static items in the SPIn. The dynamic errors that occurred at higher rates may be more understandable. For example, one error item was based on consistency between two items regarding the participant's income: if the

person was indicated to be reliant on social assistance in one item, then a corresponding item should indicate that their income is low for their needs. The allowance that those on Alberta Works (Alberta's social assistance program) receive is well below the Low Income Cut-Off, otherwise known as the poverty line, as defined by Statistics Canada (Government of Alberta, 2021; Statistics Canada, 2021b). However, it is possible that some of the justice system professionals who use the SPIn are unaware of how difficult it can be to survive on such a low income, or else may not agree that the Alberta Works allowance is low. This issue may be easily addressed by providing training to officers about social assistance programs and the challenges associated with having a low income. The other dynamic items with high error rates may also be indicative of areas where additional training and guidance may be helpful. For example, training could be provided to officers regarding the challenges associated with gaining employment when one has less than a high school education, or clarifying how to interpret pro-social community participation.

Notwithstanding these items with elevated error rates, it appears that, despite its length and complexity, the SPIn is a tool that can be scored consistently and with a high degree of accuracy. These results are encouraging, as there remains resistance in some jurisdictions to incorporating longer and more complex tools such as the SPIn into their risk assessment and case planning processes (Bonta & Andrews, 2017). The relative value of simple actuarial tools as compared to longer, more complex tools remains a source of debate (Lussier & Davies, 2011). However, as demonstrated by the results of the present study, such tools can be scored with a high degree of consistency. I can further surmise that the training being provided by Orbis Partners to professionals in Alberta is effective, given that so few errors were identified.

Differences in Error Rate by Sex. Exploratory analyses were conducted to determine whether errors were occurring at different rates based on the sex of the participant. While no prior research has investigated the accuracy of scoring based on the sex of the individual under assessment, other evidence of sexist attitudes in the justice system has been found (e.g., Gaader et al., 2004). As such, investigating the potential impact of sex on error rate was determined to be worthwhile. The findings presented here indicate that errors occur at slightly higher rates for female participants as compared to male participants. Significantly higher rates of error were observed for females at both the item-level, and in error totals. However, these differences were too small to be considered important. More research is needed to determine whether the small differences observed in the present study are indicative of a real problem. Nonetheless, it appears as though sexism is not having a significant impact on the accuracy of SPIn scores.

Differences in Error Rate by Indigenous Status. Canada has long history of abuse and systemic discrimination against Indigenous people, and the impact of this maltreatment is highly visible in the justice system (McGuire & Murdoch, 2021). Discriminatory and racist practices have been identified at all levels of the justice system, from policing all the way to risk assessment and case management practices (e.g., Babchishin et al., 2012; McKay, 2021). As such, exploration of the relationship between error rate and Indigenous status was warranted. The results presented in the current study indicate that errors are in fact occurring at higher rates for Indigenous people, as compared to their non-Indigenous counterparts. Small but consistent effects were observed at both the item-level and when looking at errors overall. There are several possible explanations for the observed pattern.

First, it must be noted that due to the lack of previous research using this method, it is possible that the results found here occurred by chance and will not be replicated. However,

given the large sample and consistency of the pattern across all analyses, it seems unlikely that these results were found by chance. Instead, I hypothesize that the results may be evidence of the ongoing discrimination that Indigenous people face in the Canadian justice system. It is plausible that the professionals who are conducting these assessments may, at times, be inadvertently allowing their personal biases to impede their diligence in completing risk assessments. It is unlikely that probation and parole officers are deliberately biasing the results of their assessments; rather, they may be less attentive when conducting assessments with Indigenous peoples. Past research using proxies for rater conscientiousness and scoring accuracy have demonstrated the implications for the predictive accuracy of assessments. For example, Chadwick (2014) found that DRAOR assessments completed by those who had received formal training had significantly higher predictive accuracy than those assessments completed by professionals who had not been trained in the tool's use. Similarly, Hanson et al. (2015) found that complete assessments that were completed—the proxy used for rater conscientiousness in this study—had significantly greater predictive accuracy than those which were incomplete. Extensive research has also documented the experiences justice-involved Indigenous Canadians with overt and covert racism and stereotyping (e.g., Clark, 2019; McKay, 2021). Common assumptions include that Indigenous people are lazy, substance abusers; they are often treated as though they are dangerous criminals, even when their offences are minor (Cesaroni et al., 2019). Such stereotypes, along with the known fact that Indigenous people tend to recidivate at higher rates than non-Indigenous people, may lead some professionals to unconsciously write off their Indigenous clients as hopeless, or otherwise impossible to help. An officer may unintentionally be more or less conscientious when completing risk assessments based on their own personal impression of the client's risk for reoffence. Further research will be necessary to determine

whether this adequately explains the higher error rate found in Indigenous assessments. Alternatively, it must be acknowledged that given the small effects observed, it is possible that these differences in error rate will not be replicated in future studies.

Differences in Error Rate Based on Type of Error Item. It was hypothesized that certain kinds of SPIn items would have higher rates of error based on how complex or challenging they may be to score. More specifically, I suggested that static items—which typically have simple numeric or yes/no responses—would have the lowest rate of error, dynamic-continuum items—scored on a poled scale from high risk to high strength—would have the highest rate of error. I further hypothesized that dynamic-checklist items would occur at a rate somewhere between static and dynamic-continuum errors, as these items have more straightforward scoring than dynamic-continuum items, but still require more rater discretion than static items. The results presented here directly contradicted this hypothesis. The highest rate of error was among static items, followed by dynamic-checklist items, and dynamic-continuum items had the lowest rate of error. Several possible explanations for this pattern should be considered.

The hypothesis that static items would have the lowest rate of error was based in the assumption that items that are more difficult to score—those which require more subjective evaluation or rater discretion—would be more prone to error than those that have simple, straightforward answers. This does not appear to be the case: the most challenging items appear to be less error prone. I suggest that this might be due to rater attentiveness or conscientiousness. Static items, which require simple responses, may be easier to misread, mis-score, or generally overlook as they require little mental effort to score. Conversely, dynamic items, particularly dynamic-continuum items which require the rater to make a subjective evaluation of the client,

may be less prone to inattention due to the greater effort required. It is possible that the same factors that lead to lower data quality in self-report survey measures (e.g., Meade & Craig, 2012) do not create problems in tools such as the SPIn. Furthermore, the SPIn and other similar risk assessments are used by trained professionals to guide and support their work with clients; it is reasonable to assume that such raters would be more diligent in their scoring of complex items than, for example, undergraduate students participating in a study for course credit.

It is also possible that professionals are less attentive when scoring historic, static items as compared to dynamic items due to their relevance in case management. Static items are extremely useful for predicting the likelihood that an individual will reoffend, but they are less useful in case planning and management. Dynamic items provide information about areas of criminogenic need that may be targeted through intervention, and thus are much more useful to probation and parole officers. For this reason, professionals may be more diligent in scoring dynamic items.

Predictive Accuracy and Error Rate

It was hypothesized that overall, assessments with higher rates of error would have lower predictive accuracy than assessments with lower rates of error. The known relationship between reliability and predictive validity, as well as preliminary results using a similar technique (Robinson, 2019), indicated that support for this hypothesis should have been found. Instead, an antithetical pattern of results was discovered in the full sample. Across six measures of recidivism at both one- and three-year follow-up periods, Pre-Screen assessments that contained errors had greater predictive accuracy than assessments that contained no errors.

These results are surprising and appear to defy logic. However, I propose that, in reality, there were no important differences in predictive accuracy based on error rate. In a sample as

large as the one employed in the present study it is not difficult to find statistically significant results. Thus, it is more important to attend to the size of the differences. In the full sample, the largest difference in AUC values was .03—which may otherwise be interpreted as a difference of 3% in the overall predictive accuracy of the SPIn Pre-Screen. This difference is quite small and may not be of any practical importance.

A more complex picture emerged when the predictive accuracy of the SPIn was assessed as a function of error within each demographic sub-group. The same counter-intuitive pattern of results observed in the full sample was found among non-Indigenous males and females: assessments containing errors had slightly higher predictive accuracy than assessments with no errors, though with very small effects. However, among Indigenous participants, there were no clear differences in the predictive accuracy of assessments with and without errors. It is unclear why the effects of error were different for Indigenous and non-Indigenous participants. However, as Indigenous participants had the highest rate of error, I suggest that this provides further evidence that errors have no meaningful impact on the predictive accuracy of assessments, at least when they occur at low rates as observed in this study.

It appears that the assumption made by classical test theory, that errors in scoring balance out and leave the mean relatively unaffected, is likely true for the SPIn (DeVellis, 2012). With and without errors, the SPIn Pre-Screen was a good predictor of all kinds of recidivism with medium to large effect sizes. Errors in completed assessments do not appear to have much—if any—impact on their predictive accuracy.

Differences in Predictive Accuracy by Sex and Error Rate. The predictive accuracy of the SPIn Pre-Screen was evaluated and compared based on the sex of participants. Sex-based comparisons were made within Indigenous and non-Indigenous sub-groups as well as within sub-

groups based on the presence or absence of errors. Only small differences in predictive accuracy were observed among non-Indigenous participants. In assessments that contained errors female participants had larger AUCs, though the difference was only statistically significant for any return to custody, where AUCs were .05 larger for females at both one- and three-year follow-up. Differences were smaller in assessments without errors and about half of the AUCs were larger for females, while the rest were larger for males. This is an interesting finding: it appears that the errors may be the root of the small sex-based differences in predictive accuracy of the SPIn for non-Indigenous participants.

Consistent with the results for non-Indigenous participants, only small sex-based differences in predictive accuracy were observed among Indigenous participants. Though none of the differences were statistically significant, in assessments that contained error most AUCs were larger for females than for males. Even smaller differences were observed in assessments without errors, and about half of the AUCs were larger for males while the other half were larger for females. Taken together, it appears that small sex differences in predictive accuracy only exist in assessments that contain error. SPIn Pre-Screen assessments with no errors demonstrate comparable predictive validity, regardless of the participant's sex.

Differences in Predictive Accuracy by Indigenous Status and Error Rate. Finally, differences in the predictive accuracy of the SPIn Pre-Screen based on error and Indigenous status were evaluated. Comparisons were made between Indigenous and non-Indigenous participants within sexes and error groups. In assessments that contained errors, the SPIn Pre-Screen was a better predictor for non-Indigenous females than Indigenous females across all measures of recidivism, though only half of these differences were large enough to reach statistical significance. Differences were much smaller in assessments that contained no errors, and five AUCs were actually larger for Indigenous females. Results were much the same among male participants. In assessments that contained errors, most AUCs were larger for non-Indigenous males and half of these differences were statistically significant. Smaller differences were observed in assessments without errors, though all AUCs remained larger for non-Indigenous participants. The difference was only statistically significant for violent recidivism. As seen with sex-based differences, it appears that errors in scoring may be driving some of the differences in the SPIn Pre-Screen's predictive accuracy between Indigenous and non-Indigenous participants.

In sum, the SPIn Pre-Screen was a more effective predictor for non-Indigenous participants. However, it appears that these differences in predictive accuracy may be largely attributable to errors in scoring; differences in predictive accuracy based on Indigenous status were either smaller or absent in assessments without errors. This is an encouraging result: the SPIn itself is functioning well for Indigenous and non-Indigenous participants alike, and the small differences in predictive accuracy may be reduced or even eliminated, if errors in scoring can be reduced or eliminated.

There are several ways to interpret these results, and more research is required before any firm conclusions may be drawn. First, it should be acknowledged that given the unique nature of the methods used in the present study and the small effect sizes observed, it is possible these results will not be replicated in other samples or with other measures. However, I suggest that these results may be indicative of the impact racial bias may have on a professional's conscientiousness while conducting risk assessments. Racial biases may be addressed through training; this will be discussed in greater detail in subsequent sections.

Limitations

There are several limitations of the present study that are worthy of consideration. First and foremost, the strategy of using inter-item cross-consistency relationships to identify errors in completed risk assessments has never been tested in published research. Only one previous unpublished report was identified, and only minimal details were provided regarding their analytic strategy (Robinson, 2019). What little previous research exists looking at alternative methods for detecting scoring errors suggests that proxies for scoring accuracy can be useful for distinguishing between high- and low-quality assessments (e.g., Hanson et al., 2015; Meade & Craig, 2012), but much more research is needed. Without replication, it is impossible to say conclusively whether this new technique for assessing intra-rater consistency will be useful or informative in other contexts or with other measures.

It is also important to note that the sample used in the present study overlaps with the samples used in most other SPIn studies (Jones et al., 2015; Jones & Robinson, 2017a, 2018). It is essential that new research be conducted with the SPIn in new samples. New assessment data from Alberta, or from one of many American jurisdictions that use the SPIn, should be used in

future SPIn research. Without such research, it is impossible to know whether the results seen with Alberta community corrections will extend to other locations and demographics.

Finally, the scope of the present study was limited by the nature of the dataset provided by the Alberta Solicitor General. No information about the assessors was provided, and as such many worthwhile questions could not be addressed. Had we been able to link assessments to the officers who completed them, we may have been able to investigate whether certain individuals were making more errors than others. Furthermore, interactions between the sex and race of the assessor and the sex and race of the client may have provided more informative results regarding the impact of sexism and racism on rater accuracy. The utility of such research will be discussed in greater detail in the following section.

Recommendations for Practice and Future Research

Despite the unexpected nature of results observed in the present study, several recommendations can be made. First, it appears that generally, the SPIn itself can be scored consistently and accurately. Despite the length of the Full Assessment, and the complexity of some items, very few errors in scoring arose. Only a few individual SPIn items showed elevated error rates. In my opinion, the static items with elevated error rates do not need to be adjusted, as the directions are already clear and easy to follow. However, the guidelines for the dynamic items with elevated error rates may be worth revisiting. For example, a clearer definition of what constitutes a low income for one's needs may help reduce errors in this area.

I also recommend that Orbis Partners make use of these inter-item cross-consistency relationships to help professionals improve their scoring. By embedding the cross-consistency code I developed into the Case Works software used alongside the SPIn, errors in scoring could be detected when they are first made. A pop-up could appear if inconsistent scores are entered so

that the professional using the tool could rectify their mistake prior to proceeding. Furthermore, Orbis may be able to track the number of errors or inconsistencies made by those using the SPIn. If an individual's rate of error becomes problematic, they may be flagged for additional training or mentorship—a service that Orbis Partners already offers to jurisdictions that use their tools.

I also have several recommendations for the Alberta Solicitor General. Most important will be to address the tentative evidence of sex- and race-based bias in SPIn scoring. Despite the generally small differences in error rate noted, it was found that errors in scoring were related to increased disparities in predictive accuracy as a function of both sex and Indigenous status. If my hypothesis is correct, and rater bias is indeed driving some of the errors made in scoring the SPIn, thereby causing differences in predictive accuracy based on participant sex and Indigenous status, it may be possible to ensure that future SPIn assessments are more equitable. By addressing any underlying biases that probation and parole officers may have, the accuracy and predictive validity of the SPIn could be improved. The best strategy for addressing these biases is less clear. Plenty of anti-racist, feminist, and intersectional training exists; a brief search reveals that a wide variety of options—online and in-person, in both public and private sectors—are available in Alberta and across Canada (e.g., Canada School of Public Service, 2021; Calgary Anti-Racism Collective, 2021). Unfortunately, at this time no research has been conducted to evaluate the impact of such training on an individual's biases. Many measures of racist and anti-racist attitudes have been developed and validated (e.g., Grigg & Manderson, 2016; Knowles & Hawkman, 2020), however no studies to date have used these tools to examine changes in attitudes following training. Given the challenges associated with creating anti-racist training programs that don't perpetuate and reinforce racism through white saviourism, it is essential that the impact of these trainings be evaluated (James, 1995; Jeffery, 2005). Thus, along with the

recommendation to implement training to address biases, I also suggest that such measures be used to compare attitudes before and after training. The topic of anti-racist education is sensitive and politically charged, particularly in recent years (e.g., the debate surrounding teaching Critical Race Theory in American schools), but we cannot allow this to block progress. Addressing bias in risk assessment scores is an important step towards equity in correctional practice.

Finally, several important questions for future research have arisen from the results of the present study. To address limitations previously discussed, future research with the SPIn should be conducted with new samples. The dataset provided by the Alberta Solicitor General is large and rich with information, and as such has been used in most other SPIn research to date (Jones et al., 2015; Jones & Robinson, 2017a, 2018; Wanamaker, 2020). It is essential that new samples be used to determine if the results of past research are generalizable to other jurisdictions.

It would also be valuable to use cross-consistency items to evaluate professionals who use the SPIn and gain a better understanding of the factors that influence inattentive or inconsistent scoring. Studies that have examined the interaction between the sex and race of justice system professionals and the sex and race of their justice-involved clients have yielded mixed results. Some studies indicate that these factors interact to influence decision-making (e.g., Leiber et al., 2016), while others revealed no interaction (e.g., Munoz et al., 2021). The methodology used in the present study may be employed to elucidate how the intersections of race and sex influence the risk assessment process.

Conclusion

The present study was the first to investigate the utility of inter-item cross-consistency relationships to identify errors in completed risk assessments, and to explore how these errors impact predictive validity. Despite the unexpected nature of many of the results presented here,

several important patterns were noted. Errors occur very infrequently. Further, errors appear to have little—if any—impact on the predictive validity of the SPIn overall. However, both the sex and Indigenous status of participants were found to be related to error rate and predictive validity. Female participants had slightly more errors in their assessments than male participants, though with very small effects. Indigenous participants also had higher rates of error than non-Indigenous participants. These results appear to suggest that professionals may be unintentionally allowing their personal biases to influence their attentiveness and accuracy in completing risk assessments. Additionally, small differences in the predictive accuracy of the SPIn Pre-Screen were observed as a function of sex and Indigenous status in assessments that contained errors. However, these differences were reduced or eliminated in assessments that contained no errors. This is a particularly important finding, as it indicates that errors in scoring—that are potentially driven by the biases of probation and parole officers—may be at the root of differences in the predictive accuracy of risk assessments. In sum, intra-rater consistency may be a useful proxy for scoring accuracy; future research would benefit from using this technique to clarify the relationship between a rater's sex and race, and the sex and race of their clients.

References

Austin, J., Coleman, D., Peyton, J., & Johnson, K.D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, D.C.: Institute on Crime, Justice, and Corrections at The George Washington University

Babchishin, K. M., Blais, J. & Helmus, L. (2012). Do Static Risk Factors Predict Differently for Aboriginal Sex Offenders? A Multi-site Comparison Using the Original and Revised Static-99 and Static-2002 Scales. *Canadian Journal of Criminology and Criminal Justice, 54*(1), 1-43. https://doi.org/10.3138/cjccj.2010.E.40

Baird, C., Healy, T., Johnson, K., Bogie, A., Dankert, E. W. & Scharenbroch, C. (2013). *A Comparison of Risk Assessment Instruments In Juvenile Justice.* National Council on Crime and Delinquency.

Belknap, J. (2015). *The invisible woman: Gender, crime and justice* (4th ed.). Stamford, CT: Cengage Learning.

Bonta, J. & Andrews, D. A. (2017). *The Psychology of Criminal Conduct* (6th ed.). New York, NY: Routledge.

Brown, S. L. & Motiuk, L. L. *The Dynamic Factors Identification and Analysis (DFIA) Component of the Offender Intake Assessment (OIA) Process: A Meta-Analytic, Psychometric and Consultative Review.* Correctional Service of Canada: Ottawa, ON.

Canada School of Public Service. (2021). *Tools, Training and Resources to Combat Racism & Discrimination in the Workplace*. https://doi.org/10.1080/07418820400095901

Calgary Anti-Racism Education (CARED). (2021). *The CARED Collective: About Us.* https://doi.org/10.1080/07418820400095901

Cesaroni, C., Grol, C. & Fredericks, K. (2019). Overrepresentation of Indigenous youth in Canada's Criminal Justice System: Perspectives of Indigenous young people. *Australian & New Zealand Journal of Criminology, 52*(1), 111-128. https://doi.org/10.1177/0004865818778746

Chadwick, N. C. (2014). *Validating the Dynamic Risk Assessment for Offender Re-entry (DRAOR) in a sample of US probationers and parolees* (Master's thesis). Carleton University, Ottawa, Canada.

Chadwick, N. C. (2020). *Examining Trajectories of Change on Risk and Protective Factors among White and Black Men Offenders on Community Supervision in Iowa* (Doctoral dissertation). Carleton University, Ottawa, Canada.

Clark, S. (2019). *Overrepresentation of Indigenous People in the Canadian Criminal Justice System: Causes and Responses*. Department of Justice Canada. Retrieved from https://publications.gc.ca/collections/collection_2021/jus/J4-99-2019-eng.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). Hillside, NJ: Lawrence Erlbaum Associates.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., Farrington, D. P., & Rogers, R. D. (2009). Gender differences in structured risk assessment: Comparing the accuracy of five instruments. *Journal of Consulting and Clinical Psychology, 77*(2), 337-348. http://dx.doi.org/10.1037/a0015155

Connelly, K. & Heesacker, M. (2012). Why Is Benevolent Sexism Appealing? Associations with System Justification and Life Satisfaction. *Psychology of Women Quarterly, 36*(4), 432-443. https://doi.org/10.1177/0361684312456369

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antracist Politics. *University of Chicago Legal Forum, 1*(8), 139-167. Retrieved from https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8/?utm_source=chicagounbound.uchicago.edu%2Fuclf%2Fvol1989%2Fiss1%2F8&utm_medium=PDF&utm_campaign=PDFCoverPages

Cutroni, L. & Anderson, J. (2021). Lady Injustice: The Moderating Effect of Ambivalent Sexism in a Mock Case of Intimate Partner Homicide. *Criminal Justice and Behaviour, 48*(3), 373-390. https://doi.org/10.1177/0093854820967704

Department of Justice Canada. (2019). *Indigenous overrepresentation in the criminal justice system.* JustFacts, Research and Statistics Division, Department of Justice Canada.

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services, 13*, 206-222. https://doi.org/10.1037/ser0000075

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care, 44*(11), S50-S59. https://doi.org/134.117.10.200

DeVellis, R. F. (2012). *Scale Development: Theory and Applications* (3rd edition). Sage Publications.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., …, Kostas-Polston, E. (2007). A Psychometric Toolbox for Testing Validity and Reliability. *Journal of Nursing Scholarship, 39*(2), 155-164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Dick, C. (2020). Sex, Sexism, and Judicial Misconduct: How the Canadian Judicial Council Perpetuates Sexism in the Legal Realm. *Feminist Legal Studies, 28*, 133-153. https://doi.org/10.1007/s10691-020-09431-5

Duwe, G., & Rocque, M. (2017). Effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy, 16*, 235-269. https://doi.org/10.1111/1745-9133.12270

Ewert v. Canada, 2018 Supreme Court of Canada 30, [2018] 2 S.C.R. 165

Gaader, E., Rodriguez, N. & Zatz, M. S. (2004). Criers, Liars, and Manipulators: Probation Officers' Views of Girls. *Justice Quarterly, 21*(3), 547-578. https://doi.org/10.1080/07418820400095901

Geck, C. (2012). *The Youth Assessment Screening Instrument: A Psychometric Evaluation with Canadian Male Youthful Offenders* (unpublished master's thesis). Carleton University, Ottawa, ON.

Genesereth, M., Chaudhri, V. K., Brachman, R., Rossi, F. & Stone, P. (2020). *Introduction to Logic Programming.* Morgan & Claypool Publishers.

Goforth, C. (2015). *Using and Interpreting Cronbach's Alpha*. Research and Data Services, University of Virginia Library. Retrieved from: https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/

Gonzalez, R. (2009). *Data Analysis for Experimental Design*. The Guilford Press.

Government of Alberta (2021). *Income Support*. https://www.alberta.ca/income-support.aspx

Gravetter, F. J. & Forzano, L.-A. B. (2016). *Research Methods for the Behavioural Sciences* (5th ed.). Cengage Learning.

Gravetter, F. J. & Wallnau, L. B. (2014). *Essentials of Statistics for the Behavioural Sciences* (8th ed.). Cengage Learning.

Grigg, K. & Manderson, L. (2016). The Australian Racism, Acceptance, and Cultural-Ethnocentrism Scale (RACES): item response theory findings. *International Journal for Equity in Health, 15*(49). http://dx.doi.org.proxy.library.carleton.ca/10.1186/s12939-016-0338-4

Gutierrez, L., Wilson, H. A., Rugge, T., & Bonta, J. (2013). The prediction of recidivism with Indigenous offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology and Criminal Justice, 55*, 55-99. https://doi.org/10.3138/cjccj.2011.E.51

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hamilton, M. (2019). The Sexist Algorithm. *Behavioural Sciences & the Law, 37*(2), 145-157. https://doi.org/10.1002/bsl.2406

Hannah-Moffat, K. & Shaw, M. (2001). *Taking Risks: Incorporating Gender and Culture into the Classification and Assessment of Federally Sentenced Women in Canada.* Status of Women Canada.

Hanson, R. K., Helmus, L.-M. & Harris, A. J. R. (2015). A Prospective Study Using STABLE-2007, Static-99R, and Static-2002R. *Criminal Justice and Behaviour, 42*(12), 1205-1224. https://doi.org/10.1177/0093854815602094

Helmus, L. M. & Babchishin, K. M. (2017). Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy. *Criminal Justice and Behaviour, 44*(1), 8-25, https://doi.org/10.1177/0093854816678898

Herzog, S. & Oreg, S. (2008). Chivalry and the Moderating Effect of Ambivalent Sexism: Individual Differences in Crime Seriousness Judgments. *Law & Society Review, 42*(1), 45-74. https://doi.org/10.1002/bsl.2406

IBM Corp. (Released 2020). *IBM SPSS for Mac, Version 27.* Armonk, NY: IMB Corp

James, C. E. (1995). Multicultural and Anti-Racism Education in Canada. *Race, Gender & Class, 2*(3), 31-48. Retrieved from https://www.jstor.org/stable/41674707?seq=1&cid=pdf-

Jeffery, D. (2005). 'What good is anti-racist social work if you can't master it'?: exploring a paradox in anti-racist social work education. *Race, Ethnicity and Education, 8*(4), 409-425. https://doi.org/10.1080/13613320500324011

Jones, N. J., Brown, S. L., Robinson, D. & Frey, D. (2015). Incorporating Strengths Into Quantitative Assessments of Criminal Risk for Adult Offenders. *Criminal Justice and Behaviour, 42*(3), 321-338. https://doi.org/10.1177/0093854814547041

Jones, N. J. & Robinson, D. (2017a). *The Validity of the Service Planning Instrument (SPIn) for Alberta Community Corrections*. Ottawa, ON: Orbis Partners Inc.

Jones N. J. & Robinson, D. (2017b). *The Validity of the Service Planning Instrument for Clark County, Washington Adult Probation.* Ottawa, ON: Orbis Partners Inc.

Jones, N. J. & Robinson, D. (2018). Service Planning Instrument. In J. P. Singh, D. G., Kroner, J. S., Wormith, S. L. Desmarais & Z. Hamilton (Eds.), *Handbook of Recidivism Risk/Needs Assessment Tools*. (First Edition, pp. 181-198). John Wiley & Sons Ltd.

Kennealy, P. J., Skeem, J. L. & Hernandez, I. R. (2017). Does Staff See What Experts See? Accuracy of Front Line Staff in Scoring Juvenile's Risk Factors. *Psychological Assessment, 29*(1), 26-34. https://doi.org/10.1037/pas0000316

Knowles, R. T. & Hawkman, A. M. (2020). Anti-Racist Quantitative Research: Developing, Validating, and Implementing Racialized Teaching Efficacy and Racial Fragility Scales. *The Urban Review, 52*, 238-262. http://dx.doi.org.proxy.library.carleton.ca/10.1186/s12939-016-0338-4

Lee, S. C., Hanson, R. K. & Blais, J. (2020). Predictive accuracy of the Static-99R and Static-2002R risk tools for identifying Indigenous and White individuals at high risk for sexual recidivism in Canada. *Canadian Psyhology/Psychologie Canadienne, 61*(1), 42-57. https://doi.org/10.1037/cap0000182

Leiber, M. J., Peck, J. H. & Beaudry-Cyr, M. (2016). When does Race and Gender Matter? The Interrelationships between the Gender of Probation Officers and Juvenile Court Detention and Intake Outcomes. *Justice Quarterly, 33*(4), 614-641. https://doi.org/10.1080/07418825.2014.958185

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the level of service inventory revised (LSI-R). *Federal Probation, 68*(3), 34-38.

Lussier, P. & Davies, P. (2011). A Person-Oriented Perspective on Sexual Offenders, Offending Trajectories, and Risk of Recidivism: A New Challenge for Policymakers, Risk Assessors, and Actuarial Prediction? *Psychology, Public Policy, and Law, 17*(4), 530-561. https://doi.org/10.1037/a0024388

Malakieh, J. (2019). Adult and youth correctional statistics in Canada, 2017/2018. *Juristat, 85*(2), retrieved from https://www150.statcan.gc.ca/n1/en/pub/85-002-x/2019001/article/00010-eng.pdf?st=TX_fAGOr

McGuire, M. M. & Murdoch, D. J. (2021). (In)-justice: An exploration of the dehumanization, victimization, criminalization, and over-incarceration of Indigenous women in Canada. *Punishment & Society*, *0*(0), 1-22. https://doi.org/10.1177/14624745211001685

McKay, J. (2021). Systemic Racism in Policing in Canada: Report of the Standing Committee on Public Safety and National Security. House of Commons Standing Committee on Public Safety and National Security. Retrieved from https://publications.gc.ca/collections/collection_2021/parl/xc76-1/XC76-1-1-432-6-eng.pdf

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433, https://doi.org/10.1037/met0000144

Meade, A. W. & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods, 17*(3), 437-455. https://doi.org/10.1037/a0028085

Miller, J. & Maloney, C. (2013). Practitioner Compliance with Risk/Needs Assessment Tools. *Criminal Justice and Behaviour, 40*(7), 716-736. https://doi.org/10.1177/0093854812468883

Multon, K. D. & Coleman, J. S. M. (2018). In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, https://doi.org/10.4135/9781506326139.n344

Munoz, C. G., Perrault, R. T. & Vincent, G. M. (2021). Probation Officer Assessments of Risk when the Youth Look Different: Contributions of Structured Professional Judgment to Concerns About Racial Bias. *Youth Violence and Juvenile Justice, 19*(2), 206-226. https://doi.org/10.1177/1541204020954264

Nuffield, J. (1982). Parole Decision-Making in Canada: Research Towards Decision Guidelines. Ottawa, ON: Solicitor General of Canada.

Office of the Correctional Investigator (OCI) (2019) Office of the Correctional Investigator annual report 2018–2019. Retrieved from: www.oci-bec.gc.ca/cnt/rpt/annrpt/annrpt20182019-eng.aspx

Orbis Partners, Inc. (2000). *Youth Assessment and Screening Inventory (YASI).* Ottawa, ON: Author.

Orbis Partners, Inc. (2003). *The Service Planning Instrument (SPIn).* Ottawa, ON: Author.

Perley-Robertson, B., Helmus, M. & Forth, A. (2019). Predictive accuracy of static risk factors for Canadian Indigenous offenders compared to non-Indigenous offenders: implications for risk assessment scales. *Psychology, Crime & Law, 25*(3), 248-278. https://doi.org/10.1080/1068316X.2018.1519827

Public Safety Canada. (2018). Corrections and Conditional Release Statistical Overview. *Public Safety Canada Portfolio Corrections Statistics Committee.* Retrieved from https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ccrso-2018/ccrso-2018-en.pdf

Rice, M. E. & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behaviour, 29,* 615-620, https://doi.org/10.1007/s10979005-6832-7

Robinson, D. (2019). [Unpublished data illustrating AUC differences in completed Youth Assessment Screening Instrument (YASI) with high and low error rate]. Orbis Partners, Ottawa, ON.

Rocque, M., & Plummer-Beale, J. (2016). In the eye of the beholder? An examination of the inter-rater reliability of the LSI-R and YLS/CMI in a correctional agency. *Journal of Criminal Justice, 42*, 568-578. https://doi.org/10.1016/j.jcrimjus.2014.09.011

Salisbury, E. J. & Van Voorhis, P. (2009). Gendered Pathways: A Quantitative Investigation of Women Probationer's Paths to Incarceration. *Criminal Justice and Behaviour, 36*(6), 541-566. https://doi.org/10.1177/0093854809334076

Simourd, D. (2006). *Validation of risk/needs assessment in the Pennsylvania Department of Corrections: Final Report*. Hampden Township, PA: Department of Corrections.

Statistics Canada (2021a). Table 35-10-0183-01 Incident-based crime statistics, by detailed violations, police services in Alberta. https://doi.org/10.25318/3510018301-eng

Statistics Canada (2021b). Table 11-10-0241-01 Low income cut-offs (LICOs) before and after tax by community size and family size, in current dollars. https://doi.org/10.25318/1110024101-eng

Steinbach, B. & Posthoff, C. (2013). *Boolean Differential Equations.* Morgan & Claypool Publishers.

Stevens, K. L., Austin, A., Wheeler, D. & Malec, T. (2021). The role of the defendant gender on juror decision-making within a mock sex trafficking case among a jury-eligible community sample. *The Journal of Sexual Aggression,* 1-13. https://doi.org/10.1080/13552600.2021.1973127

Stewart, L. A., Wardrop, K., Thompson, J., Derkzen, D. & Motiuk, L. (2017). *Reliability and Validity of the Dynamic Factors Identification and Analysis-Revised*. Correctional Service of Canada: Ottawa, ON.

Stroh, M., Eichinger, M., Giza, A., Hirschmann, N., Bögelein, N., Pitsela, A. & Neubacher, F. (2016). Are Female Offenders Underreported Compared to Male Offenders? A German-Greek Comparison of Crime Reporting, Rating of Offence Seriousness and Personal Experiences of Victimisation. *European Journal of Criminal Policy and Research, 22*(4), 635-653. https://doi.org/10.1007/s10610-016-9302-5

Viljoen, S., Nicholls, T., Roesch, R., Gagnon, N., Douglas, K. & Brink, J. (2016). Exploring Gender Differences in the Utility of Strength-Based Assessment Measures. *International Journal of Forensic Mental Health, 15*(2), 149-163. http://dx.doi.org/10.1080/14999013.2016.1170739

Walters, G. D. (2011). Predicting recidivism with the psychological inventory of criminal thinking styles and level of service inventory-revised: Screening version. *Law and Human Behaviour, 35,* 211-229. https://doi.org/10.1007/s10979-010-9231-7

Wanamaker, K. (2020). *A Multi-Wave Longitudinal Examination of How Strengths and Risks Inform Risk Assessment Treatment profiles for Justice-Involved Men and Women Using the Service Planning Instrument (SPIn)* (Doctoral dissertation). Carleton University, Ottawa, Canada.

Wattanaporn, K. A. & Holtfreter, K. (2014). The Impact of Feminist Pathways Research on Gender-Responsive Policy and Practice. *Feminist Criminology, 9*(3), 191-207. https://doi.org/10.1177/1557085113519491

Wesely, J. K. & Miller, J. M. (2018). Justice System Bias Perceptions of the Dually Marginalized: Observations from a Sample of Women Ex-offenders. *Victims & Offenders, 13*(4), 451-470. https://doi.org/10.1080/15564886.2017.1362614

Yesberg, J. A., & Polaschek, D. L. L. (2015). Assessing dynamic risk and protective factors in the community: examining the validity of the Dynamic Risk Assessment for Offender Re-entry. *Psychology, Crime & Law, 21*, 80-99. https://doi.org/10.1080/1068316X.2014.935775

Appendix A

Table 1

Summary of SPIn Studies to Date

Study	SPIn components evaluated	Population	Sample	Follow-Up Period	ROC Results	Logistic Regression Results
Jones et al., 2015	Pre-Screen Risk and Strength	Provincially sentenced adults on community supervision in Alberta (overlap with sample in current study) between 2009 and 2011	*N* = 3656 (*n* = 694 female, *n* = 2962 male; *n* = 635 Indigenous, *n* = 3021 non-Indigenous)	18 months	Any new offence: AUC values ranging from .75 (Indigenous participants) to .77 (full sample, male participants)	Risk total: significant predictor in all models except for female participants Strength total: significant predictor in all models Risk x Strength: significant predictor in all models except for Indigenous participants
Jones & Robinson, 2017a	Pre-Screen and Full Assesment	Provincially sentenced adults on community supervision in Alberta between 2009 and 2012	*N* = 46,794 (*n* = 9636 female, *n* = 37,158; *n* = 9586 Indigenous, *n* = 37,208 non-Indigenous)	1 year, 3 years, 6 years	One year: AUCs ranging from .64 (any conviction) to .75 (any custody)	No logistic regression

		(partially overlaps with sample in current study)			Three year: AUCs ranging from .64 (any conviction) to .74 (any custody) Six year: .64 (any conviction) to .74 (any custody)	
Jones & Robinson, 2017b	Pre-Screen Risk and Strength, Full Assessment Risk and Strength	Adults on probation in Clark County, Washington between 2010 and 2016	*N* = 2248 (*n* = 499 female, *n* = 1749 male; *n* = 1993 White, *n* = 162 Black, *n* = 93 other)	3 years, 6 years	AUC values for Pre-Screen Risk scores ranging from .64 (one-year follow-up) to .70 (six year follow up) AUC values for Pre-Screen Strength score ranging from .62 (two and six year follow-ups) to .63 (one, three, four, and five year follow-ups) Domain AUCs ranging from small to medium effects	No logistic regression

Jones & Robinson, 2018	Pre-Screen Risk and Strength	Provincially sentenced adults in custody in Alberta between 2009 and 2014	*N* = 501 (*n* = 91 female, *n* = 410 male; *n* = 208 Indigenous, *n* = 293 non-Indigenous)	3 years	Any new offence: AUC values ranging from .73 (female participants) to .78 (male participants) Domain risk and strength scores statistically significant predictors with small to medium effects	Risk total and all domain risk scores significant predictors in all models Strength total and all domain strength scores significant predictors in all models
		Provincially sentenced adults on community supervision in Alberta between 2009 and 2011 (same sample as Jones et al., 2015)	*N* = 3656 (*n* = 694 female, *n* = 2962 male; *n* = 635 Indigenous, *n* = 3021 non-Indigenous)	18 months	Overall predictive accuracy not reported. Domain risk and strength scores statistically significant predictors with small to medium effects	Risk total and all domain risk scores significant predictors in all models Strength total and all domain strength scores significant predictors in all models

Appendix B

Description of the Service Planning Instrument

(SPIn; Orbis Partners, 2003)

Table 1

SPIn Summary Table with Item Labels

Domain	Overview	Pre-Screen Items	Full Assessment Items
A) Criminal history Static risk: 0-20	Primarily based on information from official sources (i.e. conviction records, police records, etc.) Should begin the session by asking open-ended questions about the event that lead them to their current situation with the CJS (e.g., first contact with the CJS, reasons they were involved, etc.) – elicit a narrative description.	All static items: 1. Age at first arrest 2. Previous adult convictions (1, 2, 3+) 3. Incarcerations as an adult (1, 2, 3+) 4. Delinquency adjudications (i.e. youth convictions) (1, 2, 3+) 5. Incarcerations as a delinquent (1, 2, 3+) 6. Variety of offences a. Assault/violence b. Robbery c. Burglary d. Fraud e. Other property f. Drug g. DWI offenses h. Sex offences	Identical to pre-screen

Domain	Overview	Pre-Screen Items	Full Assessment Items
		i. Other	
B) Response to supervision Static risk: 0-43	Items should be assessed based on official records (conviction records, police records, etc.).	All static items: 1. Failures to appear in court (yes/no) 2. Technical violations while on supervision (yes/no) 3. New offences on supervision (yes/no) 4. Transfer to custody while on supervision (yes/no) 5. Absconded from probation or parole (yes/no) 6. Escaped or attempted escape (yes/ no)	Additional static items: 7. Current institutional disciplinary infractions (0, 1, 2-4, 5+) 8. Previous institutional disciplinary infractions (0, 1, 2-4, 5+) 9. Ever placed in segregation for discipline (yes/no) 10. Any failures from temporary releases (yes/no)
C) Aggression/Violence Static risk: 0-41 Dynamic risk: 0-8 Dynamic strength: 0-8	Violent behaviour includes threats, force, or physical harm caused to another person. Responses to static items largely come from file-based sources but can also include other violence that is documented by another	All static items: 1. Violent behaviour in the last 6 months (yes/no) 2. Previous violent behaviour or convictions including current offence (1, 2, 3+) 3. Any violence towards unknown victims (yes/no)	Additional static items: 6. Any violent behaviour before age 16 (yes/no) 7. Weapon offences (0, 1, 2, 3+) 8. Any history of violence associated with mental disorder (yes/no) 9. Violent behaviour while incarcerated for the current conviction (yes/no/NA)

Domain	Overview	Pre-Screen Items	Full Assessment Items
	credible source (e.g. a social worker). Dynamic items to be coded based on the interview and additional information gathered from other sources. Use open-ended questions to elicit information. Items should be scored based on the presence of attitudes, regardless of whether they have resulted in corresponding behaviours. Should also take into account offending behaviour (e.g. a DV offender may not have aggressive attitudes in other areas of life, but still hold aggressive attitudes re: their partner).	4. Perpetrator of domestic violence (yes/no) 5. Violations of protection or no contact orders (yes/no)	10. Threatening behaviour while incarcerated for the current conviction (yes/no/NA) Dynamic items (all scored on a 5-point scale ranging from – to ++): 11. Anger/frustration tolerance "highly volatile with reputation for fits of anger and rage" to "never gets upset over small things or has tantrums" 12. Believes in use of physical violence to solve conflict/arguments "believes violence is usually the only option for solving conflict" to "believes violence is always wrong" 13. Believes in use of verbal aggression to solve conflict/arguments "believes verbal aggression is usually necessary to solve conflict" to "believes verbal aggression is always wrong" 14. Frequently in conflict with others "constantly in conflict with others" to "never in conflict with others" 15. Motivation to address aggressive behaviour – "uncooperative and unwilling to work on positive

Domain	Overview	Pre-Screen Items	Full Assessment Items
			change" to "actively committed and working on change"
D) Substance use Static risk: 0-9 Dynamic risk: 0-28	Assessment of substance use as it relates to disruptions in the individual's life. Disrupted functioning may be seen in the following areas: employment/education, family conflict, peer relationships, mental/physical health. General guideline: if the use is more than once a week, impairs normal functioning, and is considered to need treatment, then there has been a disruption. Should also consider how their substance use contributes to their behaviour – consider situations where the dominant pattern of offending behaviour is	Dynamic assessment of the: A) Frequency of use (daily, several days per week, rarely, none), and B) How their use impacts functioning (negative weight given to use that disrupts functioning, contributes to criminal behaviour, or if there are indications of use while in custody; positive weight given to attempts to cut back) SUBSTANCES: - Alcohol - Marijuana - Cocaine/crack - Ecstasy, other "club drugs" - Heroin - Hallucinogens - Inhalants - Amphetamines - Methamphetamine - Prescription drug misuse - Other	Static item: 2. Previous substance use treatment (yes/no/NA) Dynamic items: 3. Primary motivation for use: a. N/A – No problem b. Peer pressure c. Coping with stress d. Coping with trauma e. Physical addiction f. Other 4. Motivation to address substance abuse – "uncooperative and unwilling to work on positive change" to "actively committed and working on change" (Rated on a five-point scale ranging from -- to ++)

Domain	Overview	Pre-Screen Items	Full Assessment Items
	influenced by their substance use.		
E) Social influences Dynamic risk: 0-26 Dynamic strength: 0-15	Items assessed throughout the interview by asking open-ended questions. Avoid asking direct questions regarding gang associations.	Dynamic items (each 'check all that apply'): 1. Gang association a. Belongs to a gang b. Associates with gang members c. Family member(s) belong to gangs d. No gang associations 2. Peer relationships a. No consistent peer relationships b. One or more peers with negative, antisocial influence c. One or more peers with antisocial history d. Only antisocial peers	Dynamic items (each 'check all that apply'): 3. Antisocial peers (negative, neutral, or positive) a. Associates with antisocial peers at high risk of frequent or serious offending b. Associates with antisocial peers at lower risk of frequent or serious offending c. Avoids antisocial individuals d. No antisocial peers e. None of the above 4. Social activity (negative, positive) a. Engages in antisocial activities with peers (e.g., crime, drug use, excessive drinking) b. Primarily engages in unconstructive activities with peers (bars, idle time)

Domain	Overview	Pre-Screen Items	Full Assessment Items
		e. One or more peers with positive prosocial influence f. One or more close friends with positive prosocial influence g. None of the above	c. No participation in organized prosocial activities d. Engages in prosocial structured activities (e.g., hobbies, sports, clubs, classes, organized activities, etc.) e. Engages in regularly scheduled prosocial activities f. None of the above 5. Community participation (negative, positive) a. No participation in prosocial community organizations (e.g., church, volunteering, community organizations, committees, etc.) b. Interest in participation in prosocial community organizations c. Participation in prosocial community organizations d. Highly involves in prosocial community organizations 6. Neighbourhood (negative, positive)

Domain	Overview	Pre-Screen Items	Full Assessment Items
			a. Known drug dealers b. Gang activity c. High concentration of criminals d. Fearful of personal safety e. Poor access to resources (medical, recreation, parks) f. No particular neighbourhood concerns g. Positive, prosocial neighbourhood 7. Motivation to address social influences – "uncooperative and unwilling to work on positive change" to "actively committed and working on change" (scored on a 5-point scale ranging from -- to ++)
F) Family Static risk: 0-9 Dynamic risk: 0-26 Static strength: 0-4 Dynamic strength: 0-14	Assessed in the interview using open-ended questions. Meant to assess both risk presented by family situation (unstable/violent dynamics) and protective nature of family (positive contact, closeness,	Dynamic item (scored on a 5-point scale ranging from -- to ++): 1. Marital relationship – "high degree of instability and conflict, offender expresses high dissatisfaction" to "high degree of stability, satisfaction, and	Dynamic items (all scored on a 5-point scale ranging from -- to ++): 4. Par–ing skills – "low interest in parenting role or major parenting skill deficits are recognized or evident" to "confident and proficient parenting skills are evident" 5. Custody arrangements (for non-custodial children) – "expresses major dissatisfaction with

Domain	Overview	Pre-Screen Items	Full Assessment Items
	commitment to maintaining/building healthy relationships)	commitment to the relationship" Mixed item ('check all that apply'): 2. Marital risk factors a. Perpetrated domestic violence (static – all others dynamic) b. Perpetrated domestic violence with current partner c. Expresses safety and protection issues with regard to spouse d. Partner with antisocial history e. Partner has prosocial influence f. N/A or none of the above Dynamic item (rated on a 5-point scale ranging from -- to ++):	custody arrangements" to "highly satisfied with custody arrangements" 6. History in family of origin (check all that apply) a. Absent mother in childhood b. Absent father in childhood c. Primarily raised in a single-parent home d. High marital conflict among parents e. Violence among caregivers f. Victim of physical/sexual abuse g. Kicked out of home h. Frequent conflicts with parents in childhood i. Foster or other placements j. Parental substance abuse problems k. Parental mental disorders l. Strong/positive family of origin m. None of the above 7. Current relationships with family of origin (rated on a 5-point scale ranging from -- to

Domain	Overview	Pre-Screen Items	Full Assessment Items
		3. Attachment to children – "high degree of conflict or absence of positive contact with children" to "highly rewarding relationships with major expressions of commitment to parenting role"	++) – "h– level of conflict in relationship with parent or siblings" to "highly positive relationships with family of origin, commitment to family participation, support for 8. Pro-social models in the family ('check all that apply') a. No prosocial models in the family of origin, marital, or extended family (-) b. No contact with prosocial family models (-) c. Some antisocial family members or members with a criminal history (history = static) (-) d. Accessible prosocial models among family members (+) e. Attachments to prosocial models in family (+) f. None of the above 9. Family involvement during incarceration (STATIC for those who have been released, dynamic for those currently incarcerated) (rated on a 5-point scale from -- to ++) – "no

Domain	Overview	Pre-Screen Items	Full Assessment Items
			contact/visits with any family members" to "regular and frequent contact with one or more family members" 10. Motivation to address family (rated on a 5-point scale from -- to ++) – "cooperative and unwilling to work on positive change" to "actively committed and working on change"
G) Employment Static risk: 0-9 Dynamic risk: 0-14 Static strength: 0-3 Dynamic strength: 0-12	Strictly relating to regular, paid employment with conventional relationship between employer and employee – this excludes "under the table" or other informal work without clear, structured expectations and guidelines.	Mixed item: 1. Employment history a. Unemployed at the time of current offence b. Never employed more than 1 year at a one time c. Never employed more than 6 months d. Frequently quits jobs (i.e. 3 or more times) e. Fired f. Interpersonal conflicts with staff or employers	Static item: 3. Education – check all that apply a. Less than 9th grade b. Less than 12th grade c. Literacy issues d. High school graduate e. Some post-secondary training f. College degree g. Advanced degree Dynamic items (scored on a 5-point scale ranging from – to ++): 4. Employment plans – "no plans for finding employment, no job leads" to "secured job offer to begin or resume work or currently employed" 5. Employment performance – "clear evidence of poor work habits (unreliability, tardiness,

Domain	Overview	Pre-Screen Items	Full Assessment Items
		g. Difficulty finding employment h. None of the above Dynamic item (scored on a 5-point scale ranging from -- to ++): 2. Emp–ment motivation – "expresses no interest in finding employment or remaining employed" to "intrinsically motivated to find and maintain employment, enjoys work"	poor performance)" to "excellent job performance and commitment to superior work" 6. Marketability – "no job skills, lack of job experience, or significant barriers to employment" to "highly attractive candidate with marketable skills or high demand occupation" 7. Job search skills – "poorly informed of labor market opportunities, unaware of job search skills, poor presentation or potential employers" to "confident about job search techniques and can present extremely well to potential employers" 8. Motivation to address employment – "uncooperative and unwilling to work on positive change" to "actively committed and working on change"
H) Attitudes Dynamic risk: 0-14 Dynamic strength: 0-14	This domain requires an overall "situational analysis" that lead to the current CJS contact. Attitudes are not meant	Dynamic items (scored on a 5-point scale ranging from -- to ++): 1. Law-abiding attitudes – "openly admits	Dynamic items (scored on a 5-point scale ranging from -- to ++): 3. Attitude when engaged in criminal activity – "confident and proud of behaviour" to

Domain	Overview	Pre-Screen Items	Full Assessment Items
	to be questioned directly, and instead assessed throughout the interview through open discussion of thoughts and feelings surrounding the offence. Interested in the WHO, WHAT, WHERE, WHEN, WHY, and HOW. Attitudes towards the current offence, victims, other involved individuals, police officers, and the CJS as a whole. Items should be scored based on the interviewer's assessment of the individual, rather than solely based on the individual's expressed attitudes. It is important to also consider behaviour, as this is not	unwillingness to demonstrate law-abiding behaviour" to "clearly positive commitment toward law-abiding behaviour" (about awareness and respect for social rules) 2. Accepts responsibility – "openly accepts or is proud of behaviour" to "voluntarily accepts full responsibility for criminal behaviour (also consider minimizations, justifications, or attempts to blame others for behaviour)	"nervous, afraid, or worried about outcome" 4. Commitment to criminal lifestyle – "expresses total commitment to a criminal lifestyle" to "expresses eagerness to disassociate self with criminal lifestyle" 5. Attitudes toward the criminal justice system – "views all criminal justice authorities with contempt" to "indicates respect for the role of criminal justice authorities" 6. Understands impact of behaviour – "total lack of empathy for harm caused to others" to "fully understands the nature of harm caused to others" 7. Willingness to make amends – "unwilling to make amends" to "eagerly indicates plans for making amends" 8. Readiness to change – "hostile or unwilling to make positive behavioural change" to "actively committed to working on change" 9. Program involvement – "unwilling and hostile toward participation" to "enthusiastic

Domain	Overview	Pre-Screen Items	Full Assessment Items
	always in line with expressed attitudes.		about the effects of participation in programs" 10. Motivation to address attitudes – "uncooperative and unwilling to work on positive change" to "actively committed and working on change"
I) Social/cognitive skills Dynamic risk: 0-18 Dynamic strength: 0-18	Similar offence or situational analysis required as with attitudes domain. This domain is primarily concerned with these skills as they relate to the offending behaviour, but also should be considered in other areas of life (e.g. employment, family, etc.).	Dynamic items (scored on a 5-point scale ranging from -- to ++): 1. Impulsivity – "cannot identify triggers that cause problem behaviours" to "uses self-control techniques to avoid trouble" (primarily concerned with ability to foresee how a situation might play out, and ability to plan to avoid negative outcomes) 2. Hostile attributions – "attributes almost all neutral actions of people as hostile and antagonistic" to "can easily tolerate criticism or hostility directed by others"	Dynamic items (scored on a 5-point scale ranging from -- to ++): 3. Consequential thinking – "does not understand there are consequences of actions" to "acts to obtain good and avoid bad consequences" 4. Social perspective-taking – "unwilling to recognize there can be other points of view" to "can accept other points of view without necessarily agreeing" 5. Problem-solving – "cannot identify when problem behaviours or situations occur" to "can apply appropriate solutions to problems" 6. Behavioural control – "believes criminal/problem behaviour is completely out of his or her control" to "recognizes problem behaviour is controllable and accepts full responsibility"

Domain	Overview	Pre-Screen Items	Full Assessment Items
			7. Positive interpersonal skills – "cannot express needs to others without an element of interpersonal conflict or lack of clarity" to "demonstrates social appeal through positive interpersonal skills" 8. Goal setting/planning – "exhibits no interest or desire to set goals and make plans for the future" to "carefully sets out realistic goals and plans and takes active steps to achieve them" 9. Motivation to address skills – "uncooperative and unwilling to work on positive change" to "actively committed and working on change"
J) Stability Static risk: 0-2 Dynamic risk: 0-13 Dynamic strength: 0-7	This domain guides the identification of obstacles to the offender learning and adopting more prosocial behaviour; also consider how the individual feels regarding their own situation and ability to manage it independently.	Dynamic items 1. Financial a. Reliance on social assistance b. Low income for household needs c. High debt load d. Other financial stressors (e.g., medical expenses)	Dynamic items: 3. Transportation – check all that apply a. Lack of access to needed transportation (for job, education, supervision, or treatment activities/responsibilities) b. Unreliable access to needed transportation c. No driver's licence or ineligible

Domain	Overview	Pre-Screen Items	Full Assessment Items
		e. Lacks health insurance or Medicare eligibility f. Income almost approaches household requirements g. Income reasonably adequate, comfortable financial situation h. None of the above 2. Accommodation a. Lacks realistic plan for accommodation b. Temporary or unstable accommodation arrangements c. Stable accommodation with spouse, family, or others d. Other stable accommodation arrangements	d. Required transportation is reliable and accessible e. Funds accessible for transportation f. None of the above 4. Life skills – check all that apply a. Lacks budgeting skills, banking, etc. b. Difficulties approaching others for service (e.g., accommodation, social assistance) c. Poor hygiene d. Unaware of community support services e. History of homelessness f. Strong skills for independent living g. None of the above 5. Motivation to address stability issues (scored on a 5-point scale ranging from -- to ++) – "u–operative and unwilling to work on positive change" to "actively committed and working on change"

Domain	Overview	Pre-Screen Items	Full Assessment Items
		e. None of the above	
K) Mental health Dynamic risk: 0-2	This domain contains only risk factors (no protective). Interviewer should code based on diagnostic information provided by a professional that is qualified to make a diagnosis (which MAY include the professional completing the assessment). Assessors are encouraged to refer for assessment and treatment if they suspect a disorder where one has not been identified/treated appropriately.	Dynamic items: 1. Presence of a MH condition a. A current condition – serious b. A current condition – serious, with a lack of compliance to treatment c. A current condition – serious, no treatment d. A current condition – stable e. No current mental health problems 2. Homicidal ideation (yes/no) Mixed item: 3. Suicidal ideation a. No ideations	Dynamic items: 5. Other indicators (check all that apply): a. Trauma or victimization as a child b. Trauma or victimization as an adult c. Self-injurious behaviour d. Eating disorders e. Physical abuse f. Sexual abuse g. Other 6. Motivation to address mental health issues (scored on a 5-point scale ranging from -- to ++) – “u–operative and unwilling to work on positive change” to “actively committed and working on change”

Domain	Overview	Pre-Screen Items	Full Assessment Items
		b. Suicidal thoughts c. Suicide attempts Static item: 4. Sexual aggression (any history yes/no)	

Appendix C

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-B (CUREB-B) has granted ethics clearance for the research project described below and research may now proceed. CUREB-B is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

Ethics Protocol Clearance ID: Project # 112550
Research Team: Julie Goodwin (Primary Investigator)
David Robinson (Other)
Dr. Shelley Brown (Research Supervisor)
Project Title: Using a Statistical Item Cross-Validation Approach to Assess the Reliability of the Service Planning Instrument (SPIn) in Alberta
Funding Source (If applicable):
Effective: **March 06, 2020** Expires: **March 31, 2021**.

Please ensure the study clearance number is prominently placed in all recruitment and consent materials: CUREB-B Clearance # 112550.

Restrictions:

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB-B via a Change to Protocol Form. All changes must be cleared prior to the continuance of the research.
3. An Annual Status Report for the renewal of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.
4. A closure request must be sent to CUREB-B when the research is complete or terminated.
5. During the course of the study, if you encounter an adverse event, material incidental finding, protocol deviation or other unanticipated problem, you must complete and submit a Report of Adverse Events and Unanticipated Problems Form, found here: https://carleton.ca/researchethics/forms-and-templates/

Failure to conduct the research in accordance with the principles of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 2nd Edition* and the *Carleton University Policies and Procedures for the Ethical Conduct of Research* may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

Please contact the Research Compliance Coordinators, at ethics@carleton.ca, if you have any questions.

CLEARED BY: **Date: March 06, 2020**

Natasha Artemeva, PhD, Chair, CUREB-B

Janet Mantler, PhD, Vice-Chair, CUREB-B

Appendix D

Research Unit, Strategic Services Branch
Correctional Services Division,
Alberta Justice and Solicitor General
10365 – 97 Street
Edmonton, Alberta
T5J 3W7
Email: JSG-CSD-ResearchUnit@gov.ab.ca
Tel: (780) 422-7022

July 06, 2020

Dr. Shelley Brown
Carleton University
Gender and Crime Laboratory
Department of Psychology
A405 Loeb Building
1125 Colonel By Drive
Ottawa, ON K1S 5B6
Phone: 613-520-2600 ext. 1505 Email: shelley.brown@carleton.ca

Dear Dr. Brown,

I am pleased to inform you that the research projects entitled 'Using a Statistical Item Cross-Validation Approach to Assess the Reliability of the Service Planning Instrument (SPIn) in Alberta' and 'Using a Statistical Item Cross-Validation Approach to Assess the Reliability of the Youth Assessment Screening Instrument (YASI)' have been approved by the Correctional Services Division. Please note this approval is contingent upon adherence to the CSD Policy and Procedures for External Researchers and the approved Research Application.

I wish you all the best with your research and look forward to reading the final report.

Sincerely,

Christie Nicholson
Manager, Strategic Research
Justice and Solicitor General
Strategic Services Branch
Correctional Services Division
Government of Alberta

Tel: 780-422-7022
E-mail: Christie.nicholson@gov.ab.ca

Appendix E

Table 1

SPIn Pre-Screen Logic Statements

Error variable	Hypothesis statement	Conclusion statement
PSError1	IF the participant has had 1 or more delinquency adjudications (A4)	THEN their age at first arrest must be less than 18 (A1)
PSError2	IF the participant has 1 or more previous convictions for violent behaviour (C2)	THEN assault/violence must be indicated in their variety of offences (A6a) AND/OR THEN robbery must be indicated in their variety of offences (A6b) AND/OR THEN sex offences must be indicated in their variety of offences (A6h)
	IF assault/violence is indicated in the participant's variety of offences (A6a) AND/OR IF robbery is indicated in the participant's variety of offences (A6b) AND/OR IF sex offences is indicated in the participant's variety of offences (A6h)	THEN the participant must have 1 or more previous convictions for violent behaviour (C2)
PSError3	IF the participant has 1 or more incarcerations as a delinquent (A5)	THEN the participant must have 1 or more delinquency adjudications (A4) AND

Error variable	Hypothesis statement	Conclusion statement
		THEN the participant must have been first arrested under the age of 16 (A1)
PSError4	IF the participant has 0 delinquency adjudications (A4)	THEN the participant must have 0 incarcerations as a delinquent (A5)
PSError5	IF the participant has been transferred to custody while on supervision (B4)	THEN the participant must have technical violations while on probation AND/OR parole (B2) AND/OR THEN the participant must have committed new offences while on supervision (probation or parole) (B3) AND/OR THEN the participant must have absconded on probation or parole (B5)
PSError6	IF the participant belongs to a gang (E1a) AND/OR IF the participant associates with gang members (E1b)	THEN the participant must have one or more peers with an anti-social influence (E2b) AND/OR THEN the participant must have one or more peers with anti-social history (E2c) AND/OR THEN the participant only has anti-social peers (E2d)
PSError7	IF the participant has perpetrated domestic violence with their current partner (F2b) AND/OR IF the participant expresses safety and protection issues with regard to their spouse (F2c)	THEN their marital relationship should be rated as a risk (F1 scored as either “- - “ or “-“) No error if neutral (0)

Error variable	Hypothesis statement	Conclusion statement
PSError8	IF the participant is indicated as having perpetrated domestic violence (F2a) AND/OR IF the participant has perpetrated domestic violence with their current partner (F2b)	THEN they should be indicated as a perpetrator of domestic violence (C4)
PSError9	IF the participant has been indicated as a perpetrator of domestic violence (C4)	THEN the participant must be indicated as having perpetrated domestic violence (F2a)
PSError10	IF the participant must rely on social assistance (J1a)	THEN they also have a low income for their household needs (J1b)
PSError11	IF the participant has a history of sex offences (A6h)	THEN they must have a history of sexual aggression (K4)
PSError12	IF the participant belongs to a gang (E1a)	THEN the participant's law abiding attitudes should be rated as a risk (H1 scored as either "- -" or "-") No error if neutral (0)
PSError13	IF the participant expresses safety and protection issues with regard to their spouse (F2d)	THEN their marital relationship should be rated as a risk (F1 scored as either "- -" or "-") No error if neutral (0)

Table 2

SPIn Full Assessment Logic Statements

Error variable	Hypothesis statement	Conclusion statement
FAError1	IF the participant has ever been placed in segregation for disciplinary reasons (B9)	THEN the participant must have current institutional disciplinary infractions (B7) AND/OR THEN the participant must have previous institutional disciplinary infractions (B8)
FAError2	IF the participant belongs to a gang (E1a) AND/OR IF the participant associates with gang members (E1b)	THEN the participant must associate with anti-social peers at high risk of frequent or serious offending (E3a)
FAError3	IF the participant participates in pro-social community organizations (E5c) AND/OR IF the participant is highly involved in pro-social community organizations (E5d)	THEN the participant must engage in pro-social structured activities (E4d) AND/OR THEN the participant must engage in regularly scheduled pro-social structured activities (E4e)
	IF the participant engages in pro-social structured activities (E4d) AND/OR IF the participant engages in regularly scheduled pro-social structured activities (E4e)	THEN the participant must participate in pro-social community organizations (E5c) AND/OR THEN the participant must be highly involved in pro-social community organizations (E5d)
FAError4	IF the participant participates in pro-social community organizations (E5c)	THEN the participant must have one or more peers with a positive pro-social influence (E2e)

Error variable	Hypothesis statement	Conclusion statement
	AND/OR IF the participant is highly involved in pro-social community organizations (E5d)	AND/OR THEN the participant must have one or more close friends with a positive pro-social influence (E2f)
FAError5	IF the participant's parenting skills are rated as a risk ('-' or '- -') (F4)	THEN the participant's attachment to their children should be rated as a risk ('-' or '- -') (F3) No error if one or both items are rated as neutral
FAError6	IF the participant experienced trauma or victimization as a child (K5a) AND/OR IF the participant has experienced trauma or victimization as an adult (K5b)	THEN the participant must have witnessed a high degree of conflict among parents (F6d) AND/OR THEN the participant witnessed violence among caregivers (F6e) AND/OR THEN the participant was a victim of physical/sexual abuse (F6f) AND/OR THEN the participant was kicked out of their home (F6g)
FAError7	IF the participant was a victim of physical abuse (K5e) AND/OR IF the participant was a victim of sexual abuse (K5f)	THEN the participant must be a victim of physical/sexual abuse (F6f)

Error variable	Hypothesis statement	Conclusion statement
	IF the participant is a victim of physical/sexual abuse (F6f)	THEN the participant must be a victim of physical abuse (K5e) AND/OR THEN the participant must be a victim of sexual abuse (K5f)
FAError8	IF the participant's job search skills are rated as a risk ('-' or '- -') (G7)	THEN the participant's employment plans should also be rated as a risk ('-' or '- -') (G4) No error if one or both items is rated as neutral
FAError9	IF the participant has some post-secondary training (G3e) AND/OR IF the participant has a college degree (G3f) AND/OR IF the participant has an advanced degree (G3g)	THEN the participant's marketability should be rated as neutral (0) or a strength (G6) ('+' or "++")
FAError10	IF the participant belongs to a gang (E1a)	THEN the participant's attitudes towards the criminal justice system should be rated as a risk ('-' or '- -') (H5) No error if this item is rated as neutral

Table 3

SPIn Full Assessment Grey Error Statements

Error variable	Hypothesis Statement	Conclusion Statement
Grey1	IF the participant's parenting skills are rated as a strength ('+' or '++') (F4)	THEN the participant's attachment to their children should be rated as a strength ('+' or '++') (F3) No error if one or both items are rated as neutral ('0')
Grey2	IF the participant's employment plans are rated as a strength ('+' or '++') (G4)	THEN the participant's employment motivation should also be rated as a strength ('+' or '++') (G2) No error if one or both items are rated as neutral
Grey3	IF the participant has less than 9th grade education (G3a) AND/OR IF the participant has less than 12th grade education (G3b) AND/OR IF the participant has literacy difficulties (G3c)	THEN the participant's marketability must be rated as neutral (0) or a risk ('-' or '- -') (G6)
Grey4	IF the participant's willingness to accept responsibility is rated as a strength ('+' or "++") (H2)	THEN the participant's ability to understand the impact of their behaviour should also be rated as a strength ('+' or "++") (H6) No error if one or both items is rated as neutral
Grey5	IF the participant's impulsivity is rated as a strength ('+' or "++") (I1)	THEN the participant's consequential thinking should also be rated as a strength ('+' or "++") (I3)

Error variable	Hypothesis Statement	Conclusion Statement
		No error if one or both items is rated as neutral
	IF the participant's impulsivity as a risk ('-' or '- -') (I1)	THEN the participant's consequential thinking should also be rated as a risk ('-' or '- -') (I3) No error if one or both items is rated as neutral
	IF the participant's consequential thinking is rated as a strength ('+' or "++") (I3)	THEN the participant's impulsivity should also be rated as a strength ('+' or "++") (I1) No error if one or both items is rated as neutral
	IF the participant's consequential thinking is rated as a risk ('-' or '- -') (I3)	THEN the participant's impulsivity should also be rated as a risk ('-' or '- -') (I1) No error if one or both items is rated as neutral
Grey6	IF the participant's consequential thinking is rated as a strength ('+' or "++") (I3)	THEN the participant's problem-solving should also be rated as a strength ('+' or "++") (I5) No error if one or both items is rated as neutral
	IF the participant's consequential thinking is rated as a risk ('-' or '- -') (I3)	THEN the participant's problem-solving should also be rated as a risk ('-' or '- -') (I5) No error if one or both items is rated as neutral
	IF the participant's problem-solving is rated as a strength ('+' or "++") (I5)	THEN the participant's consequential thinking should also be rated as a strength ('+' or "++") (I3) No error if one or both items is rated as neutral

Error variable	Hypothesis Statement	Conclusion Statement
	IF the participant's problem-solving is rated as a risk ('-' or '- -') (I5)	THEN the participant's consequential thinking should also be rated as a risk ('-' or '- -') (I3) No error if one or both items is rated as neutral
Grey7	IF the participant's hostile attributions are rated as a strength ('+' or "++") (I2)	THEN the participant's social perspective-taking should also be rated as a strength ('+' or "++") (I4) No error if one or both items is rated as neutral
	IF the participant's hostile attributions are rated as a risk ('-' or '- -') (I2)	THEN the participant's social perspective-taking should also be rated as a risk ('-' or '- -') (I4) No error if one or both items is rated as neutral
	IF the participant's social perspective-taking is rated as a strength ('+' or "++") (I4)	THEN the participant's hostile attributions should also be rated as a strength ('+' or "++") (I2) No error if one or both items is rated as neutral
	IF the participant's social perspective-taking is rated as a risk ('-' or '- -') (I4)	THEN the participant's hostile attributions should also be rated as a risk ('-' or '- -') (I2) No error if one or both items is rated as neutral
Grey8	IF the participant's hostile attributions are rated as a strength ('+' or "++") (I2)	THEN the participant's interpersonal skills should also be rated as a strength ('+' or "++") (I7) No error if one or both items is rated as neutral
	IF the participant's hostile attributions are rated as a risk ('-' or '- -') (I2)	THEN the participant's interpersonal skills should

Error variable	Hypothesis Statement	Conclusion Statement
		also be rated as a risk (‘-‘ or ‘- -‘) (I7) No error if one or both items is rated as neutral
	IF the participant’s interpersonal skills are rated as a strength (‘+’ or “++”) (I7)	THEN the participant’s hostile attributions should also be rated as a strength (‘+’ or “++”) (I2) No error if one or both items is rated as neutral
	IF the participant’s interpersonal skills are rated as a risk (‘-‘ or ‘- -‘) (I7)	THEN the participant’s hostile attributions should also be rated as a risk (‘-‘ or ‘- -‘) (I2) No error if one or both items is rated as neutral
Grey9	IF the participant’s social perspective-taking is rated as a strength (‘+’ or “++”) (I4)	THEN the participant’s interpersonal skills should also be rated as a strength (‘+’ or “++”) (I7) No error if one or both items is rated as neutral
	IF the participant’s social perspective-taking is rated as a risk (‘-‘ or ‘- -‘) (I4)	THEN the participant’s interpersonal skills should also be rated as a risk (‘-‘ or ‘- -‘) (I7) No error if one or both items is rated as neutral
	IF the participant’s interpersonal skills are rated as a strength (‘+’ or “++”) (I7)	THEN the participant’s social perspective-taking should also be rated as a strength (‘+’ or “++”) (I4) No error if one or both items is rated as neutral
	IF the participant’s interpersonal skills are rated as a risk (‘-‘ or ‘- -‘) (I7)	THEN the participant’s social perspective-taking should also be rated as a risk (‘-‘ or ‘- -‘) (I4)

Error variable	Hypothesis Statement	Conclusion Statement
		No error if one or both items is rated as neutral

Appendix F

Figure 1

Histogram of SPIn Pre-Screen Errors by Sex and Indigenous Status

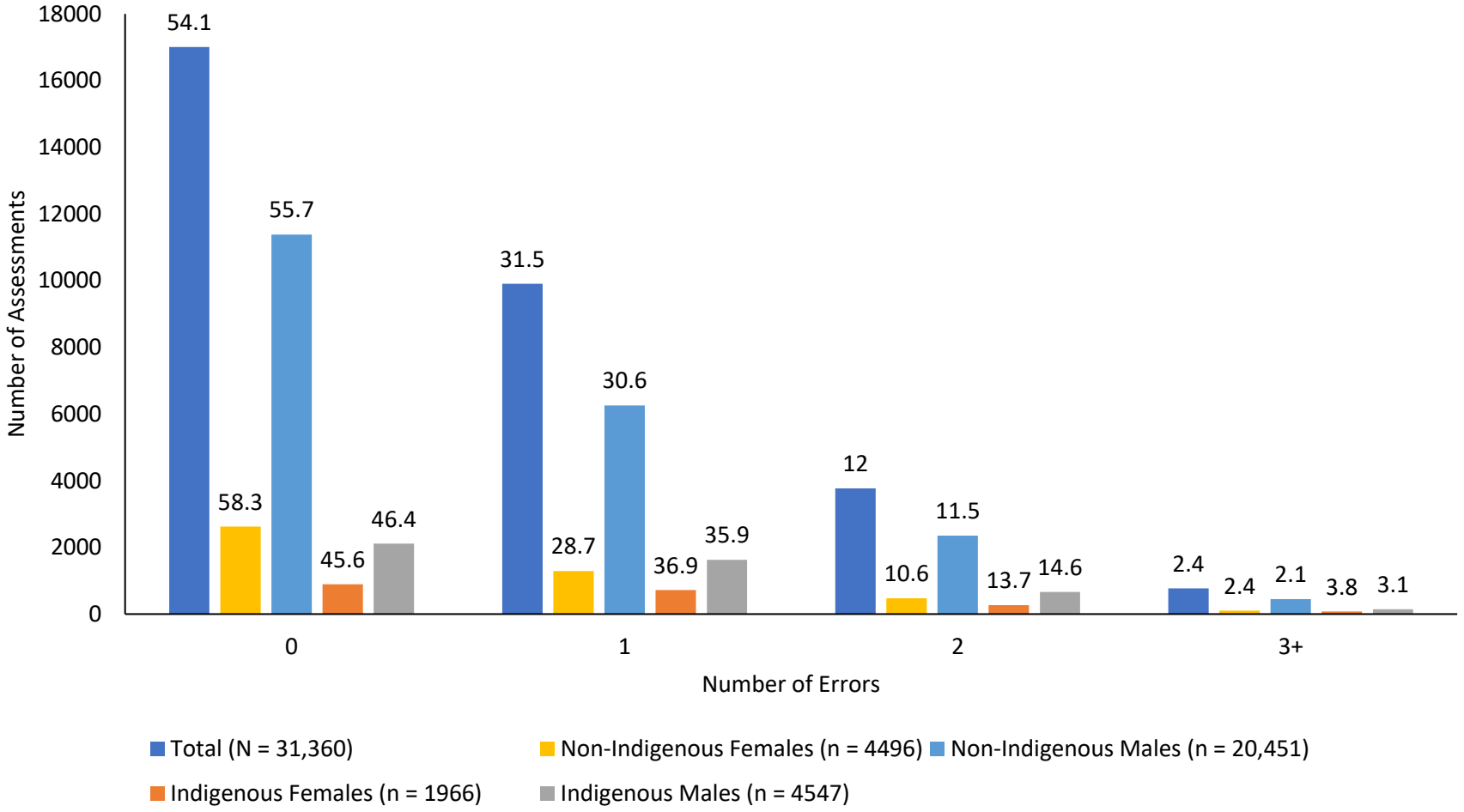


Figure 2

Histogram of SPIn Full Assessment Errors (firm and grey) by Sex and Indigenous Status

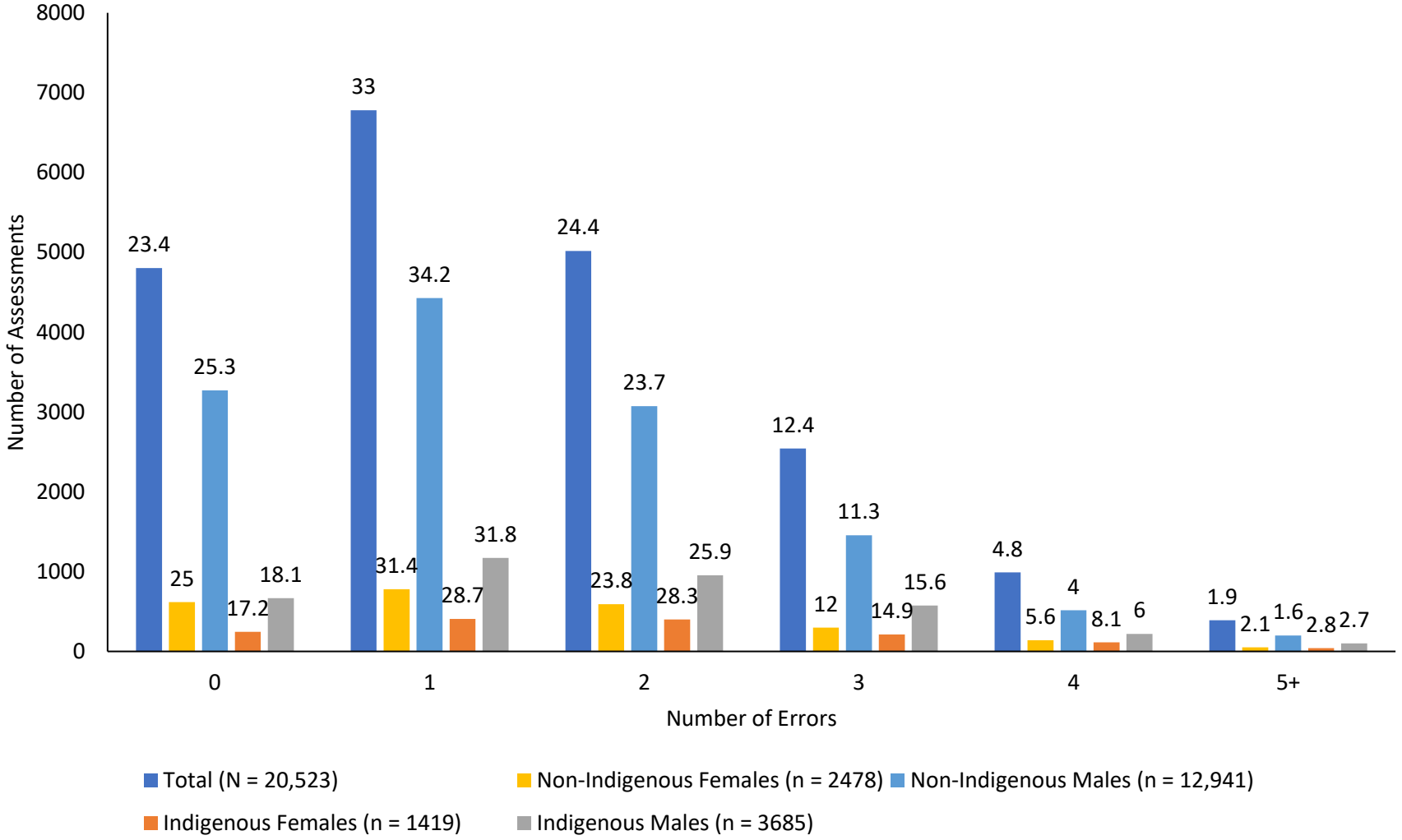


Figure 3

Histogram of SPIn Full Assessment Errors (firm only) by Sex and Indigenous Status

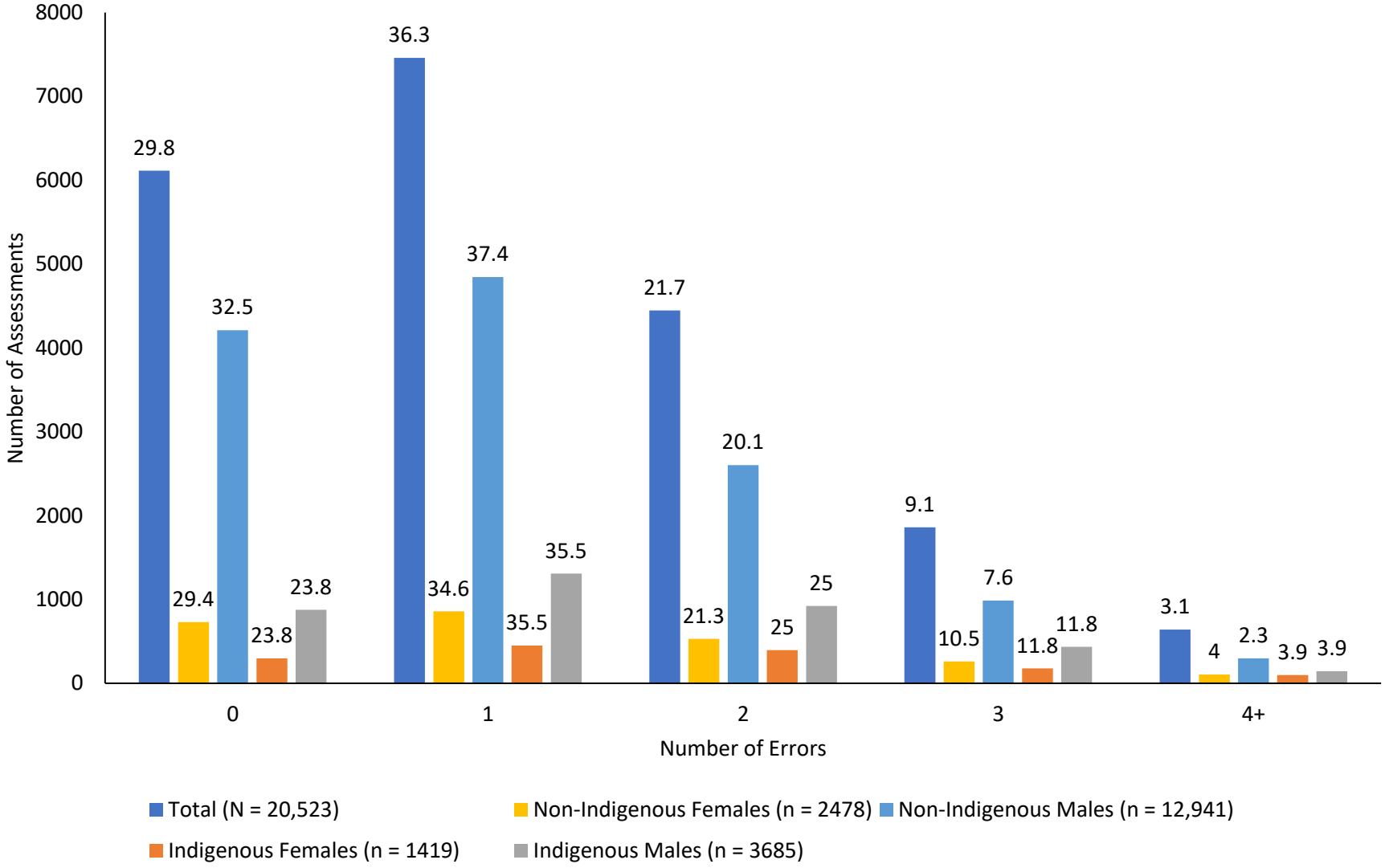


Figure 4

Histogram of Static SPIn Errors by Sex and Indigenous Status

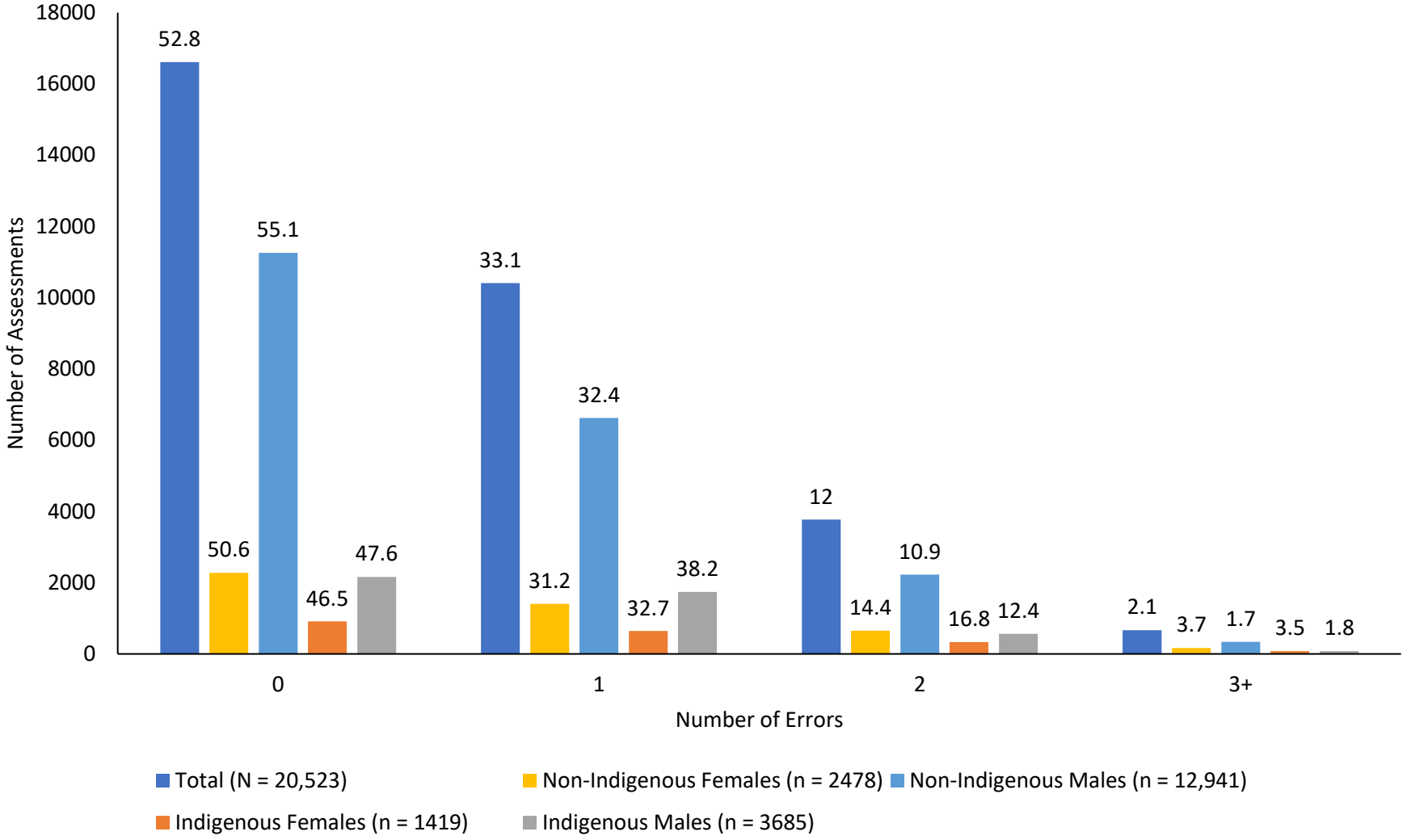


Figure 5

Histogram of Dynamic-Checklist SPIn Errors by Sex and Indigenous Status

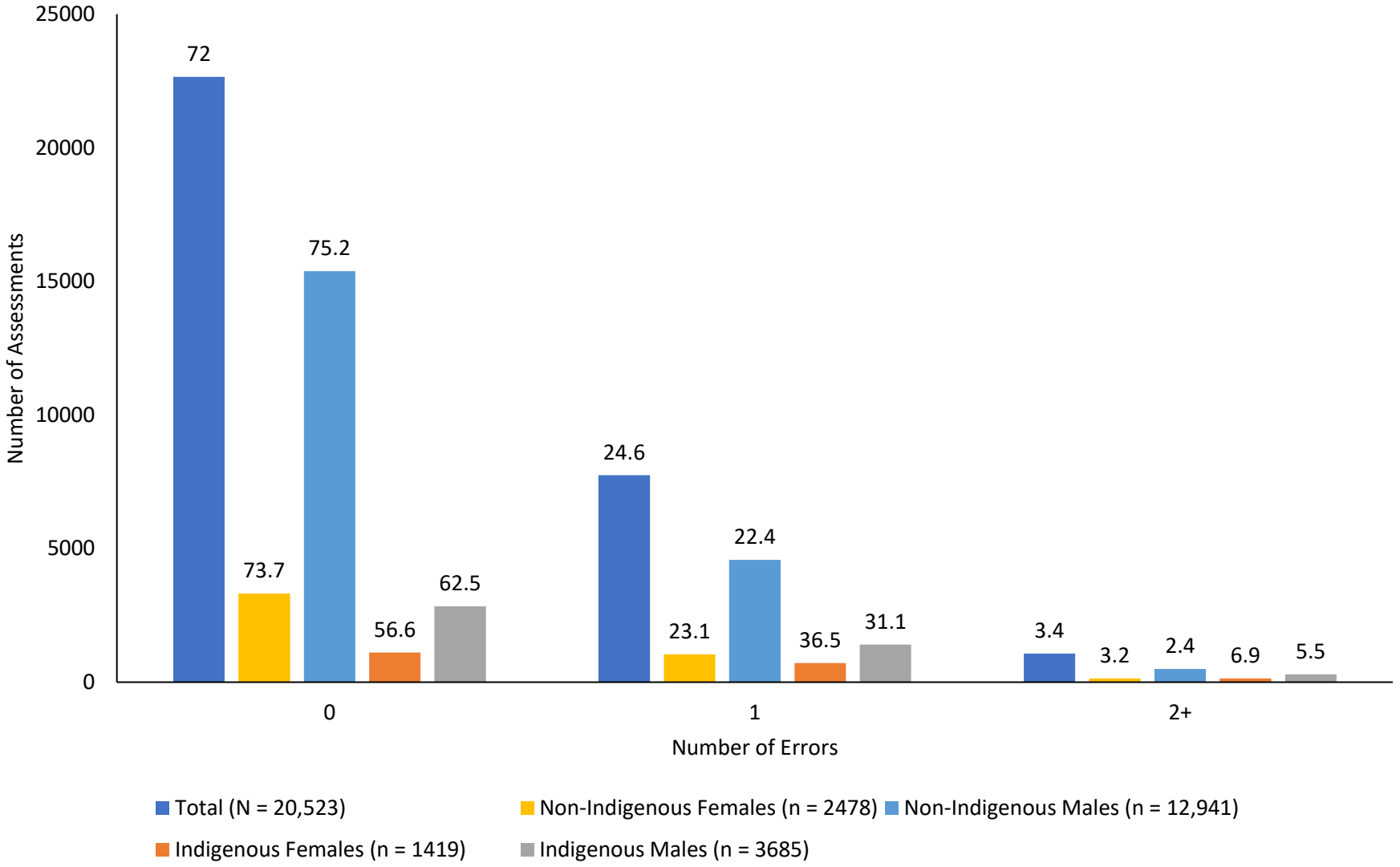


Figure 6

Histogram of Dynamic-Continuum (firm and grey) SPIn Errors by Sex and Indigenous Status

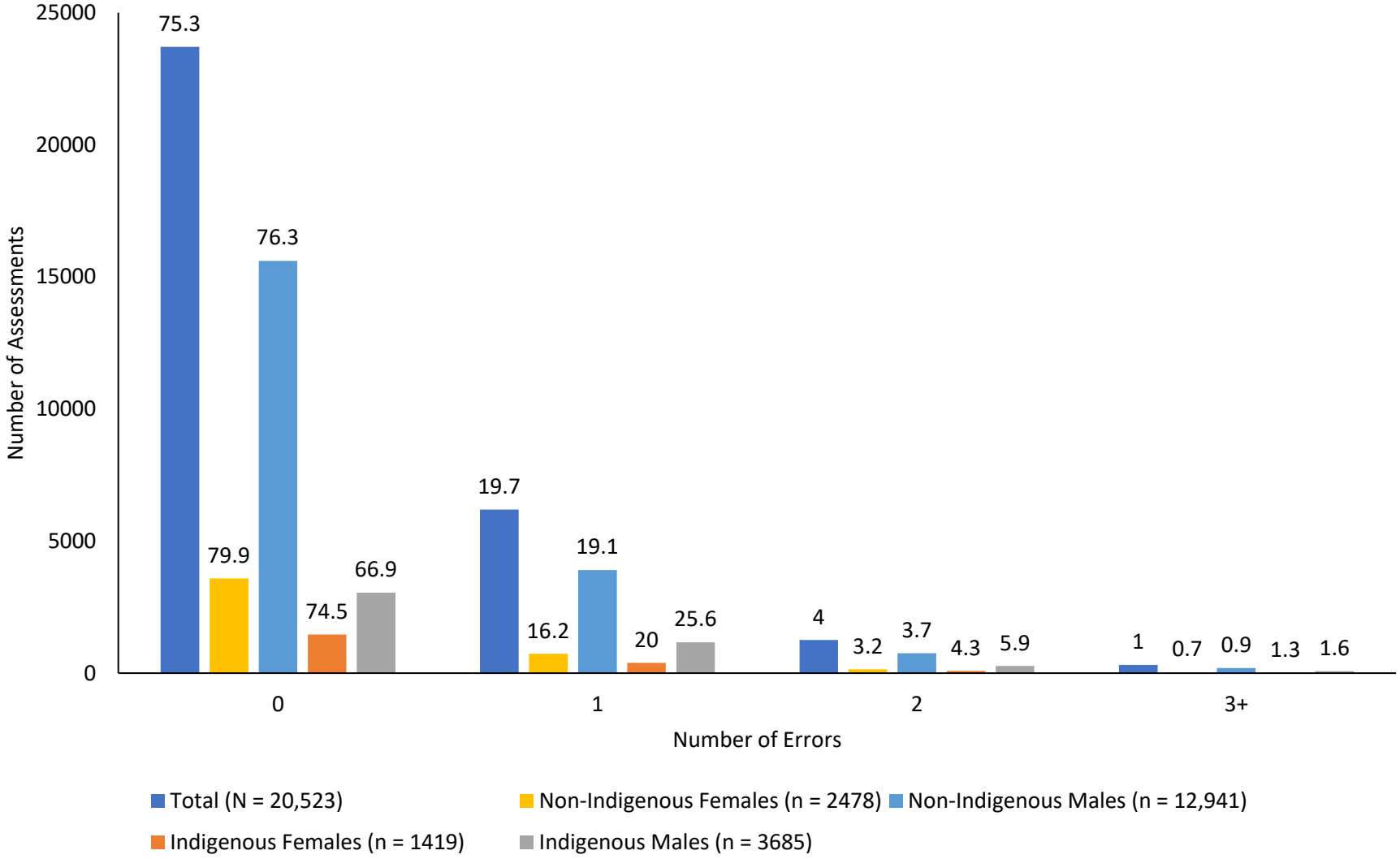
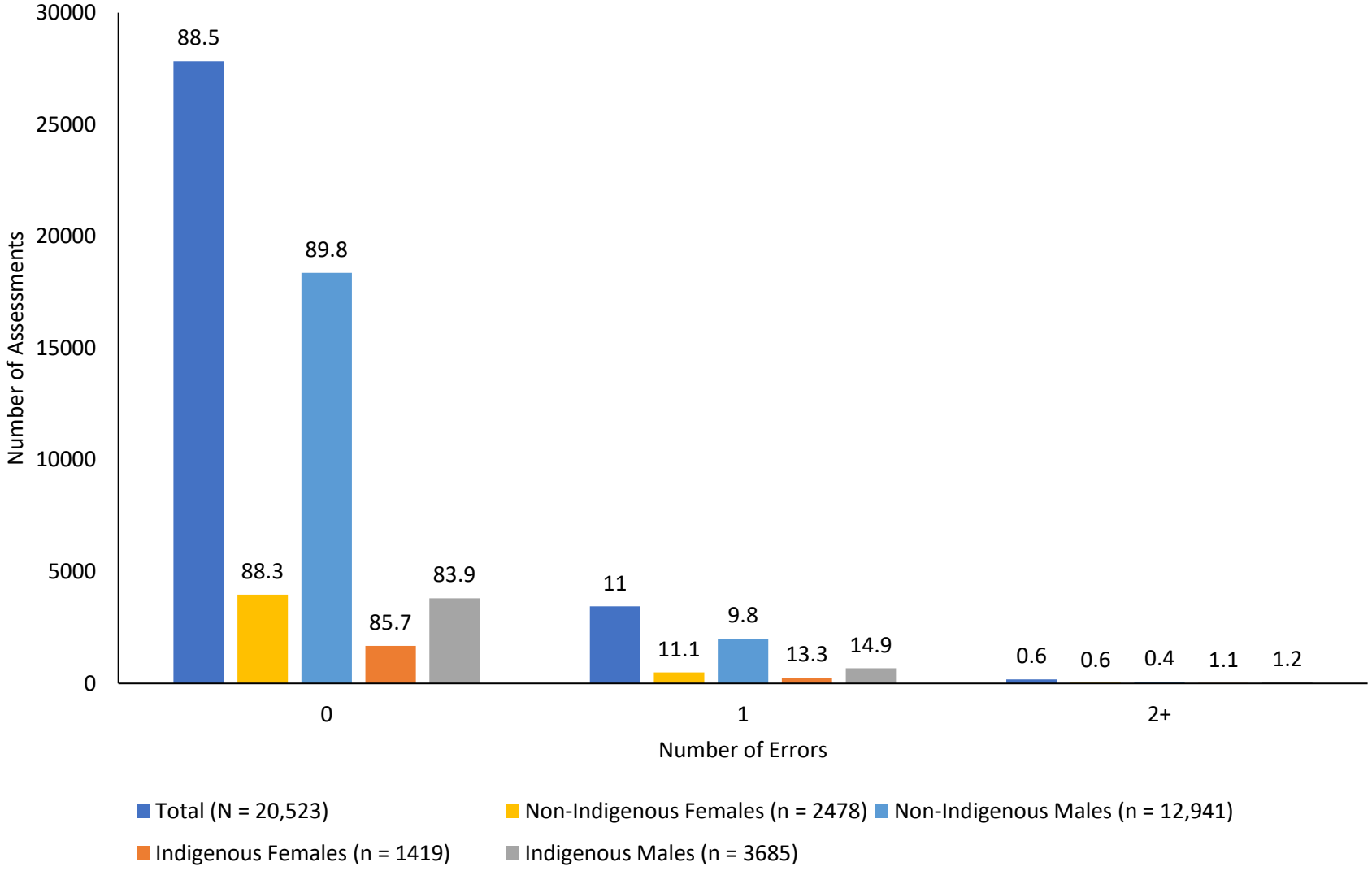


Figure 7

Histogram of Dynamic-Continuum (firm only) SPIn Errors by Sex and Indigenous Status



Appendix G

Table 1

Item-Level Differences in Error Rate by Sex and Indigenous Status

Error Variable	Type	Non-Indigenous			Indigenous		
		Female *n*/4496 (%)	Male *n*/20,451 (%)	χ^2 (φ)	Female *n*/1966 (%)	Male *n*/4547 (%)	χ^2 (φ)
Pre-Screen							
1	Static (firm)	41 (0.9)	268 (1.3)	4.79* (-.01)	29 (1.5)	74 (1.6)	0.21 (-.01)
2	Static (firm)	800 (17.8)	4087 (20.0)	11.23** (-.02)	341 (17.3)	643 (14.1)	10.98** (.04)
3	Static (firm)	20 (0.4)	247 (1.2)	20.26*** (-.03)	21 (1.1)	12 (0.3)	4.77* (-.03)
4	Static (firm)	6 (0.1)	50 (0.2)	2.03 (-.01)	5 (0.3)	12 (0.3)	0.01 (-.01)
5	Static (firm)	7 (0.2)	37 (0.2)	0.13 (-.01)	3 (0.2)	20 (0.4)	3.22 (-.02)
6	Dynamic-Checklist (firm)	21 (0.5)	100 (0.5)	0.04 (-.01)	10 (0.5)	52 (1.1)	5.87* (-.03)
7	Dynamic-Continuum (firm)	254 (5.6)	1458 (7.1)	12.63*** (-.02)	94 (4.8)	427 (9.4)	39.63*** (-.08)
8	Dynamic-Checklist (firm)	52 (1.2)	198 (1.0	1.32 (.01)	45 (2.3)	58 (1.3)	9.06** (.04)
9	Static (firm)	732 (16.3)	4166 (20.4)	39.07*** (-.04)	337 (17.1)	1190 (26.2)	62.35*** (-.10)
10	Dynamic-Checklist (firm)	577 (12.8)	1311 (6.4)	217.38*** (.09)	572 (29.1)	714 (15.7)	155.34*** (.15)
11	Static (firm)	29 (0.6)	336 (1.6)	25.46*** (-.03)	26 (1.3)	64 (1.4)	0.07 (-.01)

12	Dynamic-Continuum (firm)	6 (0.1)	49 (0.2)	1.89 (-.01)	3 (0.2)	49 (1.1)	14.83*** (-.05)
13	Dynamic-Continuum (firm)	26 (0.6)	61 (0.3)	8.32** (.02)	10 (0.5)	16 (0.4)	0.85 (.01)

		Non-Indigenous			Indigenous		
Full Assessment[a]		Female *n*/2471 (%)	Male *n*/12,941 (%)	χ^2 (φ)	Female *n*/1419 (%)	Male *n*/3685 (%)	χ^2 (φ)
1	Static (firm)	7 (0.3)	105 (0.8)	8.07** (-.02)	10 (0.7)	50 (1.4)	3.75 (-.03)
2	Dynamic-Checklist (firm)	43 (1.7)	318 (2.5)	4.74* (-.02)	45 (3.2)	210 (5.7)	13.79*** (-.05)
3	Dynamic-Checklist (firm)	528 (21.3)	3202 (24.7)	13.39*** (-.03)	270 (19.0)	846 (23.0)	9.26** (-.04)
4	Dynamic-Checklist (firm)	51 (2.1)	265 (2.0)	0.01 (< .00)	53 (3.7)	120 (3.3)	0.72 (.01)
5	Dynamic-Continuum (firm)	186 (7.5)	398 (3.1)	112.03*** (.09)	157 (11.1)	213 (5.8)	42.54*** (.09)
6	Static (firm)	323 (13.0)	902 (7.0)	104.59*** (.08)	219 (15.4)	386 (10.5)	24.11*** (.07)
7	Static (firm)	448 (18.1)	985 (7.6)	270.32*** (.13)	283 (19.9)	395 (10.7)	75.68*** (.12)
8	Dynamic-Continuum (firm)	17 (0.7)	63 (0.5)	1.60 (.01)	10 (0.7)	33 (0.9)	0.45 (-.01)
9	Dynamic-Continuum (firm)	53 (2.1)	99 (0.8)	40.21*** (.05)	19 (1.3)	26 (0.7)	4.70* (.03)

10	Dynamic-Continuum (firm)	2 (0.1)	23 (0.3)	1.21 (-.01)	2 (0.1)	24 (0.7)	5.27* (-.03)
11	Dynamic-Continuum (grey)	7 (0.3)	80 (0.6)	4.18* (-.02)	7 (0.5)	25 (0.7)	0.56 (-.01)
12	Dynamic-Continuum (grey)	6 (0.2)	35 (0.3)	0.06 (< -.01)	8 (0.6)	10 (0.3)	2.49 (.02)
13	Dynamic-Continuum (grey)	166 (6.7)	1742 (13.5)	87.71*** (-.08)	89 (6.3)	572 (15.5)	77.76*** (-.12)
14	Dynamic-Continuum (grey)	46 (1.9)	226 (1.7)	0.15 (< .010)	28 (2.0)	41 (1.1)	5.69* (.03)
15	Dynamic-Continuum (grey)	74 (3.0)	413 (3.2)	0.29 (< -.01)	41 (2.9)	106 (2.9)	0.001 (< .01)
16	Dynamic-Continuum (grey)	51 (2.1)	232 (1.8)	0.81 (.01)	36 (2.5)	81 (2.2)	0.53 (.01)
17	Dynamic-Continuum (grey)	21 (0.8))	205 (1.6)	7.82** (-.02)	21 91.5)	44 (1.2)	0.67 (.01)
18	Dynamic-Continuum (grey)	82 (3.3)	433 (3.3)	0.01 (< -.01)	47 (3.3)	134 (3.6)	0.32 (-.01)
19	Dynamic-Continuum (grey)	94 (3.8)	427 (3.3)	1.55 (.01)	51 (3.6)	123 (3.3)	0.20 (.01)

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. [a] Sample sizes are reduced for items in the Full Assessment

Table 2

SPIn Rater Assessment Errors as a Function of Sex of Participant

Error Variable	Type	Total *n*/31,460 (%)	Females *n*/6462 (%)	Males *n*/24,998 (%)	χ^2	φ
Pre-Screen						
1	Static (firm)	412 (1.3)	70 (1.1)	342 (1.4)	3.22	-.01
2	Static (firm)	5871 (18.7)	1141 (17.7)	4730 (18.9)	5.41*	-.01
3	Static (firm)	370 (1.2)	41 (0.6)	329 (1.3)	20.53***	-.03
4	Static (firm)	73 (0.2)	11 (0.2)	62 (0.2)	1.34	-.01
5	Static (firm)	67 (0.2)	10 (0.2)	57 (0.2)	1.30	-.01
6	Dynamic-Checklist (firm)	183 (0.6)	31 (0.5)	152 (0.6)	1.46	-.01
7	Dynamic-Continuum (firm)	2233 (7.1)	348 (5.4)	1885 (7.5)	36.17***	-.03
8	Dynamic-Checklist (firm)	353 (1.1)	97 (1.5)	256 (1.0)	10.53**	.02
9	Static (firm)	6425 (20.4)	1069 (16.5)	5356 (21.4)	75.33***	-.05
10	Dynamic-Checklist (firm)	3174 (10.1)	1149 (17.8)	2025 (8.1)	530.42***	.13
11	Static (firm)	455 (1.4)	55 (0.9)	400 (1.6)	20.21***	-.03
12	Dynamic-Continuum (firm)	107 (0.3)	9 (0.1)	98 (0.4)	9.68**	-.02
13	Dynamic-Continuum (firm)	113 (0.4)	36 (0.6)	77 (0.3)	8.90**	.02
Full Assessment[a]		Total *n*/20,523 (%)	Females *n*/3897 (%)	Males *n*/16,626 (%)	χ^2	φ
1	Static (firm)	172 (0.8)	17 (0.4)	155 (0.9)	9.35**	-.02
2	Dynamic-Checklist (firm)	616 (3.0)	88 (2.3)	528 (3.2)	9.13**	-.02

3	Dynamic-Checklist (firm)	4846 (23.6)	798 (20.5)	4048 (24.3)	26.22***	-.04
4	Dynamic-Checklist (firm)	489 (2.4)	104 (2.7)	385 (2.3)	1.69	.01
5	Dynamic-Continuum (firm)	954 (4.6)	343 (8.8)	611 (3.7)	187.20***	.10
6	Static (firm)	1830 (8.9)	542 (13.9)	1228 (7.7)	147.56***	.09
7	Static (firm)	2111 (10.3)	731 (18.8)	1380 (8.3)	374.15***	.14
8	Dynamic-Continuum (firm)	123 (0.6)	27 (0.7)	96 (0.6)	0.71	.01
9	Dynamic-Continuum (firm)	197 (1.0)	72 (1.8)	125 (0.8)	39.87***	.04
10	Dynamic-Continuum (firm)	51 (0.2)	4 (0.1)	47 (0.3)	4.13*	-.01
11	Dynamic-Continuum (grey)	119 (0.6)	14 (0.4)	105 (0.6)	4.06*	-.01
12	Dynamic-Continuum (grey)	59 (0.3)	14 (0.4)	45 (0.3)	0.86	.01
13	Dynamic-Continuum (grey)	2569 (12.5)	255 (6.5)	2314 (13.9)	156.78***	-.09
14	Dynamic-Continuum (grey)	341 (1.7)	74 (1.9)	267 (1.6)	1.66	.01
15	Dynamic-Continuum (grey)	634 (3.1)	116 (3.0)	519 (3.1)	0.31	-.004
16	Dynamic-Continuum (grey)	400 (1.9)	87 (2.2)	313 (1.9)	2.02	.01
17	Dynamic-Continuum (grey)	291 (1.4)	42 (1.1)	249 (1.5)	3.98*	-.01

18	Dynamic-Continuum (grey)	696 (3.4)	129 (3.3)	567 (3.4)	0.10	-.002
19	Dynamic-Continuum (grey)	695 (3.4)	145 (3.7)	550 (3.3)	1.64	.01

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. [a] Sample sizes are reduced for items in the Full Assessment.

Table 3

SPIn Rater Assessment Errors as a Function of Sex of Participant

Error Variable	Type	Total *n*/31,460 (%)	Non-Indigenous *n*/24,638 (%)	Indigenous *n*/24,998 (%)	χ^2	φ
Pre-Screen						
1	Static (firm)	412 (1.3)	309 (1.2)	103 (1.6)	4.70*	.01
2	Static (firm)	5871 (18.7)	4887 (19.6)	984 (15.1)	68.33***	-.05
3	Static (firm)	370 (1.2)	267 (1.1)	103 (1.6)	11.61***	.02
4	Static (firm)	73 (0.2)	56 (0.2)	17 (0.3)	0.30	.003
5	Static (firm)	67 (0.2)	44 (0.2)	23 (0.4)	7.59**	.02
6	Dynamic-Checklist (firm)	183 (0.6)	121 (0.5)	62 (1.0)	19.47***	.03
7	Dynamic-Continuum (firm)	2233 (7.1)	1712 (6.9)	521 (8.0)	10.12***	.02
8	Dynamic-Checklist (firm)	353 (1.1)	250 (1.0)	103 (1.6)	15.62***	.02
9	Static (firm)	6425 (20.4)	4898 (19.6)	1527 (23.4)	46.17***	.04
10	Dynamic-Checklist (firm)	3174 (10.1)	1888 (7.6)	1286 (19.7)	844.24***	.16
11	Static (firm)	455 (1.4)	365 (1.5)	90 (1.4)	0.24	-.003
12	Dynamic-Continuum (firm)	107 (0.3)	55 (0.2)	52 (0.8)	50.89***	.04
13	Dynamic-Continuum (firm)	113 (0.4)	87 (0.3)	26 (0.4)	0.37	.003
Full Assessment		Total *n*/20,523 (%)	Non-Indigenous *n*/15,419 (%)	Indigenous *n*/5104 (%)	χ^2	φ

1	Static (firm)	172 (0.8)	112 (0.7)	60 (1.2)	9.31**	.02
2	Dynamic-Checklist (firm)	616 (3.0)	361 (2.3)	255 (5.0)	92.83***	.07
3	Dynamic-Checklist (firm)	4846 (23.6)	3730 (24.2)	1116 (21.9)	11.50**	-.02
4	Dynamic-Checklist (firm)	954 (4.6)	316 (2.0)	173 (3.4)	29.61***	.04
5	Dynamic-Continuum (firm)	954 (4.6)	584 (3.8)	370 (7.2)	103.67***	.07
6	Static (firm)	1830 (8.9)	1225 (7.9)	902 (13.8)	72.14***	.06
7	Static (firm)	2111 (10.3)	1433 (9.3)	605 (11.9)	66.15***	.06
8	Dynamic-Continuum (firm)	123 (0.6)	80 (0.5)	43 (0.8)	6.74***	.02
9	Dynamic-Continuum (firm)	197 (1.0)	152 (1.0)	45 (0.9)	0.44	-.01
10	Dynamic-Continuum (firm)	51 (0.2)	25 (0.2)	26 (0.5)	18.66***	.03
11	Dynamic-Continuum (grey)	119 (0.6)	87 (0.6)	32 (0.6)	0.26	.004
12	Dynamic-Continuum (grey)	59 (0.3)	41 (0.3)	18 (0.4)	1.01	.01
13	Dynamic-Continuum (grey)	2569 (12.5)	1908 (12.4)	661 (13.0)	1.16	.01
14	Dynamic-Continuum (grey)	341 (1.7)	272 (1.8)	69 (1.4)	3.99*	-.01
15	Dynamic-Continuum (grey)	634 (3.1)	487 (3.2)	147 (2.9)	0.99	-.01

16	Dynamic-Continuum (grey)	400 (1.9)	283 (1.8)	117 (2.3)	4.19*	.01
17	Dynamic-Continuum (grey)	291 (1.4)	226 (1.5)	65 (1.3)	1.01	-.01
18	Dynamic-Continuum (grey)	696 (3.4)	515 (3.3)	181 (3.5)	0.50	.01
19	Dynamic-Continuum (grey)	695 (3.4)	521 (3.4)	174 (3.4)	0.01	.001

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

Appendix H
Differences in Predictive Accuracy By Error Rate

Table 1

SPIn Pre-Screen Predictive Accuracy Based on Number of in One-Year Follow-Up Sample

	Any Recidivism		Technical Violations		Any New Offence		Any Return to Custody		Any New Conviction		Violent Recidivism	
	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*
0 Errors (*n* = 17,011)	.70***	.005	.72***	.006	.67***	.006	.73***	.006	.62***	.007	.71***	.008
1 Error (*n* = 9909)	.70***	.006	.72***	.007	.69***	.007	.74***	.007	.66***	.008	.69***	.009
2 Errors (*n* = 3767)	.70***	.011	.75***	.011	.68***	.012	.76***	.011	.63***	.015	.67***	.015
3 + Errors (*n* = 773)	.73***	.022	.74***	.023	.72***	.023	.76***	.023	.70***	.031	.72***	.033

Table 2

SPIn Pre-Screen Predictive Accuracy Based on Number of in Three-Year Follow-Up Sample

	Any Recidivism		Technical Violations		Any New Offence		Any Return to Custody		Any New Conviction		Violent Recidivism	
	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*	AUC	*SE*
0 Errors (*n* = 17,011)	.69***	.005	.72***	.005	.67***	.005	.73***	.005	.63***	.005	.70***	.006
1 Error (*n* = 9909)	.71***	.006	.72***	.006	.69***	.006	.75***	.006	.65***	.007	.68***	.008
2 Errors (*n* = 3767)	.68***	.009	.70***	.010	.65***	.010	.72***	.010	.61***	.011	.68***	.012
3 + Errors (*n* = 773)	.71***	.025	.70***	.028	.67***	.029	.73***	.026	.67***	.030	.65***	.0360