Does Reassessment of Risk Improve Predictions? A Framework and Examination of the SAVRY and YLS/CMI

Jodi L. Viljoen Andrew L. Gray Catherine Shaffer Aisha Bhanwer
Donna Tafreshi Kevin S. Douglas
Simon Fraser University

Psychological Assessment

DOI: https://doi.org/10.1037/pas0000402

Author Note

Jodi L. Viljoen, Andrew L. Gray, Catherine Shaffer, Aisha Bhanwer, Donna Tafreshi, 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.

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 experts recommend regularly reassessing adolescents' risk for violence, it is unclear whether reassessment improves predictions. Thus, in this prospective study, we tested three hypotheses as to why reassessment might improve predictions, namely the shelf-life, dynamic change, and familiarity hypotheses. Research assistants (RAs) rated youth on the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI) every three months over a one-year period, conducting 624 risk assessments with 156 youth on probation. We then examined charges for violence and any offense over a two-year follow-up period, and youths' self-reports of reoffending. Contrary to the shelf-life hypothesis, predictions did not decline or expire over time. Instead, time-dependent area under the curve scores remained consistent across the follow-up period. Contrary to the dynamic change hypothesis, changes in youth's risk total scores, compared to what is average for that youth, did not predict changes in reoffending. Finally, contrary to the familiarity hypothesis, reassessments were no more predictive than initial assessments, despite RAs' increased familiarity with youth. Before drawing conclusions, researchers should evaluate the extent to which youth receiving the usual probation services show meaningful short-term changes in risk and if so, whether risk assessment tools are sensitive to these changes.

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

Public Significance Statement

In this study, a couple of adolescent risk assessment tools predicted reoffending among adolescent offenders. However, despite recommendations to reassess risk regularly, short-term reassessments did not improve predictions, suggesting a need for further research.

Does Reassessment of Risk Improve Predictions?
A Framework and Examination of the SAVRY and YLS/CMI

Mental health and justice professionals often use risk assessment tools to assess the likelihood that an adolescent or adult offender will reoffend (Singh et al., 2014). Many of these tools conceptualize risk as dynamic, and as such, they include factors that can change (Douglas & Skeem, 2005). This is perhaps particularly true of *adolescent* risk assessment tools (Viljoen, Cruise, Nicholls, Desmarais, & Webster, 2012). In many ways, this focus on change aligns with what we know about risk, in general, and adolescence in particular. Researchers have confirmed that risk factors and offending trajectories can change as a result of effective treatments (Henggeler, 2016), life events (Simons & Barr, 2014), and developmental processes (Sweeten, Piquero, & Steinberg, 2013). In addition, compared to adults, adolescents appear to show more rapid changes in areas such as mood (Maciejewski, van Lier, Branje, Meeus, & Koot, 2015), friendships (Poulin & Chan, 2010), and personality (Roberts, Walton, & Viechtbauer, 2006).

As a result, many researchers and professionals consider it important to reassess adolescents' risk regularly (Vincent, Guy, & Grisso, 2012; Viljoen et al., 2012). However, surprisingly, little research has tested whether reassessments improve predictions. Thus, the purpose of the present study was to examine the predictive validity of reassessments using 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). The SAVRY is a structured professional judgment tool designed to assess violence risk in youth (Borum et al., 2006), whereas the YLS/CMI is designed to assess general reoffense risk in youth (Hoge & Andrews, 2002, 2012). Both tools are widely used (Viljoen, McLachlan, & Vincent, 2010), and predict violent and any reoffending with moderate effect sizes (Olver, Stockdale, & Wormith, 2009; Olver, Stockdale, & Wormith, 2014; Singh, Grann, & Fazel, 2011).

A Framework for Investigation: Reasons that Reassessments May Improve Predictions

To provide a framework for investigation, we developed three hypotheses regarding why reassessments might improve predictions, as described below. These hypotheses are not mutually exclusive, nor are they a comprehensive list of all reasons why it may be important to reassess risk.

Shelf-Life Hypothesis: Predictive Validity Will Diminish Over Time

According to the shelf-life hypothesis, risk ratings may *expire* or become outdated over time; thus, risk assessment tools may be better able to predict short-term than long-term offending. To test this hypothesis, studies have typically compared predictive validity at two time points, one that is proximal to the assessment, and the other that is more distal (see Appendix 1, Supplemental Material). If the prediction that is most proximal to the assessment shows stronger predictive validity than the more distal assessment, this would suggest that risk assessments have a limited shelf-life, thus reinforcing the importance of reassessment.

Few studies have examined the shelf-life of the SAVRY or YLS/CMI (see Appendix 2, Supplemental Material for a review, including further details on follow-up lengths). Meyers and Schmidt (2008) did not find any significant differences in the predictive validity of the SAVRY for a 1-year versus a 3-year follow-up period. In addition, although Takahashi et al. (2013) and

Olver et al. (2012) found that the YLS/CMI was slightly better at predicting short-term than long-term offending (i.e., 6 vs. 18-month follow-up, or juvenile vs adult offending), these differences did not reach statistical significance, as confidence intervals overlapped.

Several studies have examined other adolescent risk assessment tools (e.g., sex-offense specific tools). In two studies and one meta-analysis, predictive validity was higher for shorter timeframes (i.e., juvenile vs. adult reoffending – Ralston & Epperson, 2013; Schwalbe, 2008; average follow-up of 1.4 vs. 3.3 years – Worling, Bookalam, & Litteljohn, 2012). However, two other meta-analyses found that the predictive validity did not vary based on follow-up length (Schwalbe, 2008; Viljoen, Mordell, & Beneteau, 2012). In adult studies, the findings are also mixed. In two meta-analyses, follow-up length did not moderate the predictive validity of risk assessment tools (Blair et al., 2008; Fazel et al., 2012; Singh et al., 2011), whereas in one meta-analysis tools had higher predictive validity for longer follow-up periods (Yang, Wong, & Coid, 2010), and in another, tools had lower predictive validity for longer follow-ups (Smith, Cullen, & Latessa, 2009).

These findings are difficult to interpret. First, most studies use variable rather than fixed follow-up lengths. Thus, the predictive validity for specific time periods is unclear (e.g., 6 months vs. 2 years). Second, some studies include juvenile and adult records as a proxy for follow-up length. However, this results in wide-ranging follow-ups. For instance, if a youth was 12-years old at the time of the assessment, adult records would capture offenses that occurred a minimum of 6-years later. In contrast, if a youth was 17 and a half years old, adult records would capture offenses that occurred as little as 6 months later. Third, studies have typically examined reoffending at only two time periods. Finally, rationales for selecting a particular follow-up length have been lacking, resulting in ambiguity and inconsistency across studies.

In the present study, we aimed to address these gaps by measuring reoffending outcomes for four fixed time periods: very short-term reoffending (i.e., 0 – 3 months), short-term reoffending (i.e., 3 – 6 months), mid-range reoffending (i.e., 6 – 12 months), and longer-term reoffending (i.e., 12 – 24 months). We chose these time frames because test developers often recommend reassessing risk at 3, 6, or 12 months (e.g., Prentky & Righthand, 2003; Worling & Curwen 2001; Viljoen, Nicholls, Cruise, Desmarais, & Webster, 2014). Also, to track changes in predictive validity over time, we calculated time-dependent area under the curve (AUC) scores. This study represents one of the initial applications of this approach to risk assessment (see Gray, Viljoen, & Douglas, 2015; Lloyd & Serin, 2015; Walters, Deming, & Casbon, 2015).

Dynamic Change Hypothesis: Changes in Risk will Predict Reoffending

According to the dynamic change hypothesis, tracking within-individual changes in risk may help evaluators to predict reoffending more accurately, as individuals who show reductions in risk scores might be less likely to reoffend than other individuals (Olver et al., 2007). To test this hypothesis, researchers usually administer a risk assessment tool at two or more time points. If changes in scores predict reoffending, this would suggest that assessing change aids in prediction (see Appendix 1, Supplemental Material). In essence, studies on the dynamic change hypothesis focus on *multiple assessments* and a single follow-up length, whereas studies on the shelf-life hypothesis typically focus on a single assessment and *multiple follow-ups*.

Although research on the dynamic change hypothesis has rapidly grown (see Appendix 3, Supplemental Material for a review), to our knowledge, only two studies have examined this hypothesis with youth. In one study, research assistants (RAs) rated youth's SAVRY and Juvenile Sex Offender Assessment Protocol-II (J-SOAP-II; Prentky & Righthand, 2003) scores at admission to a sex offending treatment program and again at discharge (Viljoen et al., 2015). Despite significant reductions in risk, change scores did not significantly predict reoffending; youth who improved were as likely as other youth to reoffend over an 8-year follow-up period. The authors suggested that this was because change itself is dynamic. That is, even if a youth improves or deteriorates, the degree to which they have changed may not be permanent. As such, change scores may not predict long-term offending that spans over an 8-year follow-up period. Similarly, a dissertation on the Violence Risk Scale: Youth Version (VRS:YV; Lewis, Wong, & Gordon, 2004) did not find significant associations between change scores and reoffending in a subsample of 22 youth (Stockdale, 2008).

In contrast, a number of studies with adult samples have found changes in risk total scores to predict reoffending (Beggs & Grace, 2011; Blanchard, 2013; Cohen, Lowenkamp, & VanBenschoten, 2016; de Vries Robbé et al., 2015; Hogan & Olver, 2016; Labrecque, Smith, Lovins, & Latessa, 2014; Lewis, Olver, & Wong, 2013; Michel et al., 2013; Olver, Christofferson, Grace, & Wong, 2014; Olver, Nicholaichuk, Kingston, & Wong, 2014; Olver et al., 2007; Raynor, 2007; Vose, Smith, & Cullen, 2013; Wilson, Desmarais, Nicholls, Hart, & Brink, 2013), particularly after other variables are controlled for in analyses (e.g., baseline scores; see Appendix 3, Supplemental Material). That said, a couple of studies have failed to find significant associations between change scores and reoffending (Hanson, Harris, Scott, & Helmus, 2007; Hanson, 2015; Goodman-Delahunty & O'Brien, 2014). As many of these studies have focused on sex offenders in specialized treatment programs or adult forensic psychiatric patients, their generalizability to adolescent samples is unclear. Furthermore, in most studies, changes in risk scores were rated from files, rather than a combination of interview and file information.

To extend this research, the current study tested the dynamic change hypothesis in a sample of adolescent offenders on probation. Probation is the most common disposition given to adolescent offenders in the United States (Hockenberry & Puzzanchera, 2015) and Canada (Alam, 2015). Rather than assessing risk at two time points and offending at one follow-up period, we assessed both youth risk and offending up to five times using a combination of interview and file review. Assessing risk at more than two time points allowed us to differentiate between within- and between-person differences in risk by allowing for an average level of risk per person. That is, we wanted to ask the question of whether change in an individual's level of risk, compared to what is average for that person, is predictive of changes in reoffending (see Appendix 1, Supplemental Material). For example, is it the case that having a higher level of risk than what is usual for a youth is associated with higher levels of reoffending, or is it that youth who have higher levels of risk, on average, simply reoffend more? The former would suggest that risk is dynamic in the sense that fluctuations above or below a person's average level of risk are predictive of offending. Difference scores do not allow for this kind of comparison as they only capture whether there is an increase or decrease in risk between two time points and do not provide any information as to whether a person's level of risk at a particular time point is meaningfully different than what would be considered average for that person. In addition, whereas most studies measure offending in a static manner (i.e., with a single follow-up period), assessing reoffending at multiple time points allowed us to examine changes in offending over time and whether differences in risk levels account for these changes.

Familiarity Hypothesis: Reassessments will be More Predictive than Initial Assessments

Whereas the dynamic change and shelf-life hypotheses focus on changes within *evaluatees*, our final hypothesis, the familiarity hypothesis, focuses on changes in *evaluators* and their knowledge of a youth. In particular, according to the familiarity hypothesis, evaluators may gain additional information in each assessment, which in turn could improve predictions. For instance, Hilterman et al. (2016) found that evaluators rated youth higher on historical factors such as maltreatment at the second assessment than at the initial assessment, suggesting a gain in knowledge about youths' history. During reassessments, evaluators might also obtain feedback on the accuracy of their initial assessment, enabling them to adjust their future risk ratings.

To test the familiarity hypothesis, researchers could compare the predictive validity of initial risk assessments to reassessments. If the predictive validity for *re*assessments is higher than the predictive validity of *initial* assessments or improves over time, this would suggest that familiarity enhances predictions (see Appendix 1, Supplemental Material). However, to adequately test this hypothesis and disentangle the impact of familiarity from other factors, the follow-up length between each assessment and follow-up must be equal (e.g., three months after the initial risk assessment and three months after the reassessment). Otherwise, it is not possible to determine if results are due to familiarity or differences in follow-up length (i.e., shelf-life).

Few studies meet this criterion. Although a number of studies have compared the predictive validity of the initial assessment to that of a subsequent reassessment (e.g., Olver, Christofferson, et al., 2014; O'Keefe, Klebe, & Hromas, 1998; Viljoen et al., 2015; see Appendix 4, Supplemental Material), in those studies, reoffense records were collected after the second assessment. As such, the second assessment was more proximal to reoffending than the initial assessment, making it difficult to tease apart familiarity from shelf-life. Nevertheless, these studies generally indicate that predictive validity for initial and subsequent assessments are quite similar (Hanson et al., 2007; Hanson, 2015; Olver, Christofferson, et al., 2014; Olver, Nicholaichuk, et al., 2014; Olver et al., 2007; Viljoen et al., 2015). One study, by Wilson and colleagues (2013), used an equal follow-up length of three months for each of their four reassessment points, thus providing a more direct test of the familiarity hypothesis in a sample of 30 adult male forensic psychiatric patients. Again, however, although dynamic factors predicted reoffending, reassessments did not appear to significantly improve predictions (i.e., confidence intervals overlapped).

It may be that reassessments are only better if conducted by the same evaluator who had assessed them previously. If a different evaluator assessed a youth each time, that evaluator may not gain the familiarity needed to improve their assessment. As such, in this study, we tested whether support for the familiarity hypothesis was stronger when we used a narrow definition of familiarity (i.e., reassessment by the same RA) versus a broad definition (i.e., reassessment by any RA). To control for shelf-life, we used equal follow-up periods for each assessment, using a design that was similar to Wilson et al. (2013).

The Present Study

In sum, despite recommendations to reassess youth's risk regularly, it is not clear whether reassessments improve risk predictions. The present study aimed to extend prior research in several ways. First, to provide a framework for investigation, we tested three specific hypotheses, namely the shelf-life hypothesis, the dynamic change hypothesis, and the familiarity hypothesis. Second, rather than assessing youth at two time periods, we reassessed youth up to five times. Third, rather than using a variable follow-up period, we measured reoffending outcomes at fixed time periods that were selected to correspond with the recommendations in the literature about when to reassess risk. In particular, we examined very short-term reoffending (i.e., 0 – 3 months), short-term reoffending (i.e., 3 – 6 months), mid-range reoffending (i.e., 6 – 12 months), and longer-term reoffending (i.e., 12 – 24 months). Finally, to enhance our ability to test our hypotheses, we applied statistical procedures that are not yet widely used in risk assessment research, including time-dependent AUCs and multilevel modelling (MLM). Consistent with the dynamic change hypothesis, we expected that youth who showed decreases in risk, compared to what is average for that youth, would show decreases in reoffending.

Method

To ensure transparent reporting of methodology and results, this manuscript adheres to the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement (Singh, Yang, Mulvey, & the RAGEE Group, 2014), a 50-item reporting checklist.

Participants

Our sample included 156 youth on probation in the Greater Vancouver area. Participants ranged in age from 12 to 18 years old, with a mean age of 16.41 years (*SD* = 1.14). Approximately two-thirds of participants were male (68.6%, *n* = 107). Many youth were from ethnic minority groups (61.5%, *n* = 96); 38.5% were Caucasian/European (*n* = 60), 29.5% (*n* = 46) were Aboriginal (i.e., First Nations, Métis, Inuit), 12.8% (*n* = 20) were Asian, 7.1% (*n* = 11) were East Indian/Southeast Asian, 7.1% (*n* = 11) were Hispanic, and 4.5 % (*n* = 7) were African. The ethnic and gender distribution of our sample is similar to those reported in national and provincial statistics of youth offenders, suggesting that our sample is fairly representative in this regard (Calverley, Cotter, & Halla, 2010). With respect to index offenses, over half of youth had committed a violent offense (59.6%, *n* = 93) and one-third had committed a property offense (36.5%, *n* = 57). Most youth had no charges prior to the index offense (67.9%, *n* = 106) and had received therapy or treatment programs in the previous three months (70.5%, *n* = 110).

Procedure

This study used a prospective design in which risk assessments were conducted by RAs using interview and justice file information. Ethics approval was obtained from Simon Fraser University and our research sites. All methods complied with ethical procedures.

Sampling and Consent. Youth at 11 probation offices were informed about the study by youth probation officers and study liaisons, and through 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 Greater Vancouver area. Also, 24.8% (*n* = 126) of youth were not interested in participating, and 5.1% (*n* = 26) could not be reached. Active informed consent was obtained from legal guardians. In 30 cases (5.9%), guardians could not be reached, and as such, these youth were unable to participate. Finally, seven youth participated but declined access to reoffense records; these youth were excluded from the present analyses.[1]

RAs. RAs included 11 graduate students, and 8 students with an undergraduate psychology degree who had previously completed course work and practicums with clinical or offender populations. Training included two days of didactic training on risk assessment tools, and four or more practice cases. Prior to working independently, RAs were required to demonstrate adequate inter-rater reliability, defined as scores within five points of the gold standard rating for SAVRY and YLS/CMI total scores. To examine inter-rater reliability of the SAVRY and YLS/CMI, a random sample of 20.5% of cases (n = 32) were independently coded by a second rater, who observed the interview and reviewed the same file information.

Baseline assessment. RAs conducted a standardized interview with youth at a probation office or a quiet public place, such as a coffee shop. Youth were provided a stipend of $20. Afterwards, RAs reviewed the youth's justice records, which typically contained a running log of probation officers' contacts with the youth, records of program attendance, psychiatric reports, police reports, and interviews with the family and youth. RAs then rated the SAVRY and YLS/CMI based on interview and files using the rating guidelines in the tools' manuals. As this was a true prospective study, RAs were blind to youths' subsequent charges. YLS/CMI and SAVRY items were pro-rated if 10% or fewer items were missing (Hoge & Andrews, 2002).

Follow-up assessments. Youth were reassessed every three months over a one-year period. To minimize missing follow-ups, we followed research guidelines (e.g., Ribisl et al., 1996). In particular, RAs: (a) maintained contact with participants between follow-ups assessments (e.g., called youth to check if their contact information had changed); (b) made persistent efforts to contact youth; (c) used collateral sources to assist in locating a youth (e.g., parents, service providers); (d) provided flexibility in the times, location, and method of interview (e.g., met with youth in a location near to their home); (e) problem-solved with youth about strategies to maintain contact; and (f) provided youth an additional $25 incentive for completing all the follow-ups. If efforts to complete a follow-up interview were not successful, SAVRY and YLS/CMI ratings were made based on the youths' file information. In total, 87.0% (n = 407) of the reassessments were coded from a combination of interviews and file-information, whereas 13.0% (n = 61) were coded based on file information alone.

To examine changes in risk, our goal was to complete at least one follow-up per youth. Of the 156 youth in the study, 145 youth had at least one follow-up assessment (92.9%). Specifically, 14 (9.0%) youth had one reassessment, 22 (14.1%) had two reassessments, 26 (16.7%) had three reassessments, and 83 (53.2%) had four reassessments. Youth with and without reassessments did not differ significantly in age, gender, ethnicity, index offense, prior offenses, or SAVRY and YLS/CMI Risk Total Scores. In general, youth were reassessed by the same RA, but one-quarter of the assessments were transferred to other RAs (26.9%, n = 126 of the 468 reassessments), as some RAs were no longer available.

Official offending records. Adult and youth justice records were collected through the Corrections Network System (CORNET), a province-wide justice database. RAs received a half-day training and coded three to five practice cases. Violent reoffenses included charges for "actual, attempted, or threatened infliction of bodily harm of another person" (Douglas et al., 2013, pp. 36–37). Any reoffenses included any charges, including breaches and probation violations. Similar to most adolescent risk assessment studies (Schwalbe, 2008; Viljoen et al., 2012), we examined charges rather than convictions, as convictions underestimate reoffending (Farrington, Jolliffe, Loeber, & Homish, 2007). Time-at-risk was calculated as the number of days between the baseline assessment and the date of offense. We coded the date of the offense rather than the date of the charge, given the lag between an offense and charge. For youth who did not reoffend, time-at-risk was the full two-year follow-up. During the two-year follow-up, 19.9% (n = 31) and 44.2% (n = 69) were charged with violent and any reoffenses, respectively.

Measures

The Structured Assessment of Violence Risk in Youth (SAVRY). The SAVRY (Borum et al., 2006) includes 24 risk factors in three domains: Historical (e.g., history of violence), Social/ Contextual (e.g., peer delinquency), and Individual/Clinical (e.g., risk-taking and impulsivity). Each risk factor is rated Low, Moderate, or High. The SAVRY also includes six Protective Factors (e.g., strong attachments and bonds), rated as Present or Absent. For research purposes, scores are typically summed to create a Risk Total Score (see Olver et al., 2009). In the present study, the mean Risk Total Score was 25.92 (SD = 8.47, range = 1 to 43). We also created a Dynamic Risk Total Score by summing Social/Contextual and Individual/Clinical domains. Interrater reliability was calculated using a two-way random effect model for single raters, absolute agreement (McGraw & Wong, 1996). Intraclass correlation coefficient (ICC) fell in the excellent range (ICC = .91 for SAVRY Risk Total Score; Shrout & Fliess, 1979).

Youth Level of Service/Case Management Inventory (YLS/CMI). The YLS/CMI (Hoge & Andrews, 2002) includes 42 dichotomous items, which are divided into eight subscales (i.e., Prior and Current Offenses, Family Circumstances/Parenting, Education/Employment, Peer Relations, Substance Abuse, Leisure/Recreation, Personality/Behavior, and Attitudes/ Orientation). These items are summed to create a Risk Total Score (M = 19.43, SD = 7.47, range = 0 – 35 in the present study). We also created a Dynamic Risk Total Score by summing all of the subscales except for Prior and Current Offenses. Inter-rater reliability in the present study was excellent (ICC = .82 for YLS/CMI Risk Total Score; Shrout & Fliess, 1979). We used the YLS/CMI rather than the YLS/CMI 2.0 (Hoge & Andrews, 2011), as this was the version that was available when the study began. For the most part, the coding instructions for the YLS/CMI and YLS/CMI 2.0 are identical (Hoge & Andrews, 2011). To determine whether our findings might generalize to the YLS/CMI 2.0, RAs rated 21 youth on both the YLS/CMI and YLS/CMI 2.0; the two versions had a correlation of .99 (Gray, Viljoen, & Douglas, 2015).

Self-Report of Offending (SRO). The SRO (Huizinga, Esbensen, & Weiher, 1991) is a self-report scale that assesses 23 types of offenses (e.g., purposely destroyed or damaged property; stolen something from a store; sold marijuana).[2] Previous research has found that the SRO has good psychometric properties, including measurement equivalence across ethnic groups (Knight et al., 2004). For each SRO item, youth were asked to report whether they had committed the offense never, once, a couple of times (i.e., “2 to 3 times”), or multiple times (i.e., “4 or more times”) during their lifetime, and during the past three months. To help orient youth to the past three months, RAs first interviewed youth about major events that had occurred during this time period (e.g., changes in schools, relationships, and residences). We calculated a SRO Any Offenses Total Score (for both the past three months) by summing all 23 items, and

SRO Violent Offense Total Score by summing the nine items that pertained to violence (e.g., beaten up or physically attacked somebody so badly that they probably needed a doctor).

Data Analysis Plan

General predictive validity. Prior to testing hypotheses, we examined the predictive validity of the SAVRY and YLS/CMI for the full two-year follow-up period. We calculated the area under the curve (AUC) of the receiver operating characteristic (ROC; Hanley & McNiel, 1982) using the R package "pROC" (Robin et al., 2015). The AUC reflects the probability that a randomly selected reoffender will score higher on a tool than a randomly selected non-reoffender (Hanley & McNeil, 1982). AUC values for female and male youth were compared using Venkatraman's (2000) test for uncorrelated ROC curves. To estimate the magnitude of differences between offenders and non-offenders on the tools, Cliff's delta ($\hat{\delta}$; Cliff, 1993) was calculated using the "effsize" package in R (Torchiano, 2015). Delta values can range from -1 (i.e., all observations in the non-reoffender group are larger than the reoffender group) to +1 (i.e., all observations in the reoffender group are larger than the non-reoffender group). To examine speed to reoffending, Cox Proportional Hazards survival analyses were conducted using the R package "survival" (Fox & Weisberg, 2011). Finally, rather than treating offending as a dichotomous outcome (i.e., present or absent), we also examined *number* of charges and time-at-risk. As this data was not normally distributed, we used Spearman's rank-order correlation (r_s).

Shelf-life hypothesis: Time-dependent ROC analysis. To examine the shelf-life hypothesis (i.e., whether predictive validity diminishes over time), time-dependent ROC analyses were conducted using the R package "survivalROC" (Heagerty, 2013). Time-dependent ROC analyses combine elements of ROC and survival analysis. The time-dependent AUC (AUC_t) value is defined as the probability that a risk score for a random case (i.e., a youth who has committed an offense) exceeds that for a random control (i.e., a youth who has not committed an offense) at a fixed point in time (t). Statisticians have developed several types of AUC_t; in this study, we used *cumulative*/*dynamic* AUC_t ($AUC_t^{C/D}$; Heagerty & Zhen, 2005), with nearest neighbor estimator (Heagerty, Lumley, & Pepe, 2000). $AUC_t^{C/D}$ are cumulative in the sense that they capture reoffending that occurs at any time point from the assessment to a fixed time point (e.g., 0 – 3 months, 0 – 6 months, etc.; Blanche, Dartigues, & Jacqmin-Gadda, 2013). The specificity of the $AUC_t^{C/D}$ is calculated so that if T_i (i.e., the survival time for subject i) is greater than t (i.e., a fixed point in time) for a case, then the case serves as a control. In other words, if an identified youth has not committed a reoffense by time t, they would be counted as a non-reoffender at that time point. However, once $t \geq T_i$, the youth is classified as a case (i.e., the youth is recognized as committing an offense). Sensitivity of the $AUC_t^{C/D}$ is calculated so that cases in which $t \geq T_i$ are included in the computation at each fixed time point (i.e., those who commit an offense at or prior to the fixed time point are considered cases at that time point).

Dynamic change hypothesis: MLM. To examine the dynamic change hypothesis (i.e., whether change in risk, relative to what is average for a person, predicts changes in reoffending), MLM was conducted using the GLIMMIX procedure in SAS, Version 12.1 (SAS Institute Inc., 2012). MLM for repeated measurements involves a nested structure that takes into account the dependency of measurements taken on the same individual by indicating both fixed effects (i.e., population averages) and random components (i.e., variance of fixed effects at person level) in the model (Hedeker & Gibbons, 2006). As the outcome variable (i.e., number of offenses) involves counts, we used a Poisson distribution and conducted generalized linear mixed-effects modeling. Given the nested nature of MLM, it does not require that participants be assessed at every time point, which made it well suited to our design, as there was missing data.

In our analyses, we employed two-level models with measurement occasions (i.e., Level 1 of the model, n = 624) nested within individuals (i.e., Level 2 of the model, n = 156) and specified an unstructured covariance matrix for the variance components. To reflect youths' change in risk from their own average score (i.e., within-person effects), we used SAVRY and YLS/CMI scores centered on person-level means. Following the procedures described by Raudenbush and Bryk (2002) and Singer and Willett (2003), we began with an unconditional means model in which only the intercept is entered with a variance component (i.e., allowed to vary across individuals). This was followed by an unconditional growth model in which time was added to the model with a variance component. Finally, we tested several conditional growth models in which both within- and between- effects of risk scores were added as predictors and the interactions between these effects were examined.

Familiarity hypothesis: MLM. To test the familiarity hypothesis (e.g., whether *reassessments* were more predictive than initial assessments), we conducted a set of MLM models in which familiarity with the youth was added as a moderator. We defined familiarity two ways. Within the narrower definition, familiarity was defined as whether the youth had been previously assessed by the *same* RA (1 = prior assessment with that RA, 0 = no prior assessments). Within the broader definition, familiarity was defined as whether the youth had been previously assessed by *any* RA (number of prior assessments, ranging from 0 to 4, or in other words, time). Specifically, we created a product term representing the interaction between familiarity and person mean-centered SAVRY and YLS/CMI Risk Total Scores. In addition, we compared the predictive validity of initial and average Risk Total Scores using ROC analyses.

Results

General Predictive Validity

The SAVRY and YLS/CMI Risk Total Scores and summary risk ratings significantly predicted violent and any reoffending, as did most subscale scores (see Tables 1 and 2, respectively). AUCs for Risk Total Scores reached the large range (i.e., > .71; Rice & Harris, 2005), whereas AUCs for subscale scores generally fell in the moderate range (i.e., > .64). Cliff's delta generally fell in the medium range for both subscale and total scores (i.e., .33 – .47; Romano, Kromrey, Coraggio, Skowronek, & Devine., 2006). Based on Venkatraman's test, AUCs for violent reoffending did not differ significantly by gender (i.e., SAVRY Risk Total Scores: AUCs = .74 and .78, for boys and girls, respectively; YLS/CMI Risk Total Scores: AUCs = .72 and .81), nor did AUCs for violent reoffending (i.e., SAVRY Risk Total Score: AUCs = .74 and .61, respectively; YLS/CMI Risk Total Score: AUCs = .74 and .69). Moreover, mean scores on the SAVRY and YLS/CMI Risk Total did not differ significantly by gender.

Shelf-Life Hypothesis: Time-Dependent ROC Analyses

Contrary to the shelf-life hypothesis, the SAVRY and YLS/CMI Risk Total Scores retained moderate to large predictive accuracy over the full two-year follow-up (see Figure 1). In particular, $AUC_t^{C/D}$ for *any* reoffending were nearly identical at the 3, 6, 12, and 24-month follow-ups (i.e., SAVRY $AUC_t^{C/D}$ = .71, .71, .70, and .72, respectively; YLS/CMI $AUC_t^{C/D}$ = .73,

.73, .72, and .71, respectively). With respective to *violent* reoffending, $AUC_t^{C/D}$ remained consistent over time for the YLS/CMI ($AUC_t^{C/D}$ = .70, .69, .68, and .69, respectively). On the SAVRY, $AUC_t^{C/D}$ for *violent* reoffending showed a very slight decline between 3 and 6 months but the overall magnitude of the AUC values remained just below the threshold for large effects ($AUC_t^{C/D}$ = .74, .68, .67, and .69, respectively). When we reran the analyses using SAVRY and YLS/CMI Dynamic Risk Total Scores, we found a parallel set of findings (see Figure 1).

Although AUC_t analysis only accounts for time to *first* violent or any reoffense, some youth reoffended at multiple time points. As such, we also calculated separate AUC values for each of the discrete follow-up periods (i.e., 0 – 3, 3 – 6, 6 – 12, 12 – 24 months). For instance, if a youth offended at 2 months and again at 18 months, the youth was included as an offender in the AUCs for these two time periods. In these analyses, SAVRY and YLS/CMI scores at baseline also maintained their predictive accuracy at each time period (see Table 3).

Dynamic Change Hypothesis: MLM

Model 1: Unconditional Means Model. We first tested intercept-only models for our offending outcomes (i.e., a MLM model with no predictors) to determine whether there was between-person variability on the outcome. The fit indices for the Poisson distribution for both the number of violent (χ^2 /df = 0.14) and number of any charges (χ^2 /df = 0.21) outcomes were poor, indicating less variation in the data than what is expected with a Poisson distribution. In addition, despite troubleshooting, the maximum likelihood estimation procedure was unable to reach optimization when we ran further models with time and SAVRY or YLS/CMI Risk Total Scores added as predictors. Furthermore, number of charges was quite low and showed little variability (*M* = 0.06, *SD* = 0.44 for violent reoffending and *M* = 0.53, *SD* = 1.36 for any reoffending). For these reasons, the remaining MLM analyses focused solely on self-report data on the SRO rather than official records. Although self-reported offending outcomes were available for a fewer number of reassessments due to missing data (i.e., 405 reassessments vs. 624 reassessments for official records), the SRO detected higher rates of reoffending and greater variability (*M* = 1.99, *SD* = 2.93 for violent reoffending and *M* = 6.01, *SD* = 8.69 for any reoffending). The unconditional means models for self-reported violent and any reoffending had good model fit for self-reported violent reoffending (χ^2 /df = 0.95). Although the model fit indicated overdispersion for any reoffending (χ^2 /df = 2.31), when further predictors were added in subsequent models, the model fit index substantially improved (see results below). This indicated that our hypotheses could be tested using self-reported offending outcomes.

Model 2: Unconditional Growth Model. Next, growth (i.e., change over time) in offending was examined by adding time as a predictor and allowing the slope of time to vary across individuals. The fixed-effect estimate associating time with reoffending suggested that, on average, youth reported less violent, β = -0.28, SE = 0.09, $p < .01$, and any reoffending, β = -0.13, SE = 0.06, $p < .05$, over time. In addition, the variance component for time indicated that the rate of change in violent ($\sigma^2_{\upsilon} = 0.13$, SE = 0.05, $p < .01$) and any reoffending ($\sigma^2_{\upsilon} = 0.20$, SE = 0.05, $p < .05$) differed between individuals.

Model 3: Conditional Growth Model of Within-Person Effects of Risk. Following this, we added SAVRY Risk Total Scores centered on individual means into the above model to represent within-person changes in risk levels (i.e., person mean-centered scores). Variance components were estimated for the intercept, time, and within-person effects of risk. A parallel analysis was conducted for the YLS/CMI. Contrary to the dynamic change hypothesis, fixed-effect estimates associating SAVRY and YLS/CMI Risk mean-centered scores with reoffending suggested that neither within-person effects of the SAVRY nor the YLS/CMI were predictive of changes in self-reported violent and any reoffending (see Tables 4 and 5).[3]

Model 4: Conditional Growth Model, Comparing Within-Person and Between-Person Effects of Risk. Youths' mean SAVRY or YLS/CMI Risk Total Scores across the five measurement occasions were added to the models along with time and mean-centered SAVRY or YLS/CMI Risk Total Scores. Results indicated that most variance in self-reported reoffending was accounted for by between-individual differences in Risk Total Scores such that youth who, on average, had higher risk scores committed more reoffenses (see Tables 4 and 5). In other words, mean risk scores were more predictive than within-individual change.

Model 5: Conditional Growth Model of Interaction of Within-Person and Between-Person Effects of Risk. We tested another set of models that included the interaction between within- and between-person effects of Risk Total Scores. Neither the between-person effects of the SAVRY nor the YLS/CMI moderated the relationship between within-person effects of risk and violent or any reoffending (see Tables 4 and 5). In other words, increases in risk did not predict reoffending more strongly in youth with high versus low average Risk Total Scores.

Model 6: Person-Level Variance as a Moderator (Post-Hoc Analysis). In the prior models, within-person effects of risk did not predict reoffending. A possible explanation for this finding is that youths' deviations from their average Risk Total Scores may reflect measurement error rather than true change. If this was the case, average Risk Total Scores should have weaker associations with reoffending among youth who show large deviations in Risk Total Scores than among youth with small deviations in Risk Total Scores. To test this, we computed the average squared deviation from each youth's average Risk Total Score (i.e., person-level variance), and then added time, person-level average SAVRY or YLS/CMI Risk Total Scores, person-level variance, and the person-level variance by person-level average SAVRY or YLS/CMI Risk Total Scores interaction as predictors in post-hoc MLM models. The intercept and time coefficient were allowed to vary across individuals. Contrary to this measurement error hypothesis, the fixed-effect estimates of the interactions between person-level variance and person-level average Risk Total Scores were not significant (β = -.00 to -.01, $p > .05$).

Given that SAVRY and YLS/CMI Risk Total Scores include historical factors that may not easily change, all of the above analyses were reran using SAVRY and YLS/CMI Dynamic Risk Total Scores. However, the results remained the same (analyses available upon request).

Familiarity Hypothesis: Moderator Analyses in MLM

Model 1: Conditional Growth Model of Interaction of Within-Person Effects of Risk and Familiarity, Narrowly Defined. We added time, mean-centered SAVRY or YLS/CMI Risk Total Scores, familiarity, and the familiarity by mean-centered SAVRY or YLS/CMI Risk Total Scores interaction as independent variables in MLM models. These models used the narrow definition of familiarity (i.e., whether the RA had previously assessed that youth). The intercept and slopes of time and mean-centered SAVRY scores were allowed to vary across individuals. Contrary to the familiarity hypothesis, the fixed-effect estimates of the interactions between familiarity and Risk Total Scores were not significant (see Table 6).

Model 2: Conditional Growth Model of Interaction of Within-Person Effects of Risk and Familiarity, Broadly Defined. Next, we reran the above model using the broader definition of familiarity (i.e., number of prior assessments with any RA or in other words, time). Again, the interactions between familiarity and SAVRY and person mean-centered YLS/CMI Risk Total Scores were not significant (see Table 6).

ROC Analyses. As a further test of familiarity, we compared AUCs for initial and average risk total scores, using separate one-year follow-up periods.[4] On the SAVRY, AUCs for initial versus average Risk Total Scores did not differ significantly for either violent (AUC = .74 vs. .76, respectively) or any charges (AUC = .73 vs. .79), as evidenced by overlapping confidence intervals. On the YLS/CMI, initial versus average Risk Total Scores also did not differ significantly (AUC = .71 vs. .78 for violent charge; AUC = .73 vs. .76 for any charge).

All of the above analyses were reran using SAVRY and YLS/CMI *Dynamic* Risk Total Scores. However, the results remained the same (analyses are available upon request). As a final step, we tested whether we might obtain more support for the dynamic change and familiarity hypotheses when we excluded file-only reassessments, as they might capture less information about change. However, when we reran all of our MLM analyses without file-only cases, the same pattern of results was found (analyses are available upon request).

Discussion

Similar to other studies (e.g., Olver et al., 2009; Singh et al., 2011), we found that the SAVRY and YLS/CMI significantly predicted violent and any reoffending with at least moderate effect sizes. However, our primary focus was whether reassessing risk with these tools might improve predictions. In particular, we tested three hypotheses: the shelf-life, dynamic change, and familiarity hypotheses.

According to the shelf-life hypothesis, predictive validity of reassessments may diminish or expire over time. Contrary to this hypothesis, AUC_t scores did not decline over the two-year follow-up period. Instead, predictive validity for short-term predictions (e.g., 0 – 6 months) were comparable to those for longer-term predictions (e.g., 12 – 24 months). Furthermore, SAVRY and YLS/CMI Dynamic Risk Total Scores did not expire any more quickly than Risk Total Scores. In some ways, these results are not particularly surprising. Although some studies suggest that predictive validity diminishes over time (e.g., Ralston & Epperson, 2013; Worling et al., 2012), an equal number of studies have found that follow-up length does not moderate predictive validity (e.g., Fazel et al., 2012; Schwalbe, 2007; Viljoen et al., 2012). Also, in this study we focused on reoffending over a 2-year period, as researchers often suggest reassessing youth prior to then (Vincent et al., 2012; Viljoen et al., 2014; Worling & Curwen, 2001). It is possible that risk predictions only start to expire after a much longer period of time. That said, one study reported that the SAVRY and YLS/CMI predicted for over a long mean follow-up period of 10 years (Schmidt, Campbell, & Houlding, 2011).

According to the dynamic change hypothesis, tracking within-person *changes* in risk scores may aid in predictions. Specifically, if a youth's risk level increases, relative to what is typical for that youth, it may signal an increased likelihood of reoffending. Contrary to this hypothesis, associations between self-reported offending and within-individual changes in SAVRY and YLS/CMI Risk Total and Dynamic Risk Total Scores were nonsignificant.

Furthermore, when both between-person effects and within-person effects were included in the MLM models, most of the variance in offending was accounted for by between-individual effects. In other words, knowing what a youth's risk was like, on average (i.e., between-individual effects), was more predictive than knowing the extent to which a youth's risk had changed (i.e., within-person-effects).

The failure to find a significant association between within-individual changes in risk and reoffending is consistent with a prior research with adolescents (Viljoen et al., 2015). However, a number of studies with adults have reported significant associations between changes in scores and reoffending (e.g., de Vries Robbe et al., 2014; Hogan & Olver, 2016; Michel et al., 2015; Olver et al., 2007; Vose et al., 2013; Wilson et al., 2014). Several factors may contribute to this variability in findings. First, whereas some past research has focused on whether the difference in scores at two time points predicts reoffending at a single time point, the current study examined whether changes in risk scores, relative to the person's *average* risk score, predicted *changes in reoffending*. Given that we used MLM to synthesize multiple time points for both risk scores and reoffending outcomes, it may be a more rigorous test (though see Michel et al., 2013 which reported predictive validity of change across multiple time points).

Second, the results might differ based on the nature of factors included in a particular risk assessment tool. For instance, although many of the factors on the SAVRY and YLS/CMI are potentially dynamic, they might nevertheless be relatively stable rather than acute (i.e., changing slowly and gradually rather than rapidly and in response to immediate circumstances; see Hanson et al., 2007). Another possibility is that some constructs (e.g., impulsivity) might consist of a more stable aspect (that is, a person remains relatively impulsive in general over time) and a less stable aspect (that is, that same person shows episodic acute instability as a result of "stable impulsivity"). This is consistent with a state-trait model of psychopathology (Jackson & Sher, 2003). Such periodic episodes likely cannot be detected by the design used in the current research. As such, future research should examine various approaches for measuring change, using a variety of study designs. For instance, the Violence Risk Scale (Wong & Gordon, 1999-2003) and the Short-Term Assessment of Risk and Treatability (Webster, Martin, Brink, Nicholls, & Desmarais, 2010) both have adolescent versions (Viljoen et al., 2014; Wong, Lewis, Stockdale, & Gordon, 2011) and aim to measure short-term and treatment-related changes.

Third, in our sample, changes in risk may simply reflect measurement error. In other words, youths' changes in risk may have simply been attributable to the imprecision of the tools. Indeed, we found that, in most of our models, participants' average risk scores predicted reoffending, whereas changes in risk scores did not. This may suggest that SAVRY and YLS-CMI total scores are assessing aspects of risk that are static over time, and thus, we would expect that *less* variability in reassessment ratings would improve predictions. We did not find support for this explanation in our post-hoc analyses. However, this issue merits further attention, especially as it can be difficult to differentiate true change from measurement error. To help determine if an individual has shown true or reliable change, some authors have started to use reliable change indices (e.g., Draycott, Kirkpatrick, & Askari, 2012; Olver, Beggs, & Wong, 2015; Viljoen et al., 2012).

According to the final hypothesis, the familiarity hypothesis, reassessments might improve predictions because they provide evaluators with more information about a youth.

Again, our results did not support this hypothesis. Predictive validity did not improve with increased familiarity with youth or over time. In addition, initial SAVRY and YLS/CMI risk scores achieved similar predictive validity as average risk scores. One possible explanation is that RAs had already acquired detailed information at the initial risk assessment and thus reassessment may not have contributed much new information. Another possibility is that RAs did not adequately adjust their reassessments based on new information gained via the reassessment. Overall, however, our failure to find support for this hypothesis is consistent with prior research with adults (e.g., Hanson, 2015).

In interpreting these findings, it is important to consider study limitations. One limitation is missing data. That said, our follow-up rate is comparable to other studies. For instance, in the MacArthur Violence Risk Assessment Study, 83.9% of participants completed at least one follow-up (Monahan, Steadman, & Silver, 2001). In the present study, 92.9% of participants had at least one follow-up. To minimize the effects of missing data, we coded some risk assessments based on file-information only (13.0%, n = 61), and used MLM, which incorporates all available time points (Hedeker & Gibbons, 2006). Another limitation is that, although we measured reoffending mainly through official records, this approach underestimates offending (Farrington et al., 2007). Conversely, self-reported offending, which was examined in the MLM models, also have limitations, such as memory difficulties or reporting biases. Finally, we focused only on total scores, as we were interested in tools as a whole. It may be that certain risk factors or risk domains are more predictive than other factors. Moreover, the timing of change may matter. For instance, some risk factors may have proximal effects, whereas others may have more distal or lagged effects. Thus, this is a focus of our ongoing work.

Some potential limitations in the generalizability of findings are also important to note. In particular, the youth in our sample were receiving the usual services in community settings. As such, our findings may not generalize to residential settings or samples of youth who are receiving empirically-supported interventions. We had only a small sample of female youth. Also, given that our sample was very diverse (e.g., 39% Caucasian/European, 30% Aboriginal, 13% Asian, 7% East Indian, 7% Hispanic, 5% African), we were unable to meaningfully compare findings across ethnic groups. However, this is an important area for future research, especially as we cannot assume that tools function equivalently across groups (Gutierrez, Wilson, Rugge, & Bonta, 2013; Shepherd, Adams, McEntyre, & Walker, 2014).

Although the assessment of changes in adolescents' risk scores did not improve predictions in this study, absence of evidence is not evidence of absence (Altman & Bland, 1995). Before drawing conclusions, researchers should pursue several possibilities. First, researchers should evaluate the extent to which youth show meaningful short-term changes in risk and if so, whether risk assessment tools are sensitive to these changes. If services are not effective or if tools are not sensitive to the ebb and flow of risk in everyday contexts (e.g., among youth receiving usual services), changes in risk might simply reflect measurement error. Second, researchers should explore study designs that may capture idiographic changes, such as single-case study designs (Barlow & Nock, 2009) and qualitative approaches (Carlsson, 2012). It would also be beneficial to test tools in the context of randomized control trials of interventions. Finally, researchers should test the utility of reassessments for intervention-planning. Even if reassessments do not improve risk predictions per se, they may assist in the delivery of interventions. In sum, although researchers have been quick to call some tools *dynamic* and assert that adolescents' risk rapidly changes, it may be more difficult to capture change than originally anticipated, emphasizing the need for further research.

References

Alam, S. (2015). Youth court statistics, 2013/2014. Juristat. Vol. 35, no. 1. Statistics Canada Catalogue no. 85-002-X ISSN 1209-6393.

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ: British Medical Journal (International Edition)*, *311*(7003), 485.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*(1), 19-21. doi:10.1111/j.1745-6924.2009.01088.x

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*(2), 182-192. doi:10.1037/a0022900

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology, 15*(4), 346-360. doi:10.1111/j.1468-2850.2008.00147.x

Blanchard, A. J. E. (2013). Dynamic risk factors in violence risk assessment: A multiple time-point evaluation of the HCR-20 and START (Unpublished Masters thesis). Simon Fraser University, Burnaby, BC, Canada.

Blanche, P., Dartigues, J. F., & Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, *32*(30), 5381-5397. doi: 10.1002/sim.5958

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

Calverley, D., Cotter, A., & Halla, E. (2015, November 30). *Youth custody and community services in Canada, 2008/2009.* Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2010001/article/11147-eng.htm

Carlsson, C. (2012). Using 'turning points' to understand processes of change in offending. *British Journal of Criminology*, *52*(1), 1-16.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3)*,* 494-509. doi:10.1037/0033-2909.114.3.494

Cohen, T. H., Lowenkamp, C. T., & VanBenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology and Public Policy, 15*(2), 263-296. doi:10.1111/1745-9133.12190

De Vries Robbé, M., de Vogel, V., Douglas, K. S., & Nijman, H. I. (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*(1), 53-61. doi:10.1037/lhb0000089

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

Douglas, K. S., & Skeem, J. L. (2005). Violence Risk Assessment: Getting Specific About Being Dynamic. *Psychology, Public Policy, And Law*, *11*(3), 347-383. doi:10.1037/1076-8971.11.3.347

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. doi:10.1080/14789949.2010.481720

Farrington, D. P., Jolliffe, D., Loeber, R., & Homish, D. L. (2007). How many offenses are really committed per juvenile court offender? *Victims and Offenders, 2*(3), 227-249. doi:10.1080/15564880701403934

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal, 345*(7868), 1-12. doi: http://dx.doi.org/10.1136/bmj.e4692

Fox, J., & Weiberg, S. (2011). Cox Proportion-Hazards Regression for Survival Data in R. On-line appendix, J. Fox and S. Weisberg, *An R Companion to Applied Regression* (Sage). Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/

Appendix-Cox-Regression.pdf

Goodman-Delahunty, J. & O'Brien, K. (2014). *Reoffence risk in intrafamilial child sexual offenders.* Report to the Criminology Research Advisory Council Grant: CRG 44/10-11. Canberra: Australian Institute of Criminology.

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*(1), 55-99. doi: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*(1), 29-36. doi: 10.1148/radiology.143.1.7063747

Hanson, R. K. (2015, October). *Do changes in STABLE-2007 Scores Predict Changes in Recidivism Risk?* Paper presented at the 35th Annual Research and Treatment Conference, Association for the Treatment of Sexual Abusers, Orlando, FL.

Hanson, R. K., Harris, A. J., Scott, T. L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (Vol. 5, No. 6). Ottawa, ON: Public Safety Canada.

Henggeler, S. W. (2016). Community-based interventions for juvenile offenders. In K. Heilbrun, D. DeMatteo, N. S. Goldstein, K. Heilbrun, D. DeMatteo, N. S. Goldstein (Eds.), *APA handbook of psychology and juvenile justice* (pp. 575-595). Washington, DC, US: American Psychological Association. doi:10.1037/14643-026

Heagerty, P. J. (2013). Survival ROC. *R Package Version 1.0.3*.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*(2), 337-344. doi: 10.1111/j.0006-341X.2000.00337.x

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92-105. doi: 10.1111/j.0006-341X.2005.030814.x

Hedeker, D. & Gibbons, R. D. (2006). *Longitudinal data analysis.* John Wiley & Sons: Hoboken, NJ.

Hilterman, E. B., Bongers, I., Nicholls, T. L., & van Nieuwenhuizen, C. (2016). Identifying gender specific risk/need areas for male and female juvenile offenders: Factor analyses with the Structured Assessment of Violence Risk in Youth (SAVRY). *Law And Human Behavior*, *40*(1), 82-96. doi:10.1037/lhb0000158

Hockenberry, S., & Puzzanchera, C. (2015). Juvenile Court Statistics 2013. Pittsburgh, PA: National Center for Juvenile Justice.

Hoge, R. D., & Andrews, D. A. (2002). *Youth Level of Service/Case Management Inventory: User's Manual.* Toronto, Canada: Multi Health Services.

Hoge R. D., Andrews D. A. (2011). *Youth Level of Service/Case Management Inventory 2.0 (YLS/CMI 2.0): User's manual*. Toronto, Ontario, Canada: Multi-Health Systems.

Huizinga, D., Esbensen, F., & Weiher, A. (1991). Are there multiple pathways to delinquency? Denver youth survey. *Journal of Criminal Law and Criminology, 82*(1)*,* 83-118. Doi: 0091 - 4 169/91/8201 0083

Jackson, K. M., & Sher, K. J. (2003). Alcohol use disorders and psychological distress: A prospective state-trait analysis. *Journal of Abnormal Psychology, 112*, 599-613.

Knight, G. P., Little, M., Losoya, S. H., & Mulvey, E. P. (2004). The self-report of offending among serious juvenile offenders cross-gender, cross-ethnic/race measurement equivalence. *Youth Violence and Juvenile Justice*, *2*(3), 273-295. doi:10.1177/1541204004265878

Labrecque, R. M., Smith, P., Lovins, B. K., & Latessa, E. J. (2014). The importance of reassessment: How changes in the LSI-R risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation*, *53*(2), 116-128. doi:10.1080/10509674.2013.868389

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*(2), 150-164. doi:10.1177/1073191112441242

Lloyd, C. D., & Serin, R. C. (2015, June). *Protective factors, relationship investment, and employment investment are related to reduced recidivism after accounting for static, stable, and acute risk factors*. Paper presented at the 3rd North American Correctional and Criminal Justice Psychology Conference, Ottawa, Ontario, Canada.

Maciejewski, D. F., van Lier, P. C., Branje, S. T., Meeus, W. J., & Koot, H. M. (2015). A 5-year longitudinal study on mood variability across adolescence using daily diaries. *Child*

Development, *86*(6), 1908-1921. doi:10.1111/cdev.12420

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

Meyers, J. R., & Schmidt, F. (2008). Predictive Validity of the Structured Assessment for Violence Risk in Youth (SAVRY) With Juvenile Offenders. *Criminal Justice & Behavior*, *35*(3), 344-355. doi:10.1177/0093854807311972

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), 1-13. doi:10.1080/14999013.2012.760182

Monahan, J., Steadman, H. J., & Silver, E. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence.* Cary, NC: Oxford University Press.

Mossman, D. (1994). Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*(4), 783. doi:10.1037/0022-006X.62.4.783

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*(1), 92-110. doi: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*(5), 472-499. doi:10.1177/1079063213502679

Olver, M. E., Nicholaichuk, T. P., Kingston, D. A., & Wong, S. P. (2014). A multisite examination of sexual violence risk and therapeutic change. *Journal of Consulting and Clinical Psychology*, *82*(2), 312-324. doi:10.1037/a0035340

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*(1), 156-176. doi:10.1037/a0035080

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*(4), 329-353. doi:10.1177/0093854809331457

Olver, M. E., Stockdale, K. C., & Wong, S. P. (2012). Short and long-term prediction of recidivism using the Youth Level of Service/Case Management Inventory in a sample of serious young offenders. *Law and Human Behavior*, *36*(4), 331-344. doi:10.1037/h0093927

Olver, M., Wong, S., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment*, *19*(3), 318-329. doi: 10.1037/1040-3590.19.3.318

O'Keefe, M. L., Klebe, K., & Hromas, S. (1998). *Validation of the Level of Supervision*

Inventory (LSI) for Community Based Offenders in Colorado. Denver, CO: Department of Corrections.

Poulin, F., & Chan, A. (2010). Friendship stability and change in childhood and adolescence. *Developmental Review*, *30*(3), 257-272. doi:10.1016/j.dr.2009.01.001

Prentky, R., & Righthand, S. (2003). *Juvenile Sex Offender Assessment Protocol II (J-SOAP-II) Manual.* Washington DC: U.S. Department of Justice.

Ralston, C. A., & Epperson, D. L. (2013). Predictive validity of adult risk assessment tools with juveniles who offended sexually. *Psychological Assessment, 25*(3) 905-916. doi:10.1037/a0032683

Raudenbush, S. W. Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, second edition.* Newbury Park, CA: Sage.

Raynor, P. (2007). Risk and need assessment in British probation: The contribution of LSI-R. *Psychology, Crime & Law*, *13*(2), 125-138. doi:10.1080/10683160500337592

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), 1-25. doi:10.1016/0149-7189(95)00037-2

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, *29*(5), 615. doi:10.1007/s10979-005-6832-7

Roberts, B., Walton, K., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1-25. doi:10.1037/0033-2909.132.1.1

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*(1), 77-84. doi:10.1186/1471-2105-12-77

Romano, J., Kromrey, J. D., Coraggio, J., Skowronek, J., & Devine, L. (2006). *Exploring methods for evaluating group differences in the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices?* Paper presented at the Southern Association for Institutional Research, Arlington, VA

SAS Institute Inc. (2012). SAS 12.1 [computer software]. Cary, NC: SAS Institute Inc.

Schemper, M., Wakounig, S., & Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, *28*(19), 2473-2489. doi:10.1002/sim.3623

Schmidt, F., Campbell, M. A., & Houlding, C. (2011). Comparative analyses of the YLS/CMI, SAVRY, and PCL: YV in adolescent offenders: A 10-year follow-up into adulthood.*Youth Violence And Juvenile Justice*, *9*(1), 23-42. doi:10.1177/1541204010371793

Schwalbe, C. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31*(5), 449-462. doi:10.1007/s10979-006-9071-7

Schwalbe, C. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice and Behavior, 35*(11), 1367-1381. doi:10.1177/0093854808324377

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

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*(3), 193-206. doi:10.1080/14999013.2014.922141

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499-513. doi:10.1016/j.cpr.2010.11.009

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*(1), 15-22. doi:10.1037/lhb0000090

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*(3), 281-293. doi:10.1037/law0000017

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi:10.1037//0033-2909.86.2.420.

Smith, P., Cullen, F., & Latessa, E. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology and Public Policy, 8*(1), 183-208. doi: 10.1111/j.1745-9133.2009.00551.x

Stockdale, K. C. (2008). *The validity and reliability of the Violence Risk Scale-Youth Version (VRS-YV)* (Doctoral dissertation). University of Saskatchewan, Saskatoon, Canada.

Sweeten, G., Piquero, A. R., & Steinberg, L. (2013). Age and the explanation of crime, revisited. *Journal Of Youth And Adolescence*, *42*(6), 921-938. doi:10.1007/s10964-013-9926-4

Takahashi, M., Mori, T., & Kroner, D. G. (2013). A cross-validation of the Youth Level of Service/Case Management Inventory (YLS/CMI) among Japanese juvenile offenders. *Law and Human Behavior, 37*(6)*,* 389-400. doi:10.1037/1hb0000029

Torchiano, M. (2015). Effsize: Efficient Effect Size Computation. *R Package Version 0.5*, *4*.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, *56*(4), 1134-1138. doi: 10.1111/j.0006-341X.2000.01134.x

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. doi: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*(3), 135-149. doi:10.1080/14999013.2012.737406

Viljoen, J. L., Gray, A. L., Shaffer, C., Latzman, N., Scalora, M., & 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: Journal of Research and Treatment.* Prepublished July 21, 2015, doi: 10.1177/1079063215595404

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(*5), 423-438. doi:10.1037/h0093938

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment, 17*(3)*,* 377-395. doi:10.1177/1073191109359587

Viljoen, J. L., Nicholls, T. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (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.* Washington, DC: Models for Change.

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*(12), 1383-1396. doi: 10.1177/0093854813508916

Walters, G. D., Deming, A., & Casbon, T. (2015). Predicting recidivism in sex offenders with the psychological inventory of criminal thinking styles (PICTS). *Assessment*, *22*(2), 167-177. doi:10.1177/1073191114539384

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Desmarais, S. L. (2009). *Manual for the Short-Term Assessment of Risk and Treatability* (START; Version 1.1). Port Coquitlam, British Columbia, Canada: British Columbia Mental Health and Addiction Services.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior*, *37*(6), 377-388. doi:10.1037/lhb0000025

Wong, S., & Gordon, A. E. (1999-2003). *The Violence Risk Scale.* Unpublished manuscript, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

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

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

Worling, J. R., Bookalam, D., & Litteljohn, A. (2012). Prospective validity of the Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR). *Sexual Abuse: Journal of Research and Treatment*, *24*(3), 203-223. doi: 10.1177/1079063211407080

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*(5), 740-767. doi:10.1037/a0020473

Table 1

SAVRY Predictive Validity Analyses for the Presence of Reoffending Over a Two-Year Follow-Up

						Cox Proportional Hazards Model				
SAVRY	$\hat{\delta}$	AUC	95% CI_{AUC}	r_s^{Time}	r_s^{Charge}	*B*	*SE*	Wald	HR	95% CI_{HR}
Violent Charge										
Historical	.23	.62*	[.52, .72]	-.11	.20*	0.10	0.05	4.15*	1.10	[1.00, 1.21]
Social/Contextual[a]	.34	.67***	[.57, .77]	-.11	.24**	0.23	0.08	8.35**	1.26	[1.08, 1.48]
Individual/Clinical[a]	.51	.76***	[.67, .84]	-.22**	.36***	0.29	0.07	18.05***	1.34	[1.17, 1.53]
Protective[b]	-.23	.62*	[.52, .71]	.16*	-.17*	-0.37	0.17	4.88*	0.69	[0.50, 0.96]
Risk Total Score	.42	.71***	[.61, .81]	-.17*	.32***	0.09	0.03	13.58***	1.10	[1.05, 1.15]
Risk Rating	.32	.66***	[.56, .75]	-.15	.26**	0.76	0.29	7.09**	2.14	[1.22, 3.73]
Any Charge										
Historical	.41	.70***	[.62, .79]	-.36***	.37***	0.14	0.03	20.65***	1.15	[1.09, 1.23]
Social/Contextual[a]	.32	.66***	[.58, .74]	-.29***	.31***	0.19	0.05	13.47***	1.21	[1.09, 1.34]
Individual/Clinical[a]	.50	.75***	[.67, .82]	-.41***	.45***	0.21	0.04	29.00***	1.24	[1.15, 1.34]
Protective[b, c]	-.21	.61*	[.52, .69]	.26**	-.19*	-0.27	0.10	7.82**	0.76	[0.63, 0.92]
Risk Total Score	.50	.75***	[.67, .82]	-.42***	.45***	0.09	0.02	31.60***	1.10	[1.06, 1.14]
Risk Rating	.46	.73***	[.66, .80]	-.41***	.46***	1.00	0.19	27.20***	2.70	[1.87, 3.95]

Note. $\hat{\delta}$ = Cliff's delta; AUC = area under the curve; CI_{AUC} = confidence interval of AUC; r_s^{Time} = Spearman's rho correlation with time-at-risk; r_s^{Charge} = Spearman's rho correlation with number of charges; *B* = regression coefficient; *SE* = standard error of *B*; HR = hazard ratio; CI_{HR} = confidence interval of the HR. [a] $n = 155$ [b] For ease of interpretation, scores on the Protective Factors domain were reversed for the AUC analysis such that higher scores represent a deficit in protective factors. [c] Due to violating the proportional hazards assumption, we re-ran this analysis using weighted estimation for Cox regression (Schemper, Wakounig, & Heinze, 2009). The results closely approximated those already reported in the Table ($B = -0.31$, $SE = 0.09$, Wald = 10.87, $p < .001$, HR = 0.74, 95% CI [0.61, 0.88]). * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 2

YLS/CMI Predictive Validity Analyses for the Presence of Reoffending Over a Two-Year Follow-Up

						Cox Proportional Hazards Model				
YLS/CMI	$\hat{\delta}$	AUC	95% CI_{AUC}	r_s^{Time}	r_s^{Charge}	*B*	*SE*	Wald	HR	95% CI_{HR}
Violent Charge										
Prior/Current Offenses[a]	.28	.64*	[.52, .76]	-.16*	.23**	0.36	0.11	10.86***	1.44	[1.16, 1.78]
Family Circumstances[b]	.15	.57	[.47, .68]	-.03	.12	0.15	0.11	1.93	1.16	[0.94, 1.44]
Education/Employment	.18	.59	[.47, .70]	-.08	.14	0.15	0.10	2.06	1.16	[0.95, 1.42]
Peer Relations	.36	.68***	[.60, .75]	-.28***	.27**	0.83	0.28	8.93**	2.30	[1.33, 3.96]
Substance Abuse	.21	.60	[.50, .71]	-.06	.13	0.22	0.12	3.32	1.25	[0.98, 1.59]
Leisure and Recreation	.31	.66**	[.56, .75]	-.14	.24**	0.60	0.22	7.69**	1.82	[1.19, 2.78]
Personality/Behavior	.35	.67***	[.57, .77]	-.04	.25**	0.35	0.12	9.00**	1.42	[1.13, 1.78]
Attitude/Orientation	.37	.68***	[.59, .78]	-.12	.27**	0.42	0.14	8.62**	1.52	[1.15, 2.00]
Risk Total Score	.45	.72***	[.63, .82]	-.15	.33***	0.11	0.03	13.84***	1.11	[1.05, 1.18]
Risk Rating	.35	.68***	[.58, .77]	-.14	.28***	0.80	0.26	9.67**	2.22	[1.34, 3.66]
Any Charge										
Prior/Current Offenses	.37	.68***	[.60, .77]	-.36***	.40***	0.39	0.07	28.56***	1.48	[1.28, 1.72]
Family Circumstances[b]	.22	.61*	[.53, .70]	-.19*	.19*	0.19	0.07	7.14**	1.21	[1.05, 1.40]
Education/Employment	.19	.59*	[.51, .68]	-.22**	.18*	0.18	0.07	6.82**	1.20	[1.05, 1.37]
Peer Relations	.31	.65***	[.58, .73]	-.31***	.32***	0.43	0.12	12.58***	1.54	[1.21, 1.95]
Substance Abuse	.29	.64***	[.56, .73]	-.22**	.27**	0.25	0.08	9.71**	1.28	[1.10, 1.50]
Leisure and Recreation	.26	.63**	[.55, .71]	-.24**	.25**	0.39	0.13	9.80**	1.48	[1.16, 1.90]
Personality/Behavior	.42	.71***	[.63, .79]	-.32***	.39***	0.33	0.07	20.95***	1.40	[1.21, 1.61]
Attitude/Orientation	.37	.68***	[.60, .76]	-.31***	.34***	0.43	0.10	19.04***	1.53	[1.26, 1.85]
Risk Total Score	.37	.75***	[.70, .82]	-.42***	.45***	0.11	0.02	31.73***	1.11	[1.07, 1.16]
Risk Rating	.46	.73***	[.66, .80]	-.43***	.47***	0.10	0.17	32.99***	2.66	[1.90, 3.71]

Note. $\hat{\delta}$ = Cliff's delta; AUC = area under the curve; CI_{AUC} = confidence interval of AUC; r_s^{Time} = Spearman's rho correlation with time-at-risk; r_s^{Charge} = Spearman's rho correlation with number of charges; *B* = regression coefficient; *SE* = standard error of *B*; HR = hazard ratio; CI_{HR} = confidence interval of the HR. [a] Due to violating the proportional hazards assumption, we re-ran this using weighted estimation for Cox regression (Schemper et al., 2009). The results were very similar ($B = 0.38$, $SE = 0.12$, Wald = 10.38, $p < .01$, HR = 1.46, 95% CI [1.16, 1.84]). [b] $n = 155$ * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test).

Table 3

Shelf-Life Hypothesis: Predictive Validity of SAVRY and YLS/CMI over Various Follow-ups

	Violent Charge		Any Charge	
	AUC	95% CI_{AUC}	AUC	95% CI_{AUC}
SAVRY Dynamic Risk Total				
Baseline – 3 Months	.74*	[.55, .93]	.72***	[.63, .81]
3 – 6 Months	.70**	[.56, .85]	.71***	[.61, .80]
6 – 12 Months	.75***	[.63, .87]	.70***	[.60, .79]
12 – 24 Months	.75***	[.63, .86]	.77***	[.69, .85]
SAVRY Risk Total				
Baseline – 3 Months	.78**	[.60, .95]	.73***	[.64, .82]
3 – 6 Months	.72*	[.55, .90]	.72***	[.62, .81]
6 – 12 Months	.72***	[.59, .85]	.71***	[.62, .80]
12 – 24 Months	.70**	[.58, .82]	.75***	[.67, .84]
YLS/CMI Dynamic Risk Total				
Baseline – 3 Months	.66**	[.55, .76]	.74***	[.66, .82]
3 – 6 Months	.72***	[.59, .85]	.70***	[.60, .80]
6 – 12 Months	.66**	[.54, .79]	.67***	[.58, .76]
12 – 24 Months	.73***	[.60, .86]	.70***	[.61, .79]
YLS/CMI Risk Total				
Baseline – 3 Months	.75***	[.63, .86]	.76***	[.68, .84]
3 – 6 Months	.75**	[.59, .90]	.72***	[.62, .82]
6 – 12 Months	.68**	[.56, .81]	.71***	[.62, .80]
12 – 24 Months	.72***	[.59, .86]	.72***	[.63, .81]

Note. CI_{AUC} = confidence interval of AUC. * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed test). Base rates for violent and any reoffending were as follows: baseline to 3 months was 3.8% and 23.1%, respectively; 3 – 6 months was 5.1% and 17.3%, respectively; 6 – 12 months was 9.6% and 19.2%, respectively; and 12 – 24 months was 12.8% and 25.6%, respectively.

Table 4

Dynamic Change Hypothesis: Within-Person Effects of SAVRY Risk Total on Self-Reported Reoffending Over the Subsequent 3-Months

	Within-Person SAVRY Change (Model 3)			Between vs. Within-Person Effects (Model 4)			Interaction Between and Within-Person (Model 5)		
Violent Reoffending									
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***
Intercept	0.34	0.23	125	-3.22**	0.50	124	-3.21**	0.50	124
Time	-0.35**	0.10	106	-0.18*	0.08	106	-0.18*	0.08	106
SAVRY Risk Change	-0.02	0.02	144	-0.02	0.02	144	-0.05	0.09	143
Mean SAVRY Risk	–	–	–	0.13**	0.02	144	0.13**	0.02	143
SAVRY x *Mean SAVRY*	–	–	–	–	–	–	0.00	0.00	143
Variance Components	σ^2_υ	**SE**		σ^2_υ	**SE**		σ^2_υ	**SE**	
Intercept	1.81**	0.56	–	1.59**	0.48	–	1.60**	0.48	–
Time	0.16**	0.07	–	0.12*	0.05	–	0.12*	0.05	–
Model Fit									
χ^2/df	0.68			0.77			0.77		
Any Reoffending									
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***Df***
Intercept	1.08**	0.19	127	-1.76**	0.42	124	-2.10**	0.38	124
Time	-0.17*	0.07	108	-0.13	0.07	106	–	–	–
SAVRY Risk Change	-0.05	0.03	79	-0.05	0.03	79	-0.04	0.10	80
Mean SAVRY Risk	–	–	–	0.20	0.01	73	0.11**	0.01	72
SAVRY x *Mean SAVRY*	–	–	–	–	–	–	-0.00	0.00	72
Variance Components	σ^2_υ	**SE**		σ^2_υ	**SE**		σ^2_υ	**SE**	
Intercept	2.53**	0.55	–	2.02**	0.45	–	2.31**	0.48	–
Time	0.24**	0.06	–	0.22**	0.06	–	0.23**	0.06	–
SAVRY Risk	0.01*	0.01	–	0.01*	0.01	–	0.02*	0.01	–
Model Fit									
χ^2/df	0.85			0.89			0.84		

Note. Level 2 (between-person) effects are italicized. * $p < .05$, ** $p < .01$ (two-tailed). The within-person effects representing changes in SAVRY Risk Total Scores from the persons' mean SAVRY Risk Total Score is underlined, as this was the key variable tested in these analyses.

Table 5

Dynamic Change Hypothesis: Within-Person Effects of YLS/CMI Risk Total on Self-Reported Reoffending Over the Subsequent 3-Months

	Within-Person YLS/CMI Change (Model 3)			**Between vs. Within-Person Effects (Model 4)**			**Interaction Between and Within-Person (Model 5)**		
Violent Reoffending									
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***
Intercept	0.18	0.24	128	-2.50**	0.45	125	-2.43**	0.45	125
Time	-0.30**	0.10	106	-0.19*	0.09	104	-0.22*	0.09	103
<u>YLS/CMI Risk Change</u>	-0.09	0.05	80	-0.05	0.04	81	-0.21	0.13	80
Mean YLS/CMI Risk	–	–	–	0.14**	0.02	63	0.14**	0.02	64
YLS x *Mean YLS*	–	–	–	–	–	–	0.01	0.01	64
Variance Components	**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**	
Intercept	2.08**	0.65	–	1.74**	3.24	–	1.38**	0.53	–
Time	0.17*	0.07	–	0.14*	2.29	–	0.15*	0.06	–
YLS/CMI Risk	0.04*	0.02	–	0.03*	2.33	–	0.03*	0.01	–
Model Fit									
χ^2/df	0.42			0.48			0.47		
Any Reoffending									
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***
Intercept	1.09**	0.20	128	-1.25**	0.37	125	-1.61**	0.33	125
Time	-0.18*	0.07	106	-0.14	0.07	104	–	–	–
<u>YLS/CMI Risk Change</u>	-0.06	0.03	81	-0.06	0.03	81	-0.11	0.09	81
Mean YLS/CMI Risk	–	–	–	0.12**	0.02	71	0.12**	0.02	71
YLS x *Mean YLS*	–	–	–	–	–	–	0.00	0.00	71
Variance Components	**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**	
Intercept	3.03**	0.62	–	2.65**	0.54	–	3.01**	0.58	–
Time	0.26**	0.06	–	0.25**	0.06	–	0.26**	0.06	–
YLS/CMI Risk	0.03**	0.01	–	0.03**	0.01	–	0.03**	0.01	–
Model Fit									
χ^2/df	0.52			0.55			0.54		

Note. Level 2 (between-person) effects are italicized. Under dispersion in the model may have inflated the rate of Type II errors. * $p < .05$, ** $p < .01$ (two-tailed). The within-person effects representing changes in YLS/CMI Risk Total Scores from the persons' mean YLS/CMI Risk Total Score is underlined, as this was the key variable tested in these analyses.

Table 6

Familiarity Hypotheses: Familiarity/Risk Totals Interactions and Self-Reported Reoffending Over the Subsequent 3-Months

	SAVRY						YLS/CMI					
	Narrow Definition of Familiarity (Model 1)			**Broad Definition of Familiarity (Model 2)**			**Narrow Definition of Familiarity (Model 1)**			**Broad Definition of Familiarity (Model 2)**		
Violent Reoffending												
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***
Intercept	1.04**	0.29	125	-0.70**	0.22	125	0.74*	0.32	126	0.17	0.25	126
Time	-0.55**	0.12	106	-0.35**	0.10	143	-0.46**	0.12	104	-0.29**	0.10	104
Narrow Familiarity	-0.56	0.14	142	–	–	–	-0.49**	0.17	62	–	–	–
Risk Total	0.01	0.03	142	-0.01	0.03	143	-0.07	0.06	80	-0.06	0.56	80
Risk x Familiarity	-0.07	0.04	142	–	–	–	-0.04	0.06	62	–	–	–
Risk x Time	–	–	–	0.01	0.02	143	–	–	–	0.03	0.03	63
Variance Components	**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**	
Intercept	1.75**	0.57	–	1.75**	0.56	–	2.36**	0.73	–	2.14**	0.67	–
Time	0.16*	0.07	–	0.16**	0.07	–	0.19**	0.08	–	0.17*	0.07	–
Risk[a]	–	–	–	–	–	–	0.03*	0.02	–	0.04**	0.02	–
Model Fit (χ^2 /df)	0.65			0.69			0.39			0.42		
Any Reoffending												
Fixed Effects	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***	**β**	**SE**	***df***
Intercept	1.29**	0.23	125	0.56**	0.17	125	1.20**	0.24	126	0.54**	0.17	126
Time	-0.23**	0.08	106	-0.17*	0.07	72	-0.21*	0.08	104	-0.18*	0.08	70
Narrow Familiarity	-0.17	0.10	71	–	–	–	-0.10	0.11	69	–	–	–
Risk Total	-0.06	0.03	79	-0.07*	0.02	79	-0.06	0.04	81	-0.06	0.04	81
Risk x Familiarity	0.03	0.03	71	–	–	–	-0.01	0.04	69	–	–	–
Risk x Time	–	–	–	-0.02	0.02	72	–	–	–	0.01	0.02	70
Variance Components	**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**		**σ^2_υ**	**SE**	
Intercept	2.64**	0.57	–	2.57**	0.56	–	3.13**	0.65	–	3.08**	0.64	–
Time	0.23**	0.06	–	0.24**	0.06	–	0.27**	0.07	–	0.27**	0.07	–
Risk	0.01*	0.01	–	0.01*	0.01	–	0.03**	0.01	–	-0.06**	0.01	–
Model Fit (χ^2 /df)	0.86			0.85			0.52			0.52		

Note. Under dispersion in the model may have inflated the rate of Type II errors. * $p < .05$, ** $p < .01$ (two-tailed). The interaction terms are underlined, as these were the key variables tested in this set of analyses.

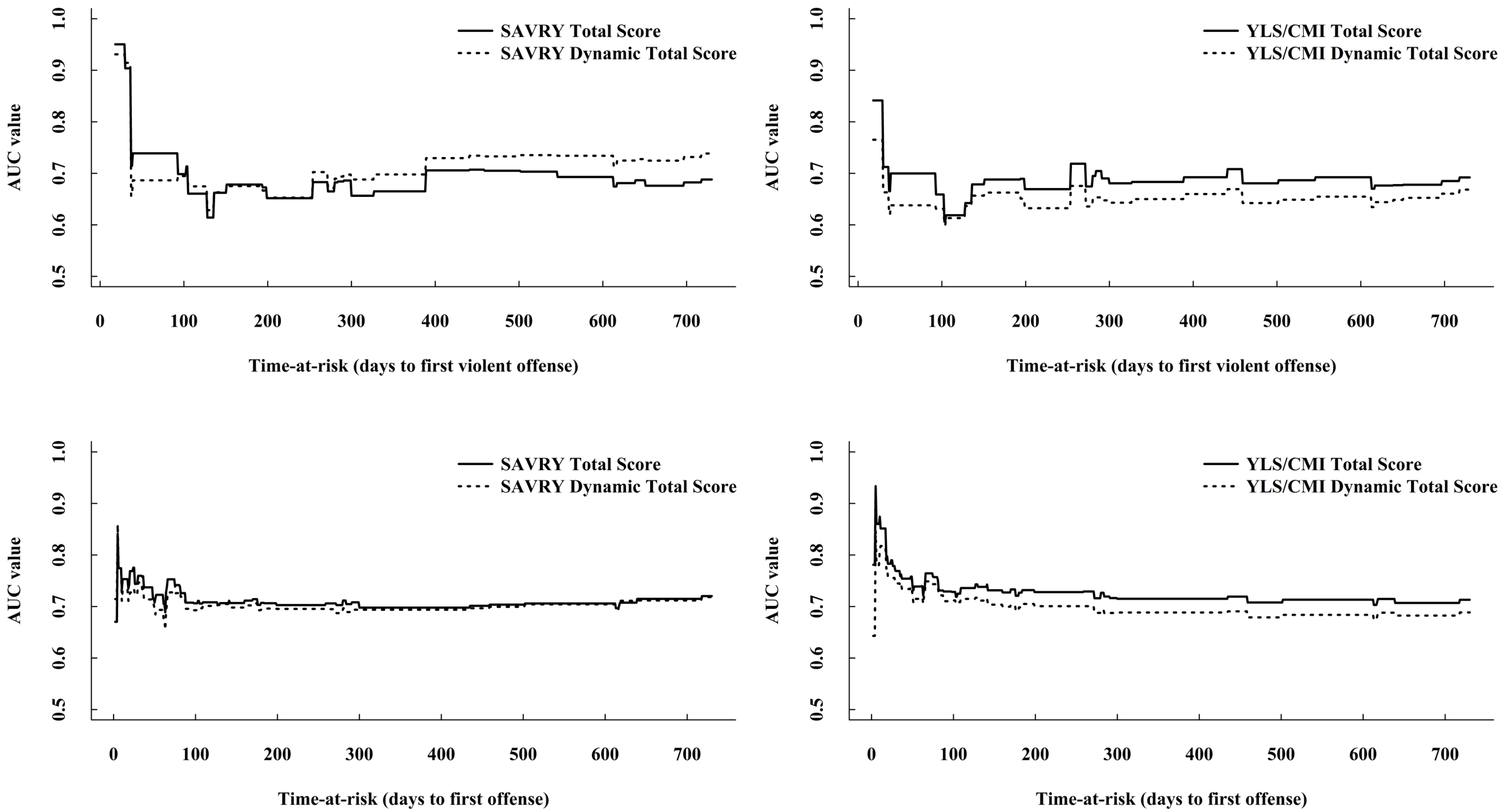

Figure 1. Shelf-life hypothesis: Time-dependent AUC analysis predicting reoffending over a two-year follow-up (*N* = 156). The extreme fluctuations in AUC values exhibited at the outset of the follow-up period is likely a reflection of the risk scores of the first few reoffenders and does not appear to reflect a meaningful trend.

Endnotes

[1] The seven excluded youth did not differ significantly from other participants in age, gender, ethnicity, index offense, prior charges, or SAVRY and YLS/CMI Risk Total Scores.

[2] In the present study, we eliminated the question "shot and killed someone" due to the low base rate of this event and because of concerns regarding confidentiality and subpoenas.

[3] In the model with self-reported violence as an outcome, SAVRY Risk Total Scores were not modeled as a variance component because the variance coefficient associating SAVRY Risk Total Scores with violent reoffending was not statistically significant ($\sigma^2_{\upsilon} = 0.01$, SE = 0.01, $p > .05$) suggesting a lack of variability in slopes (i.e., the relationship between SAVRY Risk Total Scores and violent reoffending did not differ randomly across individuals).

[4] The average Risk Total Score was calculated as the average for the baseline, 3, 6, 9, and 12-month follow-ups. Thus, to keep the follow-ups prospective, we examined whether the average risk total score predicted reoffending in the period of 12 – 24 months. To help ensure the lag was comparable for the initial risk total score, we examined whether the initial assessment predicted reoffending in the period of 6 – 18 months from the start of the study.