Are Adolescent Risk Assessment Tools Sensitive to Change? A Framework and Examination of the SAVRY and the YLS/CMI

Jodi L. Viljoen Catherine S. Shaffer Andrew L. Gray Kevin S. Douglas
Simon Fraser University

Law and Human Behavior

DOI: TBD (in press)

Author Note

Jodi L. Viljoen, Catherine S. Shaffer, Andrew L. Gray, and Kevin S. Douglas, Department of Psychology, Simon Fraser University.

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada, and a Career Investigator Award for the first author from the Michael Smith Foundation for Health Research. The authors would like to thank the youth who participated in this study as well as the many research assistants who assisted with this project.

Correspondence concerning this article should be address to Jodi Viljoen, Department of Psychology, Simon Fraser University, Burnaby, BC V5A 1S6. Contact: jviljoen@sfu.ca

Abstract

Although many adolescent risk assessment tools include an emphasis on dynamic factors, little research has examined the extent to which these tools are capable of measuring change. In this article, we outline a framework to evaluate a tool's capacity to measure change. This framework includes: (1) measurement error and reliable change, and (2) sensitivity (i.e., internal, external, and relative sensitivity). We then used this framework to evaluate the Structured Assessment of Violence Risk in Youth (SAVRY) and Youth Level of Service/Case Management Inventory (YLS/CMI). Research assistants conducted 509 risk assessments with 146 adolescents on probation (101 male, 45 female), who were assessed every 3 months over a 1-year period. Internal sensitivity was partially supported, as a modest proportion of youth showed changes over time. External sensitivity (i.e., the association between change scores and reoffending) was also partially supported. In particular, 22% of the associations between change scores and any and violent reoffending were significant at a 6-month follow-up. However, only one change score (i.e., Peer Associations) remained significant after the Bonferroni correction was applied. Finally, relative sensitivity was not supported, as the SAVRY and YLS/CMI was not more dynamic than the Psychopathy Checklist: Youth Version (PCL:YV). Specifically, the 1-year rank-order stability coefficients for the SAVRY, YLS/CMI, and PCL:YV Total Scores were .78, .75, and .76, respectively. Although the SAVRY and YLS/CMI hold promise, further efforts may help to enhance sensitivity to short-term changes in risk.

Keywords: adolescence, dynamic risk factors, offending, risk assessment, violence

Are Adolescent Risk Assessment Tools Sensitive to Change? A Framework and Examination of the SAVRY and the YLS/CMI

Risk assessment tools for violence and offending have gained widespread use (Singh et al., 2014). Considerable research has shown that these tools can predict subsequent convictions with moderate effect sizes (Skeem & Monahan, 2011; Yang, Wong, & Coid, 2010). However, little research has examined the extent to which risk assessment tools are able to assess changes in risk (Douglas & Skeem, 2005). Measuring change is important, as it might help professionals to better predict and prevent reoffending. Furthermore, attention to change may be particularly important when assessing adolescents, given that adolescence is a period of enormous change (Vincent, Guy, & Grisso, 2012; Viljoen, Cruise, Nicholls, Desmarais, & Webster, 2012).

Thus, in the present study, we examined two widely used adolescent risk assessment tools, the Structured Assessment of Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2006) and Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002), and evaluated the ability of these tools to measure change. The SAVRY and YLS/CMI have demonstrated good predictive validity (Olver, Stockdale, & Wormith, 2009; Olver, Stockdale, & Wormith, 2014; Singh, Grann, & Fazel, 2011), and include an emphasis on dynamic or modifiable factors, such as anger management difficulties, rather than solely historical factors, such as past offending. However, remarkably little research has directly examined the ability of these tools to measure changes in risk.

As a starting point, we outline a framework for evaluating a risk assessment tool's capacity to measure change (see Table 1). This framework draws from other fields such as treatment outcome research (Duff, 2012; Lambert & Vermeersch, 2013). It also draws from prior work in risk assessment (e.g., Douglas & Skeem, 2005; Monahan & Skeem, 2016).

Ability to Measure Change: A Framework for Risk Assessment Tools

In evaluating the extent to which a tool can measure change, three psychometric properties are of particular relevance (Husted, Cook, Farewell, & Gladman, 2000; Riddle & Stratford, 2013; see Table 1): (1) reliability and measurement error (e.g., Can the tool measure change in a reliable manner? When scores change, is this simply measurement error?), (2) sensitivity (e.g., Is the tool sensitive to change? That is, can it detect true changes that have occurred?), and (3) utility (e.g., Does assessing change on the tool aid in intervention planning?). The present study focused on the first two of these three psychometric properties. The third property (i.e., utility) becomes important to examine later on, after basic properties such as reliability and sensitivity to change are better understood.

Measurement Error and Reliable Change

A primary challenge in measuring change is disentangling real change from measurement error (Jacobson & Truax, 1991; see Table 1). For instance, if a youth changes a few points on a risk assessment tool, this may not be reflective of true change. Instead, it might stem from imperfect interrater reliability. As such, in evaluating the ability of a tool to measure changes in risk, an important starting point is to consider measurement error and the extent to which change is reliable (Riddle & Stratford, 2013).

To do so, one approach is to examine the standard error of measurement (SEM). SEM is

calculated using a tool's reliability and its standard deviation (Stratford, Binkley, & Riddle, 1996). SEM can be used, in turn, to estimate minimally detectable change (MDC), or in other words, the smallest real difference a tool can detect, or its error threshold (Beckerman, Vogelaar, Lankhorst, & Verbeek, 1996; Schuck & Zwingmann, 2003). A large MDC or error threshold can be problematic. If a tool's error threshold is large, even large changes on a risk tool can be uninterpretable.

Reliable change indices (RCIs; Jacobson & Truax, 1991) are closely related to MDC but extend this concept to guide the interpretation of change for a given individual. The RCI value is a *z*-value which represents the probability that, for a given individual, change would be observed based on chance alone (Riddle & Stratford, 2013). For instance, if an individual's RCI value was greater than 1.96 it would mean that the probability of obtaining that change score by chance would be less than 5%. As such, we would be able to conclude, with 95% confidence, that this individual showed reliable change.

RCIs, SEM, and MDC are widely used in fields such as the measurement of treatment outcome, neuropsychological functioning, and physical health (e.g., Duff, 2012; Lambert & Vermeersch, 2013; Wise, 2004). However, as of yet, few researchers have applied these concepts to the field of risk assessment (i.e., Draycott, Kirkpatrick, & Askari, 2012; Olver, Beggs Christofferson, & Wong, 2015; Viljoen, Beneteau, et al., 2012). For instance, no published research has investigated reliable change on the YLS/CMI, and only one study, to our knowledge, has examined reliable change on the SAVRY (Viljoen et al., 2015). The authors of that study reported that, after taking into account measurement error, a youth had to show a change of 7 to 8 points on the SAVRY Risk Total Score to be able to confidently classify this change as reliable. It is unknown if this finding will replicate in other samples.

Sensitivity to Change

In evaluating a tool's ability to measure change, another important criterion is sensitivity to change (see Table 1). Although the term sensitivity is sometimes used to refer to the accuracy of diagnostic tests (e.g., sensitivity, specificity), in this paper we use this term to refer to sensitivity or responsiveness *to change*, or in other words, the ability of a measure to detect change (Husted et al., 2000; Wright & Young, 1998).

According to one model, there are two forms of sensitivity to change, namely internal and external (Husted et al., 2000). Internal sensitivity refers to the "ability of a measure to change over a particular pre-specified time frame" (Husted et al., 2000, p. 459). Oftentimes, internal sensitivity is examined by evaluating the extent to which a tool differentiates individuals who have received various levels of treatment. For instance, if youth who received services showed greater reductions in SAVRY and YLS/CMI scores than youth who did not receive services, it would suggest that these tools show internal sensitivity to change. Besides treatment-related change, youth might show changes as a result of life events (e.g., becoming friends with an antisocial peer) or developmental processes (e.g., developing improved impulse control). Thus, risk assessment tools should presumably capture some changes in risk even among youth who do not receive treatment.

Whereas internal sensitivity to change focuses on within-individual or group-level change over time, external sensitivity to change refers to "the extent to which changes in a measure over

a specified time frame relate to corresponding changes in a reference measure" (Husted et al., 2000, p. 459). For instance, if youth were less likely to reoffend when their SAVRY and YLS/CMI risk scores decreased, this could be interpreted as evidence for external sensitivity to change. In other words, external sensitivity relates to whether change shows an expected relationship with an external standard or indicator of change (e.g., reoffense rates).

Also important is the *relative* sensitivity to change of one measure as compared to another measure (Freeman, Walters, Ingram, Salde, Hobart, & Zajicek, 2013). For instance, the SAVRY and YLS/CMI are hypothesized to be measures of dynamic factors, whereas psychopathic features, are considered to be relatively stable personality characteristics (American Psychiatric Association, 2013). As such, one would expect that the SAVRY and YLS/CMI would tap into greater levels of change than a measure of psychopathic features.

Although researchers have not yet explicitly used a framework of internal, external, and relative sensitivity to examine risk assessment tools, some studies have nevertheless examined these aspects of sensitivity. For instance, a number of studies have examined associations between change scores on *adult* risk assessment tools and reoffending (i.e., external sensitivity; e.g., Lewis, Olver, & Wong, 2013; Olver, Christofferson, Grace, & Wong, 2014; Vose, Smith, & Cullen, 2013). Many of these studies have shown that changes in risk scores significantly predict reoffending. Other studies with adult samples have examined associations between treatment and changes in risk scores, such as by comparing pre- and post-treatment scores (i.e., internal sensitivity; e.g., de Vries Robbé, de Vogel, Douglas, & Nijman, 2015; Hogan & Olver, 2016; Michel et al., 2013). Again, these studies often indicate that risk scores decrease over the course of treatment. In contrast, studies on relative sensitivity to change, or whether one tool is more sensitive than another, are rare (see Beggs & Grace, 2011). However, this is important to investigate, as tools' sensitivity to change may vary depending on whether they include factors that are modifiable and causally related to offending (see Monahan & Skeem, 2016).

Despite growing research on adult tools' sensitivity to change, only a small number of studies have examined sensitivity to change in widely used adolescent risk assessment tools, such as the SAVRY and YLS/CMI. In one study, youth on probation showed mean-level decreases in YLS/CMI scores over time, thus providing evidence for the YLS/CMI's internal sensitivity (Clarke, Peterson-Badali, & Skilling, 2016). Also, decreases in total scores and some subscales predicted decreased risk of reoffending, thus providing support for the YLS/CMI's external sensitivity.

However, in other research, the results have been less promising. In one study, youth in a residential sex offending treatment program showed significant decreases in SAVRY Risk Total Scores from admission to discharge (Viljoen et al., 2015). However, reductions in risk did not translate into reductions in reoffending. This may be because this study used a long follow-up period (i.e., 8 years), during which time youth may have continued to show increases or decreases in risk. Another study found that, despite recommendations to regularly reassess risk, reassessing risk with the SAVRY and YLS/CMI did not improve risk predictions (Viljoen et al., 2016). Specifically, reassessments did not expire or show declines in predictive validity over a two-year follow-up period, nor did assessments improve as evaluators gained familiarity with youth via repeated reassessments. This might be because of limitations in tools' sensitivity to change. As such, we examined sensitivity to change in the current study.

Present Study

In the current study, we used the above-described framework to extend research on the ability of the SAVRY and YLS/CMI to measure change. First, we investigated measurement error and reliability of change. Second, we tested sensitivity to change, including: (1) internal sensitivity (i.e., the extent to which SAVRY and YLS/CMI scores changed over time); (2) external sensitivity (i.e., whether adolescents who showed decreases in risk scores were less likely to reoffend); and (3) relative sensitivity (i.e., whether the SAVRY and YLS/CMI detected more change than a measure of psychopathic features).

Given that research and practice guidelines suggest reassessing youth on a routine basis, such as every 3 or 6 months (Vincent et al., 2012; Viljoen, Cruise, et al., 2012), we reassessed adolescents every 3 months over a 1-year period using interview and file information. As probation is the most common disposition given to adolescent offenders (Alam, 2015; Hockenberry & Puzzanchera, 2015), we chose to sample adolescents on probation. We hypothesized that the SAVRY and YLS/CMI would hold promise in measuring change and would detect more change than a measure of psychopathic features, namely the Psychopathy Checklist: Youth Version (PCL:YV; Forth, Kosson, & Hare, 2003).

Method

This manuscript adheres to the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement (Singh, Yang, Mulvey, & the RAGEE Group, 2015), a 50-item reporting checklist which aims to help ensure clear and transparent reporting of methodology.

Participants

Participants included 146 youth on community probation in a large city in Western Canada. All participants had been assessed by research assistants (RAs) on at least two occasions. The mean age at baseline was 16.36 years (*SD* = 1.15 years, range: 12 to 18 years). Most participants were male (69.2%, *n* = 101) and belonged to an ethnic minority group (63.7%, *n* = 93). In particular, 31.5% (*n* = 46) of the sample were Aboriginal (i.e., First Nations, Métis, Inuit), 13.7% (*n* = 20) were Asian, 6.2% (*n* = 9) were Southeast Asian, 6.2% (*n* = 9) were Hispanic, and 4.8% (*n* = 7) were African. With respect to index offenses, 62.3% (*n* = 91) had committed a violent offense and 36.3% (*n* = 53) had committed a property offense. Most youth had no prior charges (77.4%, *n* = 113). Youth received an average of 12.23 treatment contacts (e.g., individual or group therapy; *SD* = 15.72) from baseline to the 3-month follow-up.

Procedure

This study was conducted as part of a larger study on risk assessment (Viljoen et al., 2016). Although that study also includes a 9-month reassessment, we focused on the baseline, 3-, 6-, and 12-month follow-ups, as researchers have recommended reassessments at these time points (Vincent et al., 2012; Viljoen, Cruise, et al., 2012). Ethics approval was obtained from Simon Fraser University and the research site. All data collection methods complied with ethical procedures (i.e., American Psychological Association, 2010, 2013; Canadian Psychological Association, 2000; CIHR, NSERC, & SSHRC, 2014).

Sampling. Youth at 11 probation offices were informed about the study via youth probation officers, study liaisons, posters, and flyers. Of the youth invited to participate (*n* =

508), 32.1% (n = 163) did not meet the following eligibility criteria: (a) adjudicated for an offense and placed on probation; (b) between the ages of 12 and 18 years; and (c) residing in the Regional District (name omitted for blind review). In addition, 24.8% (n = 126) of youth were not interested in participating, and 5.1% (n = 26) could not be reached. Finally, guardian consent could not be obtained in 5.9% (n = 30) of cases and, as such, those youth were unable to participate. Nevertheless, the gender and ethnic composition of our sample were comparable to national and provincial statistics of justice-involved youth (Calverley, Cotter, & Halla, 2010), suggesting our sample is fairly representative on these factors.

RAs. RAs assessed youth at baseline and then every 3 months over a 1-year period. RAs included 11 Master's and Ph.D. students, and 8 undergraduate students who had completed relevant course work and practicums. All raters completed a 3-day workshop on the SAVRY, YLS/CMI, and PCL:YV; this training was led by the first author rather than certified through the test company (i.e., Multi-Health Systems). However, the first author had prior training, including training from some of the tool developers. In addition to didactic training, RAs completed 4 or more practice cases. These practice cases were checked using answer keys that were developed by the first author and the project managers. If the RA's total scores did not fall within 5 points of the answer key, he or she completed additional practice cases.

Assessments. For each assessment, RAs conducted a standardized interview with the youth at a probation office or a quiet public place (e.g., coffee shop) and then examined youths' justice records prior to rating the SAVRY, YLS/CMI, and PCL:YV. Youth were given a stipend of $15 for the baseline assessment and $20 for each follow-up assessment. SAVRY and YLS/CMI total scores and subscales were pro-rated if 10% or fewer items for that score were missing (Hoge & Andrews, 2002), and PCL:YV total scores were pro-rated if 25% or fewer items were missing (Forth et al., 2003). As this study was prospective, RAs were blind to youths' subsequent charges. To examine interrater reliability, 19.2% (n = 28) of the assessments were randomly sampled and coded by a second rater. For these cases, the two raters both attended the interview and reviewed the same file information, but rated tools separately.

Follow-ups. To help minimize missing follow-ups (e.g., Ribisl et al., 1996), RAs maintained contact with participants between follow-ups, made persistent efforts to contact youth, used collateral sources to assist in locating a youth (e.g., parents, service providers), provided flexibility in meeting times and locations, and provided an extra $25 incentive for completing all the follow-ups. Furthermore, if repeated efforts to meet with a youth were unsuccessful, tools were coded based on file information only (8.3% of assessments, n = 42). Of the 163 youth who completed baseline assessments, 10.4% (n = 17) did not complete any follow-ups; these youth were thus excluded from the present study, making our final sample size 146. The youth who were excluded (n = 17) did not significantly differ from the included youth (n = 146) with respect to demographic variables (i.e., age, gender, ethnicity), or offense history (i.e., index offense, prior offenses, p = .16 to .93). Of the included youth, 0.7% (n = 1), 6.2% (n = 9), and 22.6% (n = 33) were missing 3-, 6-, and 12-month follow-up data, respectively. Youth with and without missing follow-up data did not differ significantly in demographic variables (i.e., age, gender, ethnicity) or offense history (i.e., index offense, prior offenses) at any of the follow-ups (p = .09 to .99).

Official reoffending records. Adult and youth offending records were accessed via the

Corrections Network System (CORNET), a province-wide justice database. Records were successfully accessed for all but two youth. Consistent with most adolescent risk assessment studies (Schwalbe, 2008; Viljoen, Mordell, & Beneteau, 2012), we examined charges rather than convictions as official records often underestimate offending (Farrington, Auty, Coid, & Turner, 2013). Violent reoffending was defined as charges for "actual, attempted, or threatened infliction of bodily harm of another person" (Douglas, Hart, Webster, & Belfrage, 2013, pp. 36–37). Any reoffending was defined as any charges (e.g., theft, drug offenses, violent offenses, violations). To test whether changes in risk scores between baseline and the 3-month assessment predicted subsequent reoffending, we examined reoffending in the 6 months that followed the youths' 3-month assessment. We chose this follow-up period as it was proximal to changes in risk. During the follow-up period, 10.4% (n = 15) of youth committed a violent reoffense and 23.4% (n = 34) committed any reoffense. There were no significant gender or ethnic differences in rates of violent or any reoffending during the follow-up (p = .15 to .66)

Measures

The Structured Assessment of Violence Risk in Youth (SAVRY). The SAVRY (Borum et al., 2006) was designed to assess violence risk. It includes 24 risk factors that are rated as Low, Moderate, or High. These risk factors are divided into three risk domains: Historical (e.g., history of violence), Social/Contextual (e.g., peer delinquency), and Individual/Clinical (e.g., negative attitudes). The latter two sections are conceptualized as dynamic (i.e., 58% of factors are putatively dynamic). The SAVRY also includes a Protective Factors section with six dichotomous items (e.g., strong social support). The SAVRY is based on the structured professional judgment (SPJ) model, meaning that, instead of summing scores, evaluators provide a summary risk rating of Low, Moderate, or High risk. For research purposes, however, scores are typically summed to create a Risk Total Score. In a prior meta-analysis, the SAVRY was found to have moderate associations with any and violent reoffending (weighted r [r_w] = .32 and.30, respectively; Olver et al., 2009). In the present study, interrater reliability was excellent for the SAVRY Risk Total Score (ICC = .91 for a two-way random effects model, single raters, absolute agreement, n = 28 cases; Cicchetti, 1994; McGraw & Wong, 1996), good to excellent for section scores (ICCs = .84, .79, .89, .and .70 for Historical, Social/Contextual, Individual/Clinical, and Protective domains, respectively), and good for the summary risk rating (ICC = .64).

Youth Level of Service/Case Management Inventory. The YLS/CMI (Hoge & Andrews, 2002) was developed to assess general recidivism risk. It includes 42 dichotomous items, which are divided into eight subscales (e.g., Family Circumstances/Parenting, Education/ Employment). All of these subscales, except for Prior and Current Offenses, are conceptualized as dynamic (i.e., 90% of the items). In addition to totaling scores, raters make a summary risk rating using their professional judgment. The present study used the YLS/CMI rather than the YLS/CMI 2.0 (Hoge & Andrews, 2011), as this was the version that was available at the time this study was initiated. However, the correlation between total scores for the two versions is very high (r = .99, n = 21 cases; Gray, Viljoen, & Douglas, 2015). In a recent meta-analysis, the YLS/CMI demonstrated moderate associations with any and violent reoffending (r_w = .32 and .26, respectively; Olver, Stockdale, & Wormith, 2014). In the present study, interater reliability was excellent for the YLS/CMI Risk Total Score (ICC = .82 for a two-way random effects model, single raters, absolute agreement, n = 28; Cicchetti, 1994; McGraw & Wong, 1996), fair

to excellent for subscales (ICCs = .90, .54, .79, .75, .58, .60, .87, and .60 for Prior and Current Offenses, Family Circumstances/Parenting, Education/ Employment, Peer Associations, Substance Abuse, Leisure/Recreation, Personality/Behavior, and Attitudes/Orientation, respectively), and good for the summary risk rating (ICC = .71).

Hare Psychopathy Checklist: Youth Version (PCL:YV). The PCL:YV (Forth et al., 2003) is a 20-item rating scale of psychopathic traits. This measure was adapted for adolescents from the Hare Psychopathy Checklist–Revised (PCL-R; Hare, 1991, 2003). Each item is rated on a 3-point scale (i.e., 0, 1, 2), with higher scores indicating a larger number of psychopathy-related traits. In a prior meta-analysis, the PCL:YV was found to have moderate associations with any and violent reoffending (r_w = .28 and.25, respectively; Olver et al., 2009). In the present study, interrater reliability for PCL:YV Total Score fell in the excellent range (ICC = .92 for a two-way random effects model for single raters, absolute agreement, n = 29; Cicchetti, 1994; McGraw & Wong, 1996), and internal consistency was acceptable (alpha = .86). The PCL:YV was completed at baseline and at the 12-month follow-up.

Child and Adolescent Services Inventory. Treatment services were examined using a modified version of the Child and Adolescent Services Inventory (CASA; Burns, Angold, Magruder-Habib, Costello, & Patrick, 1992; Mulvey, Schubert, & Chung, 2007). Youth were asked about various services (e.g., individual therapy), including the number of times they had received that particular service in the past 3 months. To facilitate recall, we used a calendar approach (Glasner & van der Vaart, 2009; Sutton, 2010), wherein RAs asked youth about life events (e.g., birthdays, changes in residences) and recorded this information on a calendar. Youth then referred to this calendar in answering questions about the services they received during the time period. Given that we were interested in treatment-related services rather than other services (e.g., detention, foster care), we created a CASA Treatment Composite score by totaling the number of times youth had received: (1) individual therapy, (2) group therapy, (3) therapy at school, (4) family therapy, and (5) drug and alcohol treatment. The CASA has displayed good psychometric properties (Ascher, Farmer, Burns, & Angold, 1996) and has produced results consistent with official measures of service involvement (Mulvey et al., 2007).

Data Analytic Plan

Measurement error and reliable change. SEM was calculated as $SEM = S\sqrt{(1-r)}$, where *SD* was the standard deviation of baseline scores for the full sample of youth and *r* was the reliability at baseline (Jacobson & Truax, 1991). Consistent with other studies (Draycott et al., 2012; Viljoen et al., 2015), we calculated SEM using interrater reliability, as this is an important form of reliability for risk assessment tools. MDC was calculated as $MDC = 1.96\sqrt{[2(SEM)}$ (Beckerman et al., 2001; Statford et al., 1996). RCIs (95% confidence intervals) were calculated using Jacobson's and Truax's (1991) formula, $RCI = [(X_2 - X_1)/S_{diff}]$, where X_1 was the score at baseline and X_2 was the score at the relevant follow-up. S_{diff} is the standard error of measurement of X_1 and X_2, and was calculated as $S_{diff} = \sqrt{[2(SEM)^2]}$.

Internal sensitivity. Change scores were calculated for each scale as follows: [Change Scores for Risk Scales = Score at Baseline minus Score at Follow-Up] and [Change Score for Protective Scale = Score at Follow-Up minus Score at Baseline]. Thus, higher scores indicated

greater improvements. To test for group mean-level changes in risk scores and summary risk ratings from the baseline assessment to the 3-, 6-, and 12-month follow-ups, we used the Wilcoxon Signed-Rank test, a non-parametric alternative to the paired samples *t*-test (Wilcoxon, 1945). We chose non-parametric statistics because SAVRY and YLS/CMI Risk Total Scores did not have normal distributions (Conover, 1999), as indicated by the Shapiro-Wilk test and visual inspection of quantile-quantile plots (Ghasemi & Zahediasl, 2012). The effect size (ES) for the Wilcoxon test was calculated as follows: $ES = \{4[T - (\frac{R_2 + R_1}{2})]\Big/[n(n+1)]\}$, where R_1 is the sum of the positive ranks, R_2 is the sum of the negative ranks, T is the smaller of the two values (R_1 and R_2), and n is the sample size (Kerby, 2014). In addition, we calculated rank-order stability coefficients (r_s) for the 3-, 6-, and 12-month follow-ups, and examined the strength of associations between the CASA Treatment Composite score and changes in risk scores and summary risk ratings using Spearman's rho correlation coefficients (r_s), a non-parametric correlation (Mroczek, 2007). These analyses were conducted using IBM Statistics ©, Version 22 (IBM Corporation, 2013).

External sensitivity. To determine whether decreases in risk scores and summary risk ratings predicted decreased likelihood of reoffending, we conducted receiver operating characteristic (ROC) analyses (Hanley & McNiel, 1982), and Cox proportional hazards regression (i.e., survival analyses; Cox, 1972). The area under the curve (AUC) of the ROC graph represents the overall predictive accuracy of a tool (i.e., the probability that a randomly selected reoffender has a higher risk score than a randomly selected non-reoffender), whereas Cox proportional hazards regression tests the ability of a risk assessment tool to predict the time to first reoffense. As the base rate of violent reoffending was low (i.e., 10.3%, $n = 15$), we used *penalized* Cox regression models, as these models reduce bias in the estimation of the hazard ratio when base rates are small (Heinze & Schemper, 2001). These analyses were conducted using R (Ploner & Heinze, 2015; Robin et al., 2011; Therneau, 2014).

Relative sensitivity. To examine if the values for the area under the curve (AUC) of the ROC representing the association between change scores and reoffending (Hanley & McNiel, 1982) differed significantly between the SAVRY and the YLS/CMI, we used the DeLong, DeLong, and Clarke-Pearson (1988) test. To test if the SAVRY and YLS/CMI detected different levels of change, we used McNemar's test for paired proportions (McNemar, 1947). Finally, we tested whether stability coefficients for the SAVRY, YLS/CMI, and PCL:YV differed significantly using Raghunathan and colleagues' test for correlated but non-overlapping correlations (Raghunathan, Rosenthal, & Rubin, 1996).

Gender and ethnic differences. Gender differences in stability coefficients for risk total scores (i.e., internal sensitivity) were compared with *z*-tests of differences between independent correlations based on Fisher's transformation (Cohen & Cohen, 1983). Gender differences in associations between change scores on risk total scores and reoffending (i.e., external sensitivity) were tested using penalized moderated Cox proportional hazards regression analyses, in which change scores were mean-centered around zero to help reduce nonessential multicollinearity (Baron & Kenny, 1985). A parallel set of analyses were conducted to examine differences in results for ethnic minority and non-minority (i.e., Caucasian) youth.

Follow-up analyses with imputed data. Although there was no missing data on

SAVRY and YLS/CMI Risk Total Scores for the baseline assessment, rates of missingness at the follow-ups ranged from 4.8% to 30.8%. Based on Little's test, the study variables (i.e., Risk Total Scores, CASA Treatment Composite score, reoffending outcomes) were missing completely at random, χ^2 (208) = 60.37, p = 1.000 (Little, 1998), meaning that missing data is not likely to bias estimates (see Baraldi & Enders, 2010). However, an assumption of purely random missing data is often unrealistic (Enders, 2013). Also, even when data are missing at random, multiple imputation can yield more powerful statistical tests than pairwise or listwise deletion (Baraldi & Enders, 2010). As such, as an additional check, we conducted multiple imputation of risk total scores and reoffending outcomes, using an expectation maximization algorithm (Markov Chain Monte Carlo method; IBM Statistics ©, Version 22, IBM Corporation, 2013). Then, we reran our primary analyses using imputed data. Multiple imputation involves replacing missing data with a number of plausible estimates; analyses are then conducted on each separate data set and are pooled or averaged to create a single set of values (Little, Jorgensen, Lang, & Moore, 2013). To improve estimations of imputed values, we included demographic variables as auxiliary variables (i.e., age, gender, ethnicity, prior charges, CASA Treatment Composite score), and ran the imputation 100 times to maximize generalizability (Enders, 2010).

Results

Measurement Error

On the SAVRY Risk Total Score, MDC was 6.96 points out of a maximum score of 48 (95% confidence interval; see Table 2). On the YLS/CMI Risk Total Score, MDC was 9.02 points out of a maximum score of 42. Floor effects (i.e., the proportion of youth who could not show a decrease that exceeded the error threshold because their score was at the floor or lower end of the scale) and ceiling effects (i.e., the proportion of youth who could not show an increase that exceeded the error threshold because their score was at the ceiling or upper end of the scale) are shown in Table 2. Floor and ceiling effects were rare for the SAVRY Risk Total Score (0.7% and 2.7%, respectively), and relatively uncommon for the YLS/CMI Risk Total Score (12.3% and 2.7%, respectively). However, floor and ceiling effects were common for some subscales. For instance, although the MDC threshold on the Protective Factors section was 2.37, most youth (78.8%) scored 2 points or lower, and thus could not show decreases that exceeded 2.37.

Sensitivity to Change

Internal sensitivity to change. The sample showed mean-level decreases across the follow-ups on several scales that are conceptualized as dynamic (e.g., YLS/CMI Education/Employment; see Tables 3 to 5). They also showed mean-level decreases on the SAVRY and YLS/CMI summary risk ratings at 3 months (z = -2.43, p =.02, and z = -2.38, p = .02, respectively), and on the SAVRY summary risk rating at 6 months (z = -2.60, p =.009). Finally, they showed increases on historical subscales (i.e., SAVRY Historical, YLS/CMI Prior and Current Offenses; see Tables 3 to 5). However, effect sizes were small ($r \leq .20$; Kerby, 2014), and when p-values were corrected for family-wise error using a Bonferroni correction (alpha of .05/17 comparisons at each time point = .003), only the YLS/CMI Prior and Current Offenses subscale remained significant ($p < .001$). Similarly, rates of reliable change were relatively modest, particularly for the 3-month follow-up (i.e., 9.6% and 4.3% for the SAVRY and YLS/CMI Risk Total Scores, respectively; see Table 3). Also, none of the associations between treatment services and change scores were significant (see Table 6).

External sensitivity to change. In some analyses, change between the baseline and the 3-month follow-up significantly predicted reoffending in the 6 months following this change (see Table 7). For instance, decreases in SAVRY and YLS/CMI Risk Total scores significantly predicted a reduced likelihood of any reoffending after controlling for baseline scores. However, in 78% of the analyses, associations between changes in scores and reoffending were not significant (see Tables 7 and 8). For example, changes in summary risk ratings did not significantly predict any or violent reoffending (see Table 7). In addition, after correcting for the number of comparisons via a Bonferroni correction (alpha of .05/16 comparisons for each analysis = .003), the only association that remained significant was Peer Associations and any reoffending. Specifically, youth who showed increased risk in Peer Associations were more likely to engage in any reoffending ($p < .001$; see Table 7).

Relative sensitivity. There were no significant differences in the external sensitivity of the SAVRY and YLS/CMI Risk Total Scores ($Z = 0.56$, $p = .57$ and $Z = 0.43$, $p = .66$ for any and violent reoffending, respectively). With respect to internal sensitivity, the SAVRY Risk Total score detected significantly more reliable change (i.e., decreases and increases combined) than the YLS/CMI Risk Total at the 12-month follow-up ($p = .004$). However, differences in rates of reliable change at the 3- and 6-month follow-ups were not significant ($p = .06$ and .45, respectively). Contrary to expectations, the PCL:YV did not show significantly higher rank-order stability than the SAVRY or YLS/CMI ($z = 0.42$, $p = .70$ and $z = -0.50$, $p = .61$, respectively). Specifically, the 12-month stability of SAVRY, YLS/CMI, and PCL:YV Total Scores was .78, .75, and .76, respectively.

Gender differences in sensitivity. There were no significant gender differences in stability coefficients for SAVRY Risk Total scores at 6- ($z = -0.59$, $p = .56$) and 12-month follow-ups ($z = -0.72$, $p = .47$), nor for YLS/CMI Risk Total scores at 6-month follow-up ($z = -0.72$, $p = .21$). However, stability coefficients for SAVRY Risk Total scores were significantly higher for males compared to females at the 3-month follow-up ($r_s = .91$ and .81, respectively; $z = 2.30$, $p = .02$). In addition, stability coefficients for YLS/CMI Risk Total scores were significantly higher for males compared to females at both the 3-month ($r_s = .85$ and .61, respectively; $z = 2.80$, $p = .01$) and 12-month follow-ups ($r_s = .80$ and .57, respectively; $z = 2.03$, $p = .04$). Gender did not significantly moderate the association between reoffending and change scores (hazard ratio [HR] = 0.95 to 0.99, $p = .74$ to .96).

Ethnic differences in sensitivity. There were no significant differences between ethnic minority and non-minority youth in stability coefficients for SAVRY or YLS/CMI Risk Total scores at the 3- ($z = 1.49$, $p = .14$ and $z = 0.60$, $p = .54$, respectively), 6- ($z = 1.66$ $p = .10$ and $z = -0.76$; $p = .45$), or 12-month follow-ups ($z = 0.91$, $p = .36$ and $z = -0.94$, $= .35$, $p = .79$). Also, ethnicity did not significantly moderate the association between reoffending and change scores (HR = 0.86 to 0.99; $p = .18$ to .95).

Follow-up analyses with imputed data. Based on the pooled estimates from the imputed data sets, 3-month stability coefficients for SAVRY and YLS/CMI Risk Total Scores were .89 and .83, respectively, 6-month stability coefficients were .85 and .75, and 12-month stability coefficients were .76 and .73. Also, the correlations between SAVRY and YLS/CMI Risk Total Scores were -.05 and -.11, respectively, for violent reoffending, and -.10 and -.17 for any reoffending. These results were comparable to the original results (i.e., r_s values of imputed

and non-imputed data were within .02). Thus, missing data did not appear to bias our results.

Follow-up analyses with file-only assessments excluded. As a final check we reran our primary analyses excluding file-only assessments (i.e., 8.3% of assessments). The 3-month stability coefficients for SAVRY and YLS/CMI Risk Total Scores were .86 and .84, respectively, 6-month stability coefficients were .85 and .76, and 12-month stability coefficients were .78 and .76. In addition, the correlations between SAVRY and YLS/CMI Risk Total Scores were -.02 and -.10, respectively, for violent reoffending, and -.12 and -.20 for any reoffending. Again, these results were comparable to the original results (i.e., r_s values of data with and without file-only assessments were within .03). Thus, file-only assessments did not appear to bias our results.

Discussion

Studies have found support for the ability of some adult risk assessment tools, to measure changes in risk (e.g., de Vries Robbé et al., 2015; Hogan & Olver, 2016). However, research on adolescent risk assessment tools is limited. Thus, we evaluated the SAVRY and YLS/CMI's capacity to measure short-term changes in risk among youth on probation using a framework that included: (1) measurement error and reliable change, and (2) sensitivity (i.e., internal, external, and relative sensitivity). The results suggest that although the SAVRY and YLS/CMI are promising, continued efforts may help to further enhance their sensitivity to short-term change among youth on probation.

Overall, level of measurement error appeared acceptable. Specifically, youth's score had to have increased or decreased by 7 points on the SAVRY Risk Total Score (14% of the maximum possible score), and 9 points on the YLS/CMI Risk Total Score (21% of the maximum possible score) in order to conclude that the change was reliable. This is similar to previous research on the SAVRY (Viljoen et al., 2015) and on adult risk assessment tools (Draycott et al., 2012). Thus, as is expected, small changes in SAVRY and YLS/CMI scores may reflect measurement error rather than true change. However, given that ceiling and floor effects were common for some domains (e.g., SAVRY Protective Factors), it may be beneficial to expand the range of possible scores on those domains so that it is possible to capture reliable change among youth who already score quite high or low. In addition, future research should examine means by which to evaluate the measurement error of summary risk ratings; SEM or RCIs cannot be calculated for summary risk ratings, as they are ordinal rather than continuous. Moreover, in general, further work is needed on how to interpret change within an SPJ framework (e.g., SPJ ratings of change).

Internal sensitivity to change (i.e., the ability to detect change over time) was partially supported. Some youth showed reliable increases or decreases in SAVRY and YLS/CMI scores across the follow-up periods. Specifically, at the 12-month follow-up, 8% to 22% of youth showed reliable change on SAVRY and YLS/CMI Risk Total Scores. However, rates of short-term change were more modest than expected. For instance, at the 3-month follow-up, 4% to 10% of youth showed reliable change on the SAVRY and YLS/CMI Risk Total Scores. Also, rank-order stability coefficients were high (r_s = .84 to .89).

One possibility is that the youth in our sample truly were not demonstrating very much change. Indeed, interventions for adolescent offenders are often unavailable or poor in quality

(Haqanee, Peterson-Badali, & Skilling, 2015), and many youth do not receive treatments that address their criminogenic needs (Peterson-Badali, Skilling, & Haqanee, 2015; Singh et al., 2014). Consistent with this possibility, we did not find any significant associations between the treatment youth received and change, suggesting that treatments did not significantly impact risk.

Another possibility is that youth were, in fact, changing but the tools did not fully detect these changes. For instance, the SAVRY and YLS/CMI's short-term sensitivity to change might be attenuated by the time frames for ratings. Specifically, YLS/CMI items are rated based on the youths' "current situation or to conditions that were present during the previous year" (Hoge & Andrews, 2002, p. 40). The SAVRY has similar instructions (e.g., "present during the preceding year;" Borum et al., 2006, p. 18). Thus, future research could test whether short-term sensitivity improves when shorter time frames are used (e.g., past 6 months versus past year).

External sensitivity to change (i.e., associations with an external criterion, namely reoffending) was, again, partially supported. We found some significant associations between change scores and reoffending. For instance, youth who showed decreased risk in Peer Associations were less likely to engage in any reoffending. However, in most cases (i.e., 78% of the analyses), the associations between change scores and reoffending were not significant. For example, changes in summary risk ratings did not significantly predict reoffending. Also, only 2% of the analyses remained significant after a Bonferroni correction was made for the large number of comparisons. Given that prior research is mixed (i.e., Clarke et al., 2016; Viljoen et al., 2015), the SAVRY's and YLS/CMI's external sensitivity to change may vary depending on the context in which tools are used (e.g., probation versus treatment settings). However, further research is needed to identify the particular contexts in which external sensitivity may be strongest.

Finally, with respect to *relative* sensitivity, the SAVRY showed fairly similar sensitivity to change as the YLS/CMI. However, contrary to expectations, neither tool was any more dynamic than the PCL:YV. Specifically, the 1-year rank-order stability of the PCL:YV Total Score was .76 whereas the 1-year stability of the SAVRY and YLS/CMI Risk Total Scores were .78 and .75, respectively. This is higher than the 1-year stability coefficients reported in other studies of psychopathy and personality (e.g., Klimstra et al., 2009). For instance, in a study of serious adolescent offenders, the 1-year stability of scores on the PCL-YV ranged from .50 to .59 (Hawes, Mulvey, Schubert, & Pardini, 2014; see also Lynam et al., 2009). Although these findings could indicate that personality is malleable during adolescence, it also suggests that the SAVRY and YLS/CMI may be less dynamic than presumed.

In interpreting these findings, some limitations are important to note. Although this is one of the first prospective studies to examine the SAVRY and YLS/CMI's sensitivity to change, this design meant we encountered missed follow-ups. Rates of attrition were comparable to previous studies. For instance, in the MacArthur Violence Risk Assessment Study, 83.9% of participants conducted at least one follow-up, and 49.6% completed all of the follow-ups (Monahan et al., 2001). In the present study, 90.4% of potential participants had at least one follow-up, and 70.5% of study participants completed all three of their follow-ups. However, our follow-up rate was not as high as some studies. For instance, in the Pathways to Desistance Study, the average follow-up rate at each time point was 90% (Mulvey et al., 2016).

Another potential limitation of the current study related to power. Given the relatively small sample size and low base rates for reoffending (10 to 23%), statistical methods were selected in an attempt to account for these issues and maximize statistical power (i.e., use of non-parametric tests and penalized regression methods; see Blair & Higgins, 1985; Heinze, 2006). Despite limited power, significance was found for several effect sizes considered small to moderate in magnitude, with a small number remaining significant even after controlling for the familywise error rate.

Given that most participants were male, it was difficult to draw conclusions about gender differences. Moreover, although we compared sensitivity to change for youth from ethnic minority and non-minority groups, we were unable to conduct more refined analyses on any one particular ethnic group (e.g., Indigenous, Southeast Asian), given the small sample sizes. However, this is an important area for future research, especially as tools may not necessarily function equally across groups (Gutierrez, Wilson, Rugge, & Bonta, 2013; Shepherd, Adams, McEntyre, & Walker, 2014). Finally, youth were sampled from all 11 probation offices in the city which may have resulted in different types of services being received, despite there being a common service provider across the offices.

In sum, in light of these mixed findings, further research is needed. Researchers should continue to investigate sensitivity to change for the SAVRY, YLS/CMI, and other tools (e.g., Violence Risk Scale: Youth Version [Wong, Lewis, Stockdale, & Gordon, 2011], the Risk-Sophistication-Treatment Inventory [Salekin, 2004], Short-Term Assessment of Risk and Treatability: Adolescent Version [Viljoen, Nicholls, Cruise, Desmarais, & Webster, 2014]). However, given that measuring change may be difficult, researchers should also identify approaches by which to further improve tools' sensitivity to change (see Table 9 for a list of potential strategies). Finally, it will be important to investigate if assessing changes in risk holds clinical utility, such as whether it enhances professionals' ability to plan treatment. Ultimately, such efforts may help move the field of risk assessment beyond prediction, and closer to effective risk management and prevention.

References

Alam, S. (2015). Youth court statistics, 2013/2014 (Statistics Canada Catalogue no. 85-002-X). Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2015001/article/14224-eng.htm

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, DC: American Psychiatric Association.

American Psychological Association (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68*, 7-19. http://dx.doi.org/10.1037/a0029889

American Psychological Association (2010). Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010). Retrieved from http://www.apa.org/ethics/code/principles.pdf

Ascher, B. H., Farmer, E. M., Burns, B. J., & Angold, A. (1996). The child and adolescent services assessment (CASA) description and psychometrics. *Journal of Emotional and Behavioral Disorders*, *4*, 12-20. http://dx.doi.org/10.1177/106342669600400102

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37. http://dx.doi.org/10.1016/j.jsp.2009.10.001

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology, 54*, 1204-1217. http://dx.doi.org/10.1016/s0895-4356(02)00482-1

Beckerman, H., Roebroeck, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., & Verbeek, A. L. M. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, *10*, 571-578.

Beckerman, H., Vogelaar, T. W., Lankhorst, G. J., & Verbeek, A. L. M. (1996). A criterion for stability of the motor function of the lower extremity in stroke patients using the Fugl-Meyer assessment scale. *Scandinavian Journal of Rehabilitation Medicine*, *28*, 3-8. http://dx.doi.org/10.1682/jrrd.2010.10.0203jsp

Beggs, S. M., & Grace, R. C. (2011). Treatment gain for sexual offenders against children predicts reduced recidivism: A comparative validity study. *Journal of Consulting and Clinical Psychology*, *79*, 182-192. http://dx.doi.org/10.1037/a0022900

Borum, R., Bartel, P. A., & Forth, A. E. (2006). *Manual for the Structured Assessment for Violence Risk in Youth (SAVRY).* Odessa, FL: Psychological Assessment Resources.

Burns, B. J., Angold, A., Magruder-Habib, K., Costello, E. J., & Patrick, M. K. S. (1992). *The Child and Adolescent Services Assessment (CASA)*. Durham, ND: Duke University Medical Center.

Calverley, D., Cotter, A., & Halla, E. (2010). *Youth custody and community services in Canada, 2008/2009* (Statistics Canada Catalogue No. 85-002-X). . Retrieved from http://statcan.gc.ca/pub/85-002-x/2010001/article/11147-eng.htm

Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada (NSERC), & Social Sciences and Humanities Research Council of Canada (SSHRC) (2014). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. Ottawa, ON: Authors.

Canadian Psychological Association (2000). *Canadian Code of Ethics for Psychologists* (3rd ed.). Ottawa, ON: Authors.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Clarke, M. C., Peterson-Badali, M., & Skilling, T. (2016). *The relationship between changes in dynamic risk factors and the predictive validity of risk assessments among youth offenders*. Manuscript submitted for publication.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression and correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Conover, W.J. (1999). *Practical nonparametric statistics*, *3rd edition*. New York, NY: Wiley.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, *34,* 187–220. http://dx.doi.org/10.1007/978-1-4612-4380-9_37

de Vries Robbé, M., de Vogel, V., Douglas, K. S., & Nijman, H. L. (2015). Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in postdischarge community recidivism. *Law and Human Behavior*, *39*, 53-61. http://dx.doi.org/10.1037/lhb0000089

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837-845. http://dx.doi.org/10.2307/2531595

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law*, *11*, 347-383. http://dx.doi.org/10.1037/1076-8971.11.3.347

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing Risk for Violence: User Guide*. Burnaby, BC, Canada: Mental Health, Law, & Policy Institute, Simon Fraser University.

Draycott, S., Kirkpatrick, T., & Askari, R. (2012). An idiographic examination of patient progress in the treatment of dangerous and severe personality disorder: A reliable change index approach. *Journal of Forensic Psychiatry and Psychology*, *23*, 108-124. http://dx.doi.org/10.1080/14789949.2010.481720

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*, 248-261. http://dx.doi.org/10.1093/arclin/acr120

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford.

Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives*, *7*, 27-31. http://dx.doi.org/10.1111/cdep.12008

Farrington, D. P., Jolliffe, D., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Comparing delinquency careers in court records and self-reports. *Criminology*, *41*, 933-958. http://dx.doi.org/10.1111/j.1745-9125.2003.tb01009.x

Fok, C. T., & Henry, D. (2015). Increasing the sensitivity of measures to change. *Prevention Science*, *16*, 978-986. http://dx.doi.org/10.1007/s11121-015-0545-z

Forth, A. E., Kosson, D. S., Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version*. Toronto, ON: Multi Health Systems.

Freeman, J., Walters, R., Ingram, W., Slade, A., Hobart, J., & Zajicek, J. (2013). Evaluating change in mobility in people with multiple sclerosis: Relative responsiveness of four clinical measures. *Multiple Sclerosis Journal*, *19*, 1632-1639. http://dx.doi.org/10.1177/1352458513482373

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*, 486-489. http://dx.doi.org/10.5812/ijem.3505

Glasner, T., & van der Vaart, W. (2009). Applications of calendar instruments in social surveys: A review. *Quality & Quantity*, *43*, 333-349. http://dx.doi.org/10.1007/s11135-007-9129-8

Gray, A. L., Viljoen, J. L., & Douglas, K. D. (2015, March). *Assessing risk and need in male and female adolescent offenders: Predictive validity of the Youth Level of Service/Case Management Inventory.* Paper presented at the Annual Meeting of the American Psychology - Law Society, San Diego, CA.

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

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36. http://dx.doi.org/10.1148/radiology.143.1.7063747

Haqanee, Z., Peterson-Badali, M., & Skilling, T. (2015). Making 'what works' work: Examining probation officers' experiences addressing the criminogenic needs of juvenile offenders. *Journal of Offender Rehabilitation*, *54*, 37-59. http://dx.doi.org/10.1080/10509674.2014.980485

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

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

Hawes, S. W., Mulvey, E. P., Schubert, C. A., & Pardini, D. A. (2014). Structural coherence and temporal stability of psychopathic personality features during emerging adulthood. *Journal of Abnormal Psychology*, *123*, 623-633. http://dx.doi.org/10.1037/a0037078

Heinze, G., & Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics, 57,* 114-119. http://dx.doi.org/10.1111/j.0006-341X.2001.00114.x

Hockenberry, S., & Puzzanchera, C. (2015). *Juvenile court statistics 2013*. Pittsburgh, PA:

National Center for Juvenile Justice. Retrieved from https://www.ojjdp.gov/ojstatbb/njcda/pdf/jcs2013.pdf

Hogan, N. R., & Olver, M. E. (2016). Assessing risk for aggression in forensic psychiatric inpatients: An examination of five measures. *Law and Human Behavior, 40*, 233-243. http://dx.doi.org/10.1037/lhb0000179

Hoge, R. D., & Andrews, D. A. (2002). *The Youth Level of Service/Case Management Inventory manual and scoring key*. Toronto, ON: Multi-Health Systems

Hoge, R. D., & Andrews, D. A. (2011). *Youth Level of Service/case Management Inventory 2.0 (YLS/CMI 2.0): User's Manual*. Toronto, ON: Multi-Health Systems.

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, *15*, 443-445. http://dx.doi.org/10.1037/1040-3590.15.4.443

Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness: A critical review and recommendations. *Journal of Clinical Epidemiology, 53*, 459-468. http://dx.doi.org/10.1177/0049124110366231

IBM Corporation (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corporation.

Jacobson, N. S., & Truax, P. (1991). Clinical significance : A statistical approach to defining meaningful change in psychotherapy research. *Methodological Issues and Strategies in Clinical Research, 59,* 631–648. http://dx.doi.org/10.1037/10109-042

Kerby, D. S. (2014). The simple difference formula: an approach to teaching nonparametric correlation. *Innovative Teaching, 3*, 1.

Klimstra, T. A., Hale, W. I., Raaijmakers, Q. W., Branje, S. T., & Meeus, W. J. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology*, *96*, 898-912. http://dx.doi.org/10.1037/a0014746

Lambert, M. J., & Vermeersch, D. A. (2013). Psychological assessment in treatment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & ... M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 2: Testing and assessment in clinical and counseling psychology* (pp. 213-229). Washington, DC: American Psychological Association.

Lewis, K., Olver, M. E., & Wong, S. P. (2013). The Violence Risk Scale: Predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. *Assessment*, *20*, 150-164. http://dx.doi.org/10.1177/1073191112441242

Little, R. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202. http://dx.doi.org/10.1080/01621459.1988.10478722

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*, 151-162. http://dx.doi.org/10.1093/jpepsy/jst048

Lynam, D. R., Charnigo, R., Moffitt, T. E., Raine, A., Loeber, R., & Stouthamer-Loeber, M. (2009). The stability of psychopathy across adolescence. *Development and Psychopathology*, *21* 1133-1153. http://dx.doi.org/10.1017/S0954579409990083

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46. http://dx.doi.org/10.1037/1082-989x.1.1.30

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153-157. http://dx.doi.org/10.1007/BF02295996

Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Müller-Isberner, R., & ... Hodgins, S. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. *International Journal of Forensic Mental Health*, *12*, 1-13. http://dx.doi.org/10.1080/14999013.2012.760182

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12,* 489-513. http://dx.doi.org/10.1146/annurev-clinpsy-021815-092945

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P., Robbins, P., Mulvey, E., Roth, L., Grisso, T., & Banks, S. (2001). *Rethinking risk assessment. The MacArthur Study of Mental Disorder and Violence*. New York, NY: Oxford University Press.

Mroczek, D. K. (2007). The analysis of longitudinal data in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 543-556). New York, NY: Guilford Press.

Mulvey, E. P., Schubert, C. A., & Chung, H. L. (2007). Service use after court involvement in a sample of serious adolescent offenders. *Children and Youth Services Review, 29*, 518-544. http://dx.doi.org/10.1016/j.childyouth.2006.10.006

Mulvey, E. P., Schubert, C. A., Pitzer, L., Hawes, S., Piquero, A., & Cardwell, S. (2016). An examination of change in dynamic risk of offending over time among serious juvenile offenders. *Journal of Criminal Justice*, *45*, 48-53. http://dx.doi.org 10.1016/j.jcrimjus.2016.02.008

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service Scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*, 156-176. http://dx.doi.org/10.1037/a0035080

Olver, M. E., Beggs Christofferson, S. M., & Wong, S. P. (2015). Evaluation and applications of the clinically significant change method with the Violence Risk Scale-Sexual Offender Version: Implications for risk-change communication. *Behavioral Sciences & the Law*, *33*, 92-110. http://dx.doi.org/10.1002/bsl.2159

Olver, M. E., Christofferson, S. B., Grace, R. C., & Wong, S. P. (2014). Incorporating change information into sexual offender risk assessments using the Violence Risk Scale–Sexual Offender version. *Sexual Abuse: Journal of Research and Treatment*, *26*, 472-499. http://dx.doi.org/10.1177/1079063213502679

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders a meta-analysis of three assessment measures. *Criminal Justice and Behavior*, *36*, 329-353. http://dx.doi.org/10.1177/0093854809331457

Peterson-Badali, M., Skilling, T., & Haqanee, Z. (2015). Examining implementation of risk assessment in case management for youth in the justice system. *Criminal Justice and Behavior*, *42*, 304-320. http://dx.doi.org/10.1177/0093854814549595

Ploner, M., & Heinze, G. (2015). coxphf: Cox regression with Firth's penalized likelihood. *R package version 1.11*.

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178-183. http://dx.doi.org/10.1037/1082-989X.1.2.178

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*, 1-25. http://dx.doi.org/10.1016/0149-7189(95)00037-2

Riddle, D., & Stratford, P. (2013). *Is this change real? Interpreting patient outcomes in physical therapy*. Philadelphia, PA: F. A. Davis Company.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77-84. http://dx.doi.org/10.1186/1471-2105-12-77.

Salekin, R. T. (2004). *Risk-Sophistication-Treatment Inventory*. Lutz, FL: Psychological Assessment Resources.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177. http://dx.doi.org/10.1037/1082-989X.7.2.147

Schuck, P., & Zwingmann, C. (2003). The 'smallest real difference' as a measure of sensitivity to change: A critical analysis. *International Journal of Rehabilitation Research*, *26*, 85-91. http://dx.doi.org/10.1097/00004356-200403000-00015

Schwalbe, C. S. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice and Behavior, 35*, 1367-1381. http://dx.doi.org/10.1177/0093854808324377

Shepherd, S. M., Adams, Y., McEntyre, E., & Walker, R. (2014). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law*, *20*, 281-293. http://dx.doi.org/10.1037/law0000017

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*, 499–513. http://dx.doi.org/10.1016/j.cpr.2010.11.009

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., & ... Otto, R. K. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *The International Journal of Forensic Mental Health*, *13*, 193-206. http://dx.doi.org/10.1080/14999013.2014.922141

Singh, J. P., Desmarais, S. L., Sellers, B. G., Hylton, T., Tirotti, M., & Van Dorn, R. A. (2014). From risk assessment to risk management: Matching interventions to adolescent offenders' strengths and vulnerabilities. *Children and Youth Services Review, 47*, 1-9. http://dx.doi.org/10.1016/j.childyouth.2013.09.015

Singh, J. P., Yang, S., & Mulvey, E. P., & the RAGEE Group (2015). Reporting guidance for violence risk assessment predictive validity studies: The RAGEE Statement. *Law and Human Behavior*, *39*, 15-22. http://dx.doi.org/10.1037/lhb0000090

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, *20*, 38-42. http://dx.doi.org/10.1177/0963721410397271

Stratford, P. W., Binkley, J. M., & Riddle, D. L. (1996). Health status measures: Strategies and analytic methods for assessing change scores. *Physical Therapy*, *76*, 1109-1123. Retrieved from http://ptjournal.apta.org/content/76/10/1109.short

Sutton, J. E. (2010). A review of the life-events calendar method for criminological research. *Journal of Criminal Justice*, *38*, 1038-1044. http://dx.doi.org/10.1016/j.jcrimjus.2010.07.006

Therneau, T. (2014). survival: A package for survival analysis. *R package version 2.37-7.*

Viljoen, J. L., Beneteau, J. L., Gulbransen, E., Brodersen, E., Desmarais, S. L., Nicholls, T. L., & Cruise, K. R. (2012). Assessment of multiple risk outcomes, strengths, and change with the START:AV: A short-term prospective study with adolescent offenders. *International Journal of Forensic Mental Health, 11*, 165–180. http://dx.doi.org/10.1080/14999013.2012.737407

Viljoen, J. L., Cruise, K. R., Nicholls, T. L., Desmarais, S. L., & Webster, C. D. (2012). Taking stock and taking steps: The case for an adolescent version of the short-term assessment of risk and treatability. *International Journal of Forensic Mental Health*, *11*, 135-149. http://dx.doi.org/10.1080/14999013.2012.737406

Viljoen, J. L., Gray, A. L., Shaffer, C., Latzman, N. E., Scalora, M. J., & Ullman, D. (2015). Changes in J-SOAP-II and SAVRY scores over the course of residential, cognitive-behavioral treatment for adolescent sexual offending. *Sexual Abuse: A Journal of Research and Treatment*. Advance online publication, 1-33. http://dx.doi.org/10.1177/1079063215595404

Viljoen, J. L., Gray, A. L., Shaffer, C., Bhanwer, A., Tafreshi, D., & Douglas, K. S. (2016). Does reassessment of risk improve predictions? A framework and examination of the SAVRY and YLS/CMI. *Psychological Assessment*, doi:10.1037/pas0000402

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

Viljoen, J. L., Nicholls, T. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. with contributions by Douglas-Beneteau, J. (2014). *Short-Term Assessment of Risk and Treatability: Adolescent Version (START:AV) - User Guide*. Burnaby, British Columbia, Canada: Mental Health, Law, and Policy Institute.

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). *Risk assessment in juvenile justice: A guidebook for implementation.* Worcester, MA: Models for Change System Reform in Juvenile Justice. Retrieved from http://modelsforchange.net/publications/346

Vose, B., Smith, P., & Cullen, F. T. (2013). Predictive validity and the impact of change in total LSI-R score on recidivism. *Criminal Justice and Behavior*, *40*, 1383-1396. http://dx.doi.org/10.1177/0093854813508916

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80-83 http://dx.doi.org/10.2307/3001968.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*, 50-59. http://dx.doi.org/10.1207/s15327752jpa8201_10

Wong, S., Lewis, K., Stockdale, K. & Gordon, A. (2011). *The Violence Risk Scale – Youth Version*. Unpublished manuscript, Saskatoon, Saskatchewan, Canada.

Wright, J. G., & Young, N. L. (1997). A comparison of different indices of responsiveness. *Journal of Clinical Epidemiology*, *50*, 239-246. http://dx.doi.org/10.1016/S0895-4356(96)00373-3.

Yang, M., Wong, S. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*, 740-767. http://dx.doi.org/10.1037/a0020473

Table 1

Framework for Evaluating the Ability of Risk Assessment Tools to Measure Change

	General Definition	Application to Risk Assessment Tools
Measurement error and reliability	Proportion of a score which is attributable to error (e.g., low interrater reliability; see Schuck & Zwingmann, 2003)	• Measurement error should be sufficiently low so that the tool can reliably detect relatively small changes • Ceiling and floor effects should be rare
Sensitivity		
a) Internal sensitivity	Ability of a measure to detect change over a specified time period (Husted et al., 2000)	• Individuals should show decreases in risk scores and increases in protective scores following an effective treatment • Even without treatment, individuals should presumably show some change in risk over time (e.g., due to maturation or life events)
b) External sensitivity	Extent to which changes in scores meaningfully relate to changes on an external criterion of interest (Husted et al., 2000)	• Individuals should be less likely to reoffend when their risk score decreases, and more likely to reoffend when their risk score increases
c) Relative sensitivity	Whether a tool has better internal or external sensitivity than another tool (Freeman et al., 2013)	• Tools designed to measure change should capture more change than tools designed to measure putatively stable constructs (e.g., psychopathy, static risk factors)
Utility	Extent to which assessing change improves decision-making and outcomes (see Hunsley, 2003)	• Tools designed to measure change should be associated with better intervention plans (e.g., interventions that are better matched to offenders' needs) and possibly even better outcomes (e.g., reduced breaches)

Table 2

Measurement Error and Minimally Detectable Change

Measure	*M* (*SD*)	Maximum Possible Score	Standard Error of Measurement	Minimally Detectable Change	Floor Effect *n* (%)	Ceiling Effect *n* (%)
SAVRY						
Historical	10.66 (4.01)	20	1.58	4.39	9 (6.2)	21 (14.4)
Social/Contextual	6.24 (2.42)	12	1.09	3.03	22 (15.2)	29 (20.0)
Individual/Clinical	8.79 (3.54)	16	1.17	3.25	13 (9.0)	23 (15.9)
Protective	1.40 (1.57)	6	0.85	2.37	115 (78.8)	17 (11.6)
Risk Total Score	25.71 (8.55)	48	2.51	6.96	1 (0.7)	4 (2.7)
YLS/CMI						
Prior/Current Offenses	1.23 (1.47)	5	0.46	1.26	94 (64.4)	14 (9.6)
Family/Parenting	2.90 (1.74)	6	1.18	3.28	82 (56.6)	86 (59.3)
Education/Employment	2.68 (1.73)	7	0.80	2.21	79 (54.1)	23 (15.8)
Peer Associations	3.02 (1.28)	4	0.64	1.77	20 (13.7)	92 (63.0)
Substance Abuse	2.66 (1.56)	5	1.01	2.80	63 (43.2)	83 (56.8)
Leisure/Recreation	1.72 (1.02)	3	0.64	1.78	51 (34.9)	95 (65.1)
Personality/Behavior	3.07 (1.65)	7	0.59	1.63	28 (19.2)	12 (8.2)
Attitudes/Orientation	1.91 (1.24)	5	0.79	2.18	102 (69.9)	44 (30.1)
Risk Total Score	19.20 (7.59)	42	3.25	9.02	18 (12.3)	4 (2.7)

Note. *M* = mean; *SD* = standard deviation. The sample size was 145 for Social/Contextual, Individual/Clinical, and Family Circumstances, and 146 for the remaining scores. Percentages reported are the valid percentages. The floor effect reflects the number of youth in our sample that could not show reliable decreases in scores as they were already at the floor. The ceiling effect reflects the number of youth that could not show reliable increases as they were already at the ceiling.

Table 3

Change in Scores from Baseline to the 3-Month Follow-Up

		Baseline	3-Month	Group Mean-Level Change		Individual Rank-Order Stability	Proportion of Youth Showing Reliable Change		
Measure	*n*	*M* (*SD*)	*M* (*SD*)	*Z*	*ES*	r_s	Reliable Decrease *n* (%)	Reliable Increase *n* (%)	No Change *n* (%)
SAVRY									
Historical	142	10.54 (3.96)	10.78 (4.03)	-2.69**	.06	.96***	2 (1.4)	2 (1.4)	138 (97.2)
Social/Contextual	129	6.13 (2.42)	5.82 (2.51)	-1.94	-.11	.77***	7 (5.4)	0 (0.0)	122 (94.6)
Individual/Clinical	132	8.58 (3.56)	8.01 (3.63)	-2.63**	-.18	.79***	10 (7.6)	3 (2.3)	119 (90.2)
Protective	139	1.42 (1.57)	1.52 (1.68)	-1.05	.03	.78***	2 (1.4)	2 (1.4)	135 (97.1)
Risk Total Score	135	25.42 (8.57)	24.66 (8.84)	-2.10*	-.16	.89***	10 (7.4)	3 (2.2)	122 (90.4)
YLS/CMI									
Prior/Current Offenses	144	1.22 (1.47)	1.51 (1.64)	-4.25***	.08	.84***	0 (0.0)	13 (9.0)	131 (91.0)
Family/Parenting	139	2.86 (1.76)	2.90 (1.68)	-0.58	.02	.78***	0 (0.0)	0 (0.0)	139 (100.0)
Education/Employment	139	2.63 (1.74)	2.37 (1.57)	-2.18*	-.09	.69***	9 (6.5)	1 (0.7)	129 (92.8)
Peer Associations	139	2.98 (1.30)	2.78 (1.36)	-2.00*	-.04	.66***	17 (12.2)	9 (6.5)	113 (81.3)
Substance Abuse	138	2.62 (1.57)	2.49 (1.59)	-1.26	.04	.71***	5 (3.6)	3 (2.2)	130 (94.2)
Leisure/Recreation	140	1.71 (1.03)	1.75 (1.01)	-0.66	.02	.65***	6 (4.3)	6 (4.3)	128 (91.4)
Personality/Behavior	139	3.04 (1.68)	2.88 (1.75)	-1.35	-.05	.74***	16 (11.5)	10 (7.2)	113 (81.3)
Attitudes/Orientation	142	1.90 (1.25)	1.96 (1.36)	-0.84	-.03	.70***	3 (2.1)	1 (0.7)	138 (97.2)
Risk Total Score	139	19.00 (7.71)	18.59 (7.94)	-0.86	.06	.84***	4 (2.9)	2 (1.4)	133 (95.7)

Note. *M* = mean; *SD* = standard deviation; r_s = Spearman's rho stability coefficient; *Z* = *Z*-coefficient from Wilcoxon Signed-Rank test (Z-coefficients are always negative); *ES* = effect size of Wilcoxon Signed-Rank test. Percentages reported are the valid percentages for the analyses with missing data. * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 4

Change in Scores from Baseline to the 6-Month Follow-Up

		Baseline	6-Month	Group Mean-Level Change		Individual Rank-Order Stability	Proportion of Youth Showing Reliable Change		
Measure	*n*	*M* (*SD*)	*M* (*SD*)	*Z*	*ES*	r_s	Reliable Decrease *n* (%)	Reliable Increase *n* (%)	No Change *n* (%)
SAVRY									
Historical	135	10.61 (4.03)	10.84 (4.16)	-2.34*	.08	.95***	2 (1.5)	2 (1.5)	131 (97.0)
Social/Contextual	114	6.04 (2.33)	5.66 (2.53)	-2.11*	-.13	.74***	6 (5.3)	1 (0.9)	107 (93.9)
Individual/Clinical	116	8.32 (3.52)	7.83 (3.58)	-1.73	-.15	.63***	12 (10.3)	12 (10.3)	92 (79.3)
Protective	128	1.52 (1.60)	1.38 (1.53)	-0.98	-.04	.52***	4 (3.1)	9 (7.0)	115 (89.8)
Risk Total Score	118	25.10 (8.70)	24.41 (8.90)	-1.50	-.13	.86***	11 (9.3)	4 (3.4)	103 (87.3)
YLS/CMI									
Prior/Current Offenses	135	1.22 (1.46)	1.50 (1.63)	-3.74***	.09	.85***	1 (0.7)	16 (11.9)	118 (87.4)
Family/Parenting	128	2.89 (1.76)	2.79 (1.64)	-0.57	-.03	.72***	0 (0.0)	2 (1.6)	126 (98.4)
Education/Employment	128	2.63 (1.75)	2.20 (1.50)	-3.18**	-.18	.54***	12 (9.4)	3 (2.3)	113 (88.3)
Peer Associations	128	3.02 (1.30)	2.87 (1.38)	-1.16	-.02	.55***	17 (13.3)	13 (10.2)	98 (76.6)
Substance Abuse	127	2.55 (1.54)	2.35 (1.61)	-1.76	-.08	.66***	8 (6.3)	3 (2.4)	116 (91.3)
Leisure/Recreation	125	1.66 (1.02)	1.78 (0.99)	-1.60	.06	.58***	6 (4.8)	6 (4.8)	113 (90.4)
Personality/Behavior	121	2.92 (1.64)	2.73 (1.75)	-1.47	-.09	.60***	22 (18.2)	11 (9.1)	88 (72.7)
Attitudes/Orientation	129	1.83 (1.19)	1.84 (1.33)	-0.16	.01	.53***	4 (3.1)	1 (0.8)	124 (96.1)
Risk Total Score	123	18.55 (7.60)	17.85 (7.96)	-1.30	-.11	.77***	8 (6.5)	4 (3.3)	111 (90.2)

Note. *M* = mean; *SD* = standard deviation; r_s = Spearman's rho stability coefficient; *Z* = *Z*-coefficient from Wilcoxon Signed-Rank test (Z-coefficients are always negative); *ES* = effect size of Wilcoxon Signed-Rank test. Percentages reported are the valid percentages for the analyses with missing data. * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 5

Change in Scores from Baseline to the 12-Month Follow-Up

		Baseline	12-Month	Group Mean-Level Change		Individual Rank-Order Stability	Proportion of Youth Showing Reliable Change		
Measure	*n*	*M* (*SD*)	*M* (*SD*)	*Z*	*ES*	r_s	Reliable Decrease *n* (%)	Reliable Increase *n* (%)	No Change *n* (%)
SAVRY									
Historical	113	10.61 (3.98)	11.17 (4.26)	-3.24**	.17	.92***	0 (0.0)	4 (3.5)	109 (96.5)
Social/Contextual	95	6.04 (2.27)	5.52 (2.64)	-2.41*	-.20	.68***	6 (6.3)	0 (0.0)	89 (93.7)
Individual/Clinical	99	8.70 (3.49)	7.85 (3.79)	-2.25*	-.20	.60***	16 (16.2)	6 (6.1)	77 (77.8)
Protective	106	1.44 (1.56)	1.30 (1.58)	-1.07	-.04	.66***	1 (0.9)	4 (3.8)	101 (95.3)
Risk Total Score	101	25.06 (8.36)	24.31 (8.96)	-1.27	-.12	.78***	13 (12.9)	9 (8.9)	79 (78.2)
YLS/CMI									
Prior/Current Offenses	113	1.32 (1.53)	1.90 (1.80)	-4.96***	.20	.75***	0 (0.0)	18 (15.9)	95 (84.1)
Family/Parenting	110	2.95 (1.70)	2.97 (1.70)	-0.25	.01	.67***	1 (0.9)	1 (0.9)	108 (98.2)
Education/Employment	104	2.75 (1.71)	2.29 (1.55)	-2.41*	-.17	.41***	12 (11.5)	5 (4.8)	87 (83.7)
Peer Associations	106	3.02 (1.29)	2.84 (1.30)	-1.51	-.04	.62***	10 (9.4)	7 (6.6)	89 (84.0)
Substance Abuse	107	2.73 (1.50)	2.77 (1.57)	-0.56	.03	.63***	6 (5.6)	3 (2.8)	98 (91.6)
Leisure/Recreation	111	1.69 (1.01)	1.88 (0.98)	-2.13*	.09	.55***	4 (3.6)	6 (5.4)	101 (91.0)
Personality/Behavior	103	3.03 (1.64)	2.97 (1.83)	-0.24	-.02	.64***	13 (12.6)	15 (14.6)	75 (72.8)
Attitudes/Orientation	109	1.93 (1.15)	1.88 (1.35)	-0.46	-.02	.57***	2 (1.8)	0 (0.0)	107 (98.2)
Risk Total Score	100	19.06 (7.48)	19.19 (8.39)	-0.21	.02	.75***	4 (4.0)	4 (4.0)	92 (92.0)

Note. *M* = mean; *SD* = standard deviation; r_s = Spearman's rho stability coefficient; *Z* = *Z*-coefficient from Wilcoxon Signed-Rank test (Z-coefficients are always negative); *ES* = effect size of Wilcoxon Signed-Rank test. Percentages reported are the valid percentages for the analyses with missing data. * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 6

Associations between CASA Treatment Composite Score at Baseline and Subsequent Changes in Risk

Change Scores between Baseline and 3-Months	Direct r_s Correlation with CASA Treatment Composite	Partial r_s Correlation Controlling for Baseline Scores on Respective Scale
SAVRY		
Historical	-.01	-.02
Social/Contextual	.09	.06
Individual/Clinical	-.01	-.01
Protective Factors	.11	.10
Risk Total Score	.02	.02
Summary Risk Rating	.10	.05
YLS/CMI		
Prior/Current Offenses	-.03	-.04
Family/Parenting	.03	.02
Education/Employment	.13	.10
Peer Associations	-.15	-.16
Substance Abuse	.00	-.09
Leisure/Recreation	.03	-.06
Personality/Behavior	-.04	-.09
Attitudes/Orientation	.11	.08
Risk Total Score	-.01	-.05
Summary Risk Rating	.04	-.01

Note. Positive change scores indicate greater improvement. As such positive associations, indicate that more treatment is associated with higher levels of improvement. None of the associations reached or approached significance.

Table 7

Associations between Change Scores (i.e., Improvements) and Reoffending in the Subsequent 6 Months

Change Scores between Baseline and 3-Months	AUC (SE)		Direct r_s Correlation with Number of Charges		Partial r_s Correlation Controlling for Baseline Scores on Respective Scale	
	Any	Violent	Any	Violent	Any	Violent
SAVRY						
Historical	.56 (.05)	.58 (.06)	-.11	-.10	-.13	-.22
Social/Contextual	.64* (.06)	.57 (.08)	-.19*	-.07	-.28**	-.16
Individual/Clinical	.53 (.07)	.47 (.10)	-.05	.04	-.13	-.02
Protective Factors	.52 (.05)	.51 (.07)	-.02	-.02	-.07	-.05
Risk Total Score	.57 (.07)	.55 (.09)	-.11	-.05	-.17*	-.09
Summary Risk Rating	.50 (.06)	.56 (.07)	-.02	-.07	-.12	-.14
YLS/CMI						
Prior/Current Offenses	.57 (.06)	.45 (.07)	-.11	.07	-.13	.06
Family/Parenting	.50 (.06)	.50 (.08)	-.03	-.01	-.10	-.06
Education/Employment	.53 (.06)	.49 (.09)	-.01	.02	-.09	-.04
Peer Associations	.63*** (.04)	.58* (.04)	-.21*	-.11	-.30***	-.18*
Substance Abuse	.60 (.05)	.67** (.05)	-.15	-.20*	-.24**	-.23**
Leisure/Recreation	.52 (.05)	.53 (.06)	-.04	-.03	-.16	-.14
Personality/Behavior	.61* (.05)	.60 (.06)	-.19*	-.10	-.28**	-.16
Attitudes/Orientation	.57 (.06)	.58 (.07)	-.11	-.09	-.19*	-.14
Risk Total Score	.62 (.06)	.60 (.07)	-.18*	-.11	-.26**	-.16
Summary Risk Rating	.54 (.06)	.61 (.07)	-.08	-.15	-.21*	-.23**

Note. Positive change scores indicate greater improvement. Thus, if youth who showed high improvement were less likely to reoffend, a significant inverse correlation between change and reoffending would be expected. AUC = area under the curve. * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 8

Association between Change Scores and Time to Reoffense: Penalized Cox Proportional Hazards Models

	Time to Any Reoffending					Time to Violent Reoffending				
Measure	*B*	*SE*	χ^2	HR	95% CI_{HR}	*B*	*SE*	χ^2	HR	95% CI_{HR}
SAVRY (*n* = 123)										
Baseline Risk Total	0.10	0.03	16.07***	1.10	[1.05,1.16]	0.14	0.04	13.80***	1.15	[1.06,1.26]
Change Historical	-0.00	0.14	0.00	1.00	[0.76,1.30]	-0.12	0.22	0.22	0.89	[0.58,1.40]
Change Social/Contextual	-0.36	0.13	7.10**	0.70	[0.54,0.91]	-0.37	0.20	3.28	0.69	[0.47,1.03]
Change Individual/Clinical	-0.09	0.11	0.71	0.91	[0.73,1.12]	0.04	0.13	0.07	1.04	[0.78,1.32]
Change Protective Factors	0.26	0.23	1.18	1.30	[0.81,2.01]	0.28	0.21	1.45	1.32	[0.80,1.88]
	χ^2 (5) = 18.39, *p* = .002					χ^2 (5) = 14.09, *p* = .015				
YLS/CMI (*n* = 132)										
Baseline Risk Total	0.12	0.03	17.89***	1.13	[1.06,1.20]	0.14	0.05	11.80***	1.15	[1.06,1.26]
Change Prior/Current Offenses	-0.07	0.25	0.08	0.93	[0.59,1.52]	0.60	0.41	2.66	1.81	[0.90,4.35]
Change Family	-0.05	0.19	0.08	0.95	[0.65,1.36]	-0.19	0.28	0.47	0.83	[0.48,1.41]
Change Education/Employment	-0.15	0.16	1.01	0.86	[0.62,1.15]	-0.02	0.18	0.01	0.98	[0.67,1.34]
Change Peer Associations	-0.44	0.24	3.16	0.65	[0.41,1.05]	-0.07	0.34	0.04	0.94	[0.49,1.84]
Change Substance Abuse	-0.08	0.18	0.22	0.92	[0.64,1.28]	-0.29	0.24	1.55	0.75	[0.47,1.18]
Change Leisure/Recreation	-0.11	0.27	0.16	0.90	[0.51,1.50]	-0.15	0.42	0.13	0.86	[0.36,1.90]
Change Personality/Behavior	-0.31	0.19	2.84	0.73	[0.51,1.05]	-0.41	0.30	1.92	0.66	[0.36,1.18]
Change Attitudes/Orientation	-0.10	0.24	0.18	0.90	[0.56,1.43]	-0.34	0.36	0.88	0.71	[0.34,1.41]
	χ^2 (9) = 24.70, *p* = .003					χ^2 (9) = 18.69, *p* = .028				

Note. *B* = regression coefficient; *SE* = standard error of *B*; HR = hazard ratio; 95% CI = 95% confidence interval of *Exp*(*B*). * *p* < .05, ** *p* < .01, *** *p* < .001 (two-tailed test).

Table 9

Potential Strategies to Enhance Risk Assessment Tools' Sensitivity to Change

1. Include items that are sufficiently dynamic, such as selecting items that have been found to be sensitive to change in the relevant populations (e.g., offenders receiving the usual services).
2. If the goal is to examine short-term changes in risk, ensure that the time frame for item ratings is sufficiently narrow (e.g., rate items based on functioning in the past 3 or 6 months rather than functioning in the past year).
3. Provide raters with guidelines on how to evaluate if an item has changed (e.g., example interview questions, guidelines for rating change).
4. Use response scales that capture variability in items which are not restricted by floor and ceiling effects (e.g., 3- or 5-point scales vs. dichotomous scales).
5. Evaluate and compare different approaches for measuring change, such as systems for rating level of change (e.g., definite/possible improvement, no change, or definite/possible deterioration) rather than solely systems for rating level of risk (e.g., low, moderate, or high risk).

Note: These strategies are based on recommendations in the field of treatment outcome assessment (i.e., Fok & Henry, 2015; Lambert & Vermeersch, 2013).