Article

Reliability and Validity of START and LSI-R Assessments in Mental Health Jail Diversion Clients

Assessment
2019, Vol. 26(7) 1347–1361

DOI: 10.1177/1073191117704505
journals.sagepub.com/home/asm
SAGE

Evan M. Lowder[1], Sarah L. Desmarais[1], Candalyn B. Rade[1], Kiersten L. Johnson[2], and Richard A. Van Dorn[2]

Abstract

Risk assessment instruments are increasingly used in mental health jail diversion programs. This study examined the reliability and validity of Short-Term Assessment of Risk and Treatability (START) and Level of Service Inventory–Revised (LSI-R) assessments overall and by client race. Research assistants completed START and LSI-R assessments for 95 diversion clients. Arrests and jail days were collected via official records and self-report 3, 6, 9, 12, and 18 months after baseline. Assessments demonstrated good interrater reliability and convergent validity. START strength total scores and LSI-R risk estimates were the strongest predictors of recidivism. Total scores and risk estimates did not differ as a function of client race, but there were some differences in accuracy of START vulnerability and LSI-R total scores and risk estimates in predicting jail days (but not arrests), over shorter follow-ups. No such differences were found for START strength total scores across any follow-up period or recidivism measure.

Keywords

mental health jail diversion, risk assessment, protective factors, recidivism

Developed in response to rising numbers of adults with mental illnesses in jails and high rates of reoffending within this population (Baillargeon, Binswanger, Penn, Williams, & Murray, 2009; Steadman, Osher, Robbins, Case, & Samuels, 2009), over 500 mental health diversion programs exist across the United States (Case, Steadman, Dupuis, & Morris, 2009). Similar programs are found in other countries, such as Canada and Australia, as well (Richardson & McSherry, 2010; Slinger & Roesch, 2010). These programs are distinguished from traditional criminal case processing in that they seek to identify detainees with mental illness, use this information as a factor in case disposition, and connect these detainees to community-based treatment (Steadman, Barbera, & Dennis, 1994). Increasingly, diversion programs are being required to use instruments designed to assess a client's risk for violence and criminal behavior, yet few studies have investigated the use of risk instruments in these settings (Skeem & Monahan, 2011).

Because risk assessment instruments do not intrinsically hold psychometric properties, validation is critical to establishing support for their use within new settings and with specific populations (American Educational Research Association, 2014). A growing body of literature suggests justice-involved adults with mental illnesses are a unique population in the criminal justice system that have increased general risk factors for recidivism and behavioral health needs (Skeem, Winter, Kennealy, Louden, & Tatar, 2014). To date, however, very little peer-reviewed research has reported on the reliability and validity of risk assessment instruments in mental health jail diversion settings. In fact, less than a handful of studies have examined the psychometric properties of risk assessment instruments in mental health diversion programs in the United States (Barber-Rioja, Dewey, Kopelovich, & Kucharski, 2012; Canales, Campbell, Wei, & Totten, 2014; Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012). Two instruments that have been implemented in these settings are the Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Desmarais, 2009) and the Level of Service Inventory–Revised (LSI-R; Andrews & Bonta, 2001).

[1]North Carolina State University, Raleigh, NC, USA
[2]RTI International, Research Triangle Park, NC, USA

Corresponding Author:
Evan M. Lowder, Department of Psychology, North Carolina State University, 640 Poe Hall, Campus Box 7650, Raleigh NC 27695, USA.
Email: emlowder@ncsu.edu

Short-Term Assessment of Risk and Treatability

The START is a structured professional judgment instrument that guides assessment of risk for outcomes beyond violence alone, including suicide, self-harm, victimization, substance use, unauthorized leave, and self-neglect. As part of a larger project, the START additionally was being used to inform assessments of risk for general offending in the current study. Although still relatively new, the START has experienced quick uptake into practice. The START has been translated into nine languages and is being used in behavioral health and correctional settings, as well as private practices, in more than a dozen countries (Nicholls, Desmarais, Martin, Brink, & Webster, 2006). Compared with other risk assessment tools, the START is unique in its consideration of vulnerabilities *and* strengths for each item. Additionally, START assessments focus on short-term risk (i.e., weeks to months) rather than long-term risk (i.e., months to years). These distinctive aspects of the START make the approach particularly relevant to community corrections settings, such as mental health jail diversion programs, where treatment planning and intervention are prioritized and the periods of supervision are shorter.

Prior research has established generally high internal consistency and interrater reliability of START assessments. To demonstrate, in a recent systematic review, studies ($N = 7$) examining internal consistency of strength and vulnerability total scores reported Cronbach alpha coefficients of .80 to .95 and .76 to .95, respectively (O'Shea & Dickens, 2014). Findings on interrater reliability, when reported, have been similarly high. In the same review, intraclass correlations were found to be high for strength total scores (.78), vulnerability total scores (.86), and risk estimates (.82). Prior research also provides strong evidence of the validity of START assessments in predicting short- and medium-term aggression and violence among civil and forensic psychiatric inpatients (Braithwaite, Charette, Crocker, & Reyes, 2010; Chu, Thomas, Ogloff, & Daffern, 2011a, 2011b; Desmarais, Nicholls, Wilson, & Brink, 2012; O'Shea & Dickens, 2014; Wilson, Desmarais, Nicholls, & Brink, 2010). Moreover, studies have indicated strong predictive validity of START assessments among other vulnerable populations, such as among offenders with an intellectual disability (Inett, Wright, Roberts, & Sheeran, 2014).

Desmarais, Van Dorn, et al. (2012) recently investigated characteristics of START assessments completed on 96 mental health jail diversion clients in the context of routine practice. The authors found preliminary support for the use of START assessments in diversion settings; for example, findings revealed good internal consistency of START assessments ($\alpha = .95$ for strength total scores; $\alpha = .90$ for vulnerability total scores), and characteristics of START assessments, such as mean vulnerability total scores, were similar to those conducted in other settings (e.g., forensic hospitals). Moreover, findings suggested the need for investigation into an eighth risk domain to capture general offending risk in this population. However, to date, no study has examined the predictive validity of START assessments completed on mental health jail diversion program clients or in predicting general, nonviolent offending in any population. Furthermore, few studies to date have examined the psychometric properties of START assessments in correctional samples, especially those in community settings, or among racial minorities more broadly.

Level of Service Inventory–Revised

The LSI-R is an actuarial risk assessment instrument part of a larger family of Level of Service instruments developed from the Risk-Need-Responsivity model of effective offender rehabilitation and designed to estimate general offending risk (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990; Andrews, Bonta, & Wormith, 2010). These instruments additionally include the Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004), the Level of Service/Risk-Need-Responsivity (LS/RNR; Andrews, Bonta, & Wormith, 2008), and Level of Service-Inventory–Revised: Screening Version (Andrews & Bonta, 1998). Although more recently published, the LS/RNR and LS/CMI are more suitable for correctional practice and case management due to their focus on specific risks, needs, and responsivity considerations. The LSI-R, in contrast, is more readily employed in research. Indeed, whereas few studies have examined the predictive validity of LS/CMI or LS/RNR assessments in correctional settings, at least 25 published studies have reported on the psychometric properties of LSI-R assessments vis-à-vis recidivism in correctional samples (Desmarais, Johnson, & Singh, 2016), providing context and a point of reference for validation of the LSI-R in unique settings and samples.

The LSI-R includes 54 items measuring both static and dynamic risk factors across 10 separate risk domains (e.g., Financial, Education/Employment, Criminal History, etc.; Andrews & Bonta, 2001). Total scores on the instrument classify offenders into five distinct risk bins: Low, Low/Moderate, Moderate, Moderate/High, and High. In research and practice, these have been collapsed further into risk bins of Low, Moderate, and High (Canales et al., 2014; Singh, Grann, & Fazel, 2011). The LSI-R is one of the most widely used and well-known risk assessment instruments; to demonstrate, a recent survey by Multi-Health Systems reported that over 900 criminal justice agencies regularly use the LSI-R in practice (Lowenkamp, Lovins, & Latessa, 2009).

Surprisingly, despite its widespread use, limited published research has reported on the reliability of LSI-R

assessments. Initial, unpublished studies, including initial validation research conducted by the instrument's primary author, reported high levels of interrater reliability (*r* range = 0.87-0.92) for LSI-R assessments conducted in correctional samples (for a review, see Andrews et al., 2010). However, in more recent investigations, when reported in published research, levels of interrater reliability and internal consistency have been highly variable. For example, Lowenkamp et al. (2004) found LSI-R subsections to have varying rates of agreement (61.5% to 97.7%) among correctional practitioners. Furthermore, levels of internal consistency also have been found to be less reliable among racial and ethnic minorities (Schlager & Simourd, 2007), who make up about half of mental health jail diversion clients (e.g., Case et al., 2009). In contrast, a substantial body of research (i.e., over 100 studies) supports the validity of LSI-R assessments in predicting general recidivism overall (Olver, Stockdale, & Wormith, 2014; Singh et al., 2011; Vose, Cullen, & Smith, 2008), as well as across various subgroups of offenders, including male offenders, female offenders, and ethnic minorities (Olver et al., 2014). However, in some studies, LSI-R assessments have demonstrated poorer predictive validity among racial minorities (Chenane, Brennan, Steiner, & Ellison, 2015; Schlager & Simourd, 2007; Whiteacre, 2006). There is also relatively limited research on the predictive validity of LSI-R assessments among adults with mental illnesses (Ferguson, Ogloff, & Thomson, 2009).

To our knowledge, only one study has examined the validity of LSI-R assessments in the context of mental health jail diversion programs. Specifically, Canales et al. (2014) investigated the predictive validity of LS/RNR (Andrews et al., 2008) assessments conducted in a Canadian mental health court. Results showed that assessments of General Risk/Need produced strong levels of predictive validity for general recidivism, particularly among male clients. However, this study was limited in its investigation of the LS/RNR scales and not the more frequently used LSI-R. To our knowledge, no research to date has examined the predictive validity or interrater reliability of LSI-R assessments completed on mental health jail diversion clients in the United States.

Racial Bias in Risk Assessment

In the context of risk assessment, there are growing concerns about the potential for racial bias in both the predictive accuracy and utilization of these instruments in correctional practice. Concerns have been raised by notable figures such as U.S. Attorney General Eric Holder, who has argued that by focusing on static factors like socioeconomic status or criminal history, risk assessments may obstruct the administration of justice by biasing minority offenders toward higher risk classifications (Holder, 2014; Starr, 2014). Although race is not measured overtly in modern risk assessments, risk factors may serve as proxies for race in risk assessments (Harcourt, 2015) due to systematic racial disparities in wealth accumulation (Proctor, Semega, & Kollar, 2016; U.S. Census Bureau, 2014; Vornovitsky, Gottschalck, & Smith, 2011), homeownership (Callis & Kresin, 2016), and educational attainment (U.S. Census Bureau, 2016). Furthermore, racial minorities are more likely to be stopped by police (Gelman, Fagan, & Kiss, 2007) and may receive harsher criminal sentences (Sweeney & Haney, 1992) relative to their nonminority counterparts, contributing to lengthier and more substantial criminal histories. The inclusion of these factors in risk assessment suggests that racial minorities will receive higher risk scores and classifications relative to their risk of reoffending, resulting in differences in the predictive accuracy of risk assessments as a function of race.

Consequently, a growing number of research studies have investigated racial differences in the predictive accuracy of risk assessments in adults (e.g., Lowenkamp, Holsinger, & Cohen, 2015; Varela, Boccaccini, Murrie, Caperton, & Gonzalez, 2013). To date, no studies have investigated racial differences in the predictive validity of START assessments. However, several studies have investigated racial differences in the predictive validity of LSI-R assessments in offenders. For example, one study found that LSI-R assessments produced weaker predictive validity estimates for institutional misconduct risk among non-White prison inmates compared with White inmates (Chenane et al., 2015). In another study, LSI-R assessments of rearrest and parole revocation risk produced weaker predictive validity for Black prisoners on release relative to White prisoners (Ostermann & Salerno, 2016). In a final study, LSI-R assessments produced overclassification errors for Black offenders when a lower cutoff score for risk classification was employed (Whiteacre, 2006). Together, these findings suggest the potential for racial bias in the ability for a risk assessment to accurately predict likelihood of reoffending. However, no study to date has examined racial bias in the predictive validity of START assessments, and to our knowledge, no study to date has compared the potential for racial bias in predictive validity between structured professional judgment and actuarial risk assessment instruments.

The Present Study

There are limited data on the reliability and validity of risk assessments, and START and LSI-R assessments specifically, in mental health jail diversion programs. Furthermore, despite the national dialogue on race and risk assessment, relatively few studies have examined the predictive validity of risk assessments as a function of client race, in this setting or otherwise (Holder, 2014; Skeem & Lowenkamp,

2015). To those ends, the present study examined the reliability and validity of START and LSI-R assessments completed on mental health jail diversion program clients, piloting a START general offending risk estimate for the first time. Our study aims were to investigate: (a) interrater reliability of START and LSI-R assessments; (b) convergent validity between START and LSI-R assessments; (c) validity of START total scores, START general offending risk estimates, LSI-R total scores, and LSI-R final risk classification in predicting recidivism across five follow-up periods; and (d) differences, if any, in START and LSI-R assessments and their predictive validity by client race.

Method

Participants

The sample includes 95 clients (77 male; 18 female) participating in mental health jail diversion programs in a large metropolitan county in the Southern United States. Participants were drawn from postbooking jail diversion programs targeting misdemeanor- and felony-level offenders with mental illnesses. Through the diversion programs, participants received case management services for up to a 1-year period, which included referral to and coordination of community-based treatment and housing services. However, there were no ongoing status hearings for participants; revocation of diversion status occurred only in the case of a new offense. At baseline, participants were an average age of 36.05 (SD = 12.46) years. The sample comprised a mix of felony (47.4%) and misdemeanor (52.6%) offenders. Just under half (45.3%) were African American, while slightly more than half (53.7%) were Caucasian; one client identified as Asian. More than half of clients (54.7%) identified their ethnicity as Hispanic or Latino and less than half identified as non-Hispanic or Latino (44.2%); data were unavailable for one client. Participants had extensive criminal histories, with an average of 6.69 (SD = 5.42, range = 1-31) jail bookings in the 3 years prior to baseline and 16.07 (SD = 14.99, range = 3-76) lifetime bookings. More than half of participants (53.7%) reported graduating from high school or receiving a GED. Consistent with study inclusion criteria, primary diagnoses at baseline included schizophrenia (35.8%), psychosis NOS (25.3%), and bipolar disorder (20.0%). Nearly one third of the sample (30.5%) had a co-occurring substance use diagnosis at baseline. Diagnostic information was pulled from clinical records maintained by the jail diversion program.

Procedures

The institutional review board of the university approved and monitored this study for human participant protections. Participants provided consent to participate in this study prior to each interview. As part of this process, interviewers assessed a client's comprehension of the study purpose and procedures and also queried external pressures (i.e., coercive influences) to participate. Participants who were judged by the interviewer to have poor comprehension of the study purpose and procedures or who experienced undue external pressures to participate would not be enrolled in the study nor would they be interviewed at follow-up.

Data Collection. Semistructured interviews were conducted by one of four research assistants with clients at baseline and at five follow-up periods (3-month, 6-month, 9-month, 12-month, and 18-month). At baseline, research assistants completed the START and LSI-R assessments based on participant self-report, observation, and review of official records. At each follow-up, participants self-reported arrests and days incarcerated in the previous 3 months. These self-reported recidivism data were used to supplement official county-wide arrest and incarceration records, which were queried by the diversion program coordinator in the county's jail records database and provided in raw form to the research team.

Interrater Reliability. Research assistants with bachelor's degrees attended a 2-day workshop prior to the study start that included training on the START and LSI-R. The workshops and subsequent training included completion of practice cases (n = 5) on which the research assistants were required to achieve adequate agreement with the trainers. Interrater reliability was conducted on a subset of 25% of the total sample (n = 24). Interrater reliability procedures were identical to the interview procedures. In all instances, an interview was conducted with a client, and both research assistants attended the interview and coded the assessment independently.

Measures

START. The START is a structured professional judgment guide for assessing risk of seven adverse outcomes: violence to others, suicide, self-harm, victimization, substance use, unauthorized leave, and self-neglect (Webster et al., 2009). In the present study, the START additionally was used to inform an eighth estimate of risk for general offending (Desmarais, Van Dorn, et al., 2012). For each of the 20 items, strength and vulnerability demonstrated in the past 2 to 3 months are coded on a 3-point ordinal scale, from 0 (*minimally present*) to 2 (*maximally present*). Strength and vulnerability are rated independent of one another; thus, an individual may receive the same high (or low) score for both strength *and* vulnerability on any given item. Additionally, assessors may indicate key and critical items to highlight those items of particular importance to the

individual. Assessors subsequently estimate risk extending to the next 3 months as *low, moderate*, or *high* for each of the eight outcomes. These specific risk estimates are formed through structured professional judgment which considers strength and vulnerability ratings, the presence of key and/or critical items, physical health problems, and other historical factors (e.g., past violence or victimization). Strength and vulnerability total scores were calculated by summing the item ratings, with a possible range of 0 to 40. When there were five or fewer missing item ratings, we prorated total scores by summing the item ratings, dividing by the total possible scale score (i.e., 40), multiplying that amount by the number of omitted items, and adding the obtained value to the original total (see Webster et al., 2009). Assessments missing more than five-item ratings were not prorated, leading to the exclusion of two START assessments from our analytic sample. Levels of internal consistency were good for START strength total scores (Cronbach's α = .87) and START vulnerability scores (Cronbach's α = .81).

LSI-R. The LSI-R is an actuarial instrument designed to assess risk of criminal offending and technical violations in adult offenders (Andrews & Bonta, 2001). We selected this version of the Level of Service assessments due to its widespread use in research (Desmarais et al., 2016), which allowed us to investigate the predictive validity of assessments in relation to the extant literature. Additionally, because we selected this instrument for use in a research study, the case management and specific needs and responsivity components of the LS/CMI (Andrews et al., 2004) and LS/RNR (Andrews et al., 2008) assessments were not relevant to our investigation.

The LSI-R includes 54 items representing both static and dynamic risk factors for recidivism. Out of the 54 items, 13 items measure the severity of specific risk factors and are coded from 0 indicating *a very unsatisfactory situation with a very clear and strong need for improvement* (most severe) to 3 indicating *a very satisfactory situation with no need for improvement* (least severe). For these items, scores were recoded as 0 or 1 to indicate either a generally satisfactory situation (i.e., absence of the risk factor, 0) or a generally unsatisfactory situation (i.e., presence of the risk factor, 1). All other items measure the absence or presence of a risk factor and are coded as 0 or 1. Scores from the 54 items were totaled to create the LSI-R total score. Predetermined cutoffs were used to classify offenders into risk categories: Low = 0 to 13; Low-Moderate = 14 to 23; Moderate = 24 to 33; Moderate-High = 34 to 40; High = 41 to 54. Due to low cell counts, these five risk bins were collapsed into three categories following the procedure used in prior research (Singh et al., 2011); specifically, Low and Low-Moderate were coded as *Low*, Moderate was coded as *Moderate*, and Moderate-High and High were coded as *High*. This strategy also afforded better comparison with the three categories of the START General Offending risk estimate. For all predictive validity analyses involving risk estimates, *High* risk was used as the reference group. Insufficient data were available to code an LSI-R assessment for one participant. Levels of internal consistency were acceptable for LSI-R total scores (Cronbach's α = .72).

Race. Race was coded categorically and measured independently from ethnicity. All participants except one identified as either Caucasian or African American. As a result, one participant identifying as Asian was excluded from race analyses.

Recidivism. Recidivism measures were coded from raw booking data provided by jail diversion staff and included count measures of number of arrests and days incarcerated at each follow-up period. Although days incarcerated is not frequently employed as a measure of general recidivism in risk assessment validation research, this measure in meaningful in the context of jail diversion for several reasons. Primarily, jail diversion programs were created to decrease time incarcerated among justice-involved adults with mental illnesses because jails provide limited access to mental health care (Wilper et al., 2009) and may exacerbate symptoms of mental illness (Steadman et al., 1994; Torrey, Stieber, & Ezekiel, 1998). Accordingly, days incarcerated is a common and accepted indicator of recidivism in jail diversion research (Case et al., 2009; Sirotich, 2009; Steadman & Naples, 2005; Steadman, Redlich, Callahan, Robbins, & Vesselinov, 2011). Additionally, days incarcerated may be an appropriate measure of reoffending severity for groups that are at particularly high risk for reincarceration because it captures a broader range of severity in reoffending (Urban Institute, 2016). As detailed earlier, our study involves a sample of justice-involved adults with mental illnesses with substantial criminal histories (i.e., an average of 16.07 lifetime jail bookings).

To code recidivism variables, each unique booking date was treated as a separate arrest and counts of jail days were tabulated based on date of booking and date of release from jail. Additionally, participants self-reported number of arrests and days incarcerated during the follow-up periods. Because attrition rates were relatively high for self-reported data (i.e., 30% at 3-month follow-up to 81.0% at 18-month follow-up), official records were used as the primary source of recidivism data. However, when available, we incorporated self-reported data from follow-up interviews by recording client self-report estimates when a client overestimated recidivism in comparison with official records. Across follow-up periods, agreement between self-reported recidivism and official records was high for both arrests (86.8%, range: 83.3% to 91.2%) and days incarcerated (85.7%, range: 81.2% to 88.1%). Official recidivism data were unavailable for one participant who had START and LSI-R assessments.

Analyses

First, descriptive statistics were calculated on all study variables. Second, interrater reliability of START Strength and Vulnerability total scores and LSI-R total scores were evaluated using two-way mixed single effects intraclass correlation coefficients ($ICC_{3,1}$) which estimates reliability of a single rating using a fixed effects model (McGraw & Wong, 1996; Shrout & Fleiss, 1979). For intraclass correlation coefficients, values less than .40 indicate slight, between .40 and .59 fair, between .60 and .74 good, and greater than .75 excellent agreement (Cicchetti et al., 2006). Third, convergent validity was assessed via bivariate correlations between LSI-R and START total scores and via Cohen's kappa percentage agreement between LSI-R and START risk estimates. Based on Cohen's criteria, correlation values of .10 indicate a small, .30 a medium, and .50 a large effect size (Cohen, 1988). Kappa values of less than .20 reveal slight, between .21 and .40 fair, between .41 and .60 moderate, between .61 and .80 substantial, and greater than .81 almost perfect agreement (Landis & Koch, 1977). Fourth, a series of bivariate negative binomial regression analyses were conducted between total scores and recidivism at each follow-up. Because criminal justice records represent count data, generalized linear approaches are necessary to account for positive skew (i.e., excess of zeros) and absence of a normal distribution in the dependent variables (Walters, 2007). Data in this study failed to meet the variance assumptions of Poisson distributions (i.e., the variance is equivalent to the mean), thus, resulting in the negative binomial regression approach. In negative binomial models, the Wald chi-square statistic provides a significance test of the hypothesis that the regression coefficient is different from zero, controlling for other model terms. The incidence rate ratio (IRR) provides a measure of effect size and represents a factor of the dependent variable associated with a one-unit increase in the independent variable. Additional negative binomial regression analyses were conducted using dummy-coded risk estimates (with high risk as the reference group) and recidivism measures. Fifth and finally, hierarchical negative binomial models were conducted to test for interaction effects of race by total scores and risk estimates on recidivism at 12-month follow-up. In Block 1, the total score or risk estimate under investigation was entered together with race. In Block 2, a risk by race effect was added. Improvement in model fit in Block 2 was determined by change in −2 log(likelihood) statistics, which were evaluated by computing a p value for a chi-square distribution with 1 to 2 degrees of freedom.

Results

Descriptive Statistics

Total Scores. The mean START vulnerability total scores was 19.13 (SD = 6.30, range = 2-30), and the mean START strength total scores was 14.78 (SD = 7.31, range = 0-29), suggesting that participants presented with greater vulnerabilities than strengths at baseline. Mean LSI-R total scores were 28.08 (SD = 6.08, range = 15-41), suggesting that participants presented with a moderate amount of recidivism risk at baseline. Between-group comparisons showed no significant differences between African American and Caucasian participants on START strength total scores, $t(90) = 1.08$, $p = .282$, START vulnerability total scores, $t(90) = -0.85$, $p = .397$, or LSI-R total scores, $t(91) = -0.57$, $p = .571$.[1]

Table 1. Descriptive Statistics for Arrests and Jail Days for All Follow-Up Periods.

	Arrests			Jail days		
Follow-up period	*M*	*SD*	Range	*M*	*SD*	Range
3-Month	0.52	0.94	0-4	5.80	13.15	0-60
6-Month	0.82	1.31	0-7	10.51	23.56	0-153
9-Month	1.12	1.56	0-8	15.66	33.48	0-231
12-Month	1.38	1.77	0-8	19.78	37.78	0-245
18-Month	1.89	2.48	0-15	29.02	52.61	0-259

Note. $N = 94$.

Risk Estimates. Using START general offending risk estimates, less than a quarter of participants were rated as low risk (24.7%, $n = 23$), 53.8% as moderate risk (53.8%, $n = 50$), and 21.5% at high risk ($n = 20$) for recidivism. Using the three LSI-R risk bins, less than a quarter of participants were categorized at low risk (20.2%, $n = 19$), 61.7% as moderate risk ($n = 58$), and 18.1% as high risk ($n = 17$). Similar to total scores, between-group comparisons showed no significant differences between African American and Caucasian participants on START general offending risk estimates, $\chi^2(2) = 1.53$, $p = .465$, or LSI-R risk estimates, $\chi^2(2) = 1.13$, $p = .569$.

Recidivism. Full descriptive statistics for recidivism across all follow-up periods are presented in Table 1. Overall, the average number of arrests for the sample was generally low: less than two arrests per person over the 18-month follow-up period. However, there was considerable variability across participants; for example, the number of arrests ranged from 0 to 15. Findings were similar for the number of jail days across follow-up periods (see Table 1). Furthermore, each arrest was associated with an average of 15.35 jail days.

Reliability

To address Aim 1, interrater reliability was calculated on a subset of 24 participants, using interclass correlation coefficients (ICC) for START strength and vulnerability total scores and LSI-R total scores, and Cohen's kappa (and percentage agreement) for START and LSI-R risk estimates.

Table 2. Distribution of and Concordance Between LSI-R and START Risk Estimates.

	START					
	Low		Moderate		High	
LSI-R	*n*	%	*n*	%	*n*	%
Low	3	3.2	10	10.8	6	6.5
Moderate	19	20.4	31	33.3	7	7.5
High	1	1.1	9	9.7	7	7.5

Note. $N = 93$. START = Short-Term Assessment of Risk and Treatability; LSI-R = Level of Service Inventory–Revised. $\kappa = .04$ ($SE = .08$); $p = .623$.

Interrater reliability was good for START strength ($ICC_{3,1} = .61$, 95% confidence interval [CI .29, .81], $p = .001$) and vulnerability total scores ($ICC_{3,1} = .63$, 95% CI [.31, .82], $p < .001$), and was excellent for LSI-R total scores ($ICC_{3,1} = .76$, 95% CI [.53, .89], $p < .001$; Cicchetti et al., 2006). Interrater reliability was lower for START general offending risk estimates (58.3% agreement, $p = .209$) and adequate for LSI-R risk bins (66.7%, $p = .017$; Landis & Koch, 1977).

Validity

Convergent Validity. To address Aim 2, we investigated the convergent validity between LSI-R and START assessments. Results indicated a significant, positive association between START vulnerability and LSI-R total scores, $r(91) = 0.37$, $p < .001$, providing evidence for convergent validity. Similarly, START strength scores were negatively and significantly associated with LSI-R total scores, $r(91) = -0.37$, $p < .001$. Cross-tabulation of START risk estimates and LSI-R risk bins are presented in Table 2. Results showed weak agreement between START general offending risk estimates and LSI-R risk bins, $\kappa = .04$ ($SE = .08$), $p = .623$.

Predictive Validity. To address Aim 3, we investigated the predictive validity of START and LSI-R assessments using negative binomial regression analyses, the results of which are provided in Table 3. START strength total scores significantly and inversely predicted number of arrests (IRR range: 0.88-0.95) and jail days (IRR range: 0.83-0.96) across all follow-up periods. The strongest effect was observed for the prediction of jail days at 6-month follow-up (IRR = 0.83); specifically, each 1-point increase in START strength scores was associated with 1.20 times fewer jail days. START vulnerability total scores significantly and positively predicted arrests (IRR range: 1.04-1.06) and jail days (IRR range: 1.03-1.07) over follow-up periods of 6 months and longer for both. START vulnerability total scores were most robust predictors of jail days at 6- and 9-month follow-up (IRRs = 1.07). Each 1-point increase in vulnerability scores was associated with 1.07 times greater jail days served. LSI-R total scores demonstrated predictive validity over most follow-up periods, with the exception of the 18-month follow-up (see Table 3). LSI-R total scores were more robust predictors of jail days (IRR range: 1.06-1.13) than arrests (IRR range: 1.05-1.06; see columns 3 and 6 in Table 3). LSI-R total scores produced the strongest predictive validity estimates for jail days at 6-month follow-up (IRR = 1.13); each 1-point increase in LSI-R total scores was associated with 1.13 times greater jail days served.

Table 4 presents the results of analyses testing the predictive validity of START general offending risk estimates and LSI-R risk bins. Overall, START risk estimates showed relatively limited validity in predicting number of arrests across follow-up periods ($ps \geq .025$). In contrast, START risk estimates significantly predicted number of jail days from 3-month to 12-month follow-up periods. Estimates of low versus high risk showed greater discrimination (IRR range: 0.11-0.19) than found between estimates of moderate versus high risk (IRR range: 0.48-0.52), as would be expected. LSI-R risk bins similarly showed greater validity in predicting jail days than arrests, though effect sizes were generally stronger relative to START risk estimates. With respect to arrests, we found discrimination between participants in the low versus high risk bins (IRR range: 0.24-0.33), but not the moderate versus high risk bins ($ps \geq .120$). For jail days, however, there was discrimination between both low versus high risk (IRR range: 0.07-0.31) and moderate versus high risk (IRR range: 0.39-0.48) participants across most follow-up periods (see Table 4). Effect sizes were strongest at the 6-month period with respect to jail days; participants rated at high risk had 2.56 and 14.29 times more jail days relative to those rated at moderate and high risk, respectively.

To address Aim 4, we examined whether client race moderated the predictive validity of LSI-R and START assessments. Although we conducted these analyses across all five follow-up periods, we found the most significant assessment by race interaction effects at 3-month follow-up (see Table 5). In contrast, there were few significant interaction effects at the 6-, 9-, and 12-month follow-up periods and none at the 18-month follow-up periods. Of note, race did not moderate predictive validity of START strength total scores across any follow-up period or recidivism measure ($ps > .071$). (Full results not presented, but are available on request.)

Overall, there were no significant moderating effects of race on validity in predicting arrests at 3-month follow-up for any of the assessment measures ($ps \geq .186$). However, we did find some moderating effects of race on validity in predicting jail days at 3-month follow-up. Specifically, there was a significant START vulnerability total scores by race interaction ($p < .001$), such that a 1-point gain in vulnerability total scores was associated with 1.18 times more

Table 3. Validity of LSI-R and START Total Scores in Predicting Arrests and Jail Days Across Follow-Up Periods.

Follow-up period by predictor	Arrests			Jail days		
	Wald χ^2	IRR	95% CI	Wald χ^2	IRR	95% CI
START vulnerability total score						
3-Month	2.38	1.05	[0.99, 1.12]	1.05	1.02	[0.98, 1.05]
6-Month	3.38†	1.05	[1.00, 1.11]	14.26***	1.07	[1.03, 1.10]
9-Month	4.78*	1.06	[1.01, 1.11]	17.05***	1.07	[1.04, 1.11]
12-Month	5.42*	1.06	[1.01, 1.11]	6.98**	1.04	[1.01, 1.07]
18-Month	4.17*	1.04	[1.00, 1.09]	4.17*	1.03	[1.00, 1.06]
START strength total score						
3-Month	16.00***	0.88	[0.82, 0.93]	58.59***	0.84	[0.80, 0.88]
6-Month	17.27***	0.90	[0.85, 0.94]	68.75***	0.83	[0.80, 0.87]
9-Month	13.74***	0.92	[0.88, 0.96]	21.27***	0.92	[0.89, 0.95]
12-Month	12.96***	0.93	[0.89, 0.97]	6.72*	0.96	[0.93, 0.99]
18-Month	7.30**	0.95	[0.91, 0.99]	4.56*	0.96	[0.93, 1.00]
LSI-R total score						
3-Month	4.81*	1.07	[1.01, 1.14]	24.32***	1.11	[1.07, 1.16]
6-Month	5.13*	1.06	[1.01, 1.12]	31.25***	1.13	[1.08, 1.18]
9-Month	3.75†	1.05	[1.00, 1.10]	20.49***	1.10	[1.05, 1.14]
12-Month	4.76*	1.05	[1.00, 1.10]	8.38**	1.06	[1.02, 1.10]
18-Month	0.27	1.01	[0.97, 1.05]	1.30	1.02	[0.98, 1.06]

Note. N = 92 START assessments, N = 93 LSI-R assessments. START = Short-Term Assessment of Risk and Treatability; LSI-R = Level of Service Inventory–Revised; IRR = incidence rate ratio; CI = confidence interval for IRR. Validity estimates produced in separate negative binomial regression models.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 4. Validity of LSI-R and START Risk Estimates in Predicting Arrests and Jail Days Across Follow-Up Periods.

Follow-up period by predictor	Arrests						Jail days					
	Low vs. high			Moderate vs. high			Low vs. high			Moderate vs. high		
	Wald χ^2	IRR	95% CI	Wald χ^2	IRR	95% CI	Wald χ^2	IRR	95% CI	Wald χ^2	IRR	95% CI
START risk estimate												
3-Month	2.94†	0.37	[0.12, 1.15]	0.32	0.79	[0.34, 1.80]	27.58***	0.15	[0.07, 0.30]	2.48	0.64	[0.37, 1.11]
6-Month	2.85	0.43	[0.16, 1.14]	0.12	0.88	[0.42, 1.84]	39.40***	0.12	[0.06, 0.23]	7.13**	0.48	[0.28, 0.82]
9-Month	2.53	0.49	[0.20, 1.18]	<0.01	1.00	[0.50, 2.00]	45.18***	0.11	[0.06, 0.21]	5.92*	0.52	[0.30, 0.88]
12-Month	4.99*	0.39	[0.17, 0.89]	0.81	0.74	[0.39, 1.42]	26.89***	0.19	[0.10, 0.36]	5.86*	0.52	[0.31, 0.88]
18-Month	1.32	0.62	[0.28, 1.39]	1.09	0.71	[0.37, 1.35]	3.03†	0.55	[0.28, 1.08]	1.31	0.72	[0.42, 1.26]
LSI-R risk estimate												
3-Month	4.80*	0.24	[0.07, 0.86]	1.92	0.56	[0.24, 1.27]	27.68***	0.13	[0.06, 0.27]	6.37*	0.48	[0.27, 0.85]
6-Month	4.85*	0.30	[0.10, 0.88]	1.24	0.65	[0.31, 1.38]	47.93***	0.07	[0.03, 0.15]	10.99**	0.39	[0.22, 0.68]
9-Month	5.27*	0.32	[0.12, 0.85]	1.15	0.68	[0.34, 1.37]	38.32***	0.11	[0.05, 0.22]	9.32**	0.42	[0.24, 0.73]
12-Month	5.89*	0.33	[0.13, 0.81]	2.42	0.59	[0.30, 1.15]	13.57***	0.28	[0.14, 0.55]	6.98**	0.48	[0.27, 0.83]
18-Month	2.78†	0.52	[0.25, 1.12]	0.45	0.81	[0.43, 1.51]	14.20***	0.31	[0.17, 0.57]	2.28	0.67	[0.39, 1.13]

Note. N = 92 START assessments, N = 93 LSI-R assessments. START = Short-Term Assessment of Risk and Treatability; LSI-R = Level of Service Inventory–Revised; IRR = incidence rate ratio; CI = confidence interval for IRR. For risk estimates, High = 0. Validity estimates produced in separate negative binomial regression models.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

jail days among Caucasian participants relative to African American participants. LSI-R total scores similarly were moderated by race (p = .040), such that 1-point increase in LSI-R total scores was associated with 1.10 times more jail days among Caucasian participants relative to African American participants. Results also showed significant effects of START general offending and LSI-R risk estimates by race on jail days at 3-month follow-up. For the START risk estimate, Caucasian participants were incarcerated for more days relative to African American participants

Table 5. Moderation of Total Scores and Risk Estimates by Race on Recidivism at 3-Month Follow-Up.

	Arrests at 3-month				Jail days at 3-month			
Interaction term	Wald χ^2	IRR	95% CI	Δ − 2LL	Wald χ^2	IRR	95% CI	Δ − 2LL
START strength × race	0.24	0.97	[0.85, 1.10]	0.24	0.03	0.99	[0.91, 1.08]	0.03
START vulnerability × race	1.75	1.09	[0.96, 1.24]	1.80	17.76***	1.18	[1.09, 1.27]	18.57***
LSI-R total × race	0.18	1.03	[0.91, 1.16]	0.18	4.83*	1.10	[1.01, 1.20]	5.10*
START general offending (low) × race	1.23	4.61	[0.31, 68.61]	2.65	4.21*	6.09	[1.08, 34.20]	16.62***
START general offending (moderate) × race	0.17	0.70	[0.13, 3.84]		3.82†	0.33	[0.11, 1.00]	
LSI-R risk estimate (low) × race	0.16	1.78	[0.11, 29.69]	0.72	1.11	0.43	[0.09, 2.07]	5.72†
LSI-R risk estimate (moderate) × race	0.24	0.66	[0.12 3.49]		5.68*	0.25	[0.08, 0.78]	

Note. N = 91 to 92. START = Short-Term Assessment of Risk and Treatability; LSI-R = Level of Service Inventory–Revised; IRR = incidence rate ratio; CI = confidence interval for IRR. For risk estimates, High = 0. Interaction estimates produced in separate negative binomial regression models.
Δ − 2LL reflects improvement in model fit on addition of the interaction term(s) in Block 2. Block 1, not shown here, included main effects of race and risk assessment measures.
†p < .10. *p < .05. **p < .01. ***p < .001.

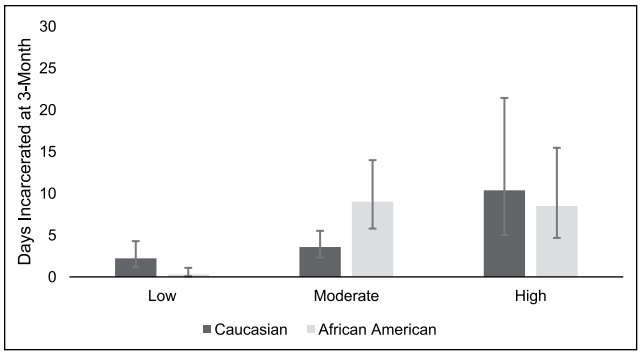

Figure 1. Jail days at 3-month follow-up by START general offending risk estimate and race.
Note. START = Short-Term Assessment of Risk and Treatability. Error bars reflect 95% confidence intervals for estimated marginal means.

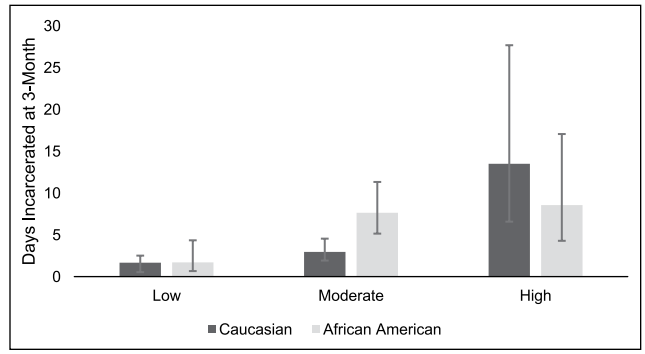

Figure 2. Jail days at 3-month follow-up by LSI-R risk estimate and race.
Note. LSI-R = Level of Service Inventory–Revised. Error bars reflect 95% confidence intervals for estimated marginal means.

in both the low and high risk categories; however, this effect was greater among participants classified at low risk (see Figure 1). For the LSI-R risk estimate, at the moderate risk level, African American participants had more jail days relative to Caucasian participants; however, at the high risk level, Caucasian participants had more jail days relative to African American participants (see Figure 2). This interaction was also observed for START general offending risk estimates, but only trended toward significance (p = .051).

Discussion

Structured risk assessments are increasingly used in mental health jail diversion programs across the United States and internationally to facilitate community-based treatment and case management of justice-involved adults with mental illnesses. However, few studies have evaluated the reliability and validity of such instruments used with mental health jail diversion clients. Thus, the goal of the present study was to examine the START and LSI-R with regard to interrater reliability, convergent validity, and predictive validity of their total scores and risk estimates over multiple follow-up periods. We additionally explored differences in START and LSI-R assessments, as well as their predictive validity, as a function of client race. Overall, findings support the use of these instruments to inform risk assessment and management of mental health jail diversion clients and provide limited evidence of racial bias. Below, we discuss the study findings in further detail.

Integration of Findings

Our first aim was to investigate the interrater reliability of START and LSI-R assessments completed on mental health jail diversion clients. Results showed good to excellent levels of interrater reliability for START vulnerability, START strength, and LSI-R total scores. To our knowledge, only one published study conducted in a correctional sample has investigated the interrater reliability of LSI-R total scores, reporting slightly lower levels of interrater reliability compared with the

present findings (Rocque & Plummer-Beale, 2014). For START assessments, these findings are consistent with levels of interrater reliability reported across seven studies with institutionalized patients (O'Shea & Dickens, 2014), but provide the first evidence supporting interrater reliability of the START strength and vulnerability total scores completed in a community corrections setting. Levels of interrater reliability were adequate for LSI-R risk estimates, but lower for START general offending risk estimates. These findings suggest the need for more training to consensus on the START general offending risk estimate prior to implementation in other criminal justice settings and echo clinicians' reports that the final risk estimates are one of the most challenging aspects of the structured professional judgment approach, and START assessments, specifically (Desmarais, 2009; Doyle, Lewis, & Brisbane, 2008). These findings also likely reflect the pilot nature of the START general offending risk estimate and the need for more detailed coding instructions.

A recent meta-analytic review of validation research conducted in the United States identified only two studies that reported on the interrater reliability of risk assessments completed in correctional settings (Desmarais et al., 2016). Though the current results are promising, further research is needed to establish the reliability of START and LSI-R assessments, as well as risk assessments more generally, in mental health jail diversion and other U.S. correctional settings.

Our second aim was to examine convergent validity between START and LSI-R assessments. Results showed strong associations between START total scores and LSI-R scores, suggesting similar constructions and operation of factors associated with recidivism risk. In contrast, there was limited convergent validity between START general offending and LSI-R risk estimates as demonstrated by poor agreement between ratings of low, moderate, and high risk. These discrepancies may reflect, at least in part, inherent differences in the assessment time frames of these two instruments. Specifically, the START risk estimates are designed to forecast outcomes over a 90-day period (Webster et al., 2009), whereas the LSI-R risk estimates forecast risk over several months to years (Fass, Heilbrun, DeMatteo, & Fretz, 2008; Kelly & Welsh, 2008; Manchak, Skeem, & Douglas, 2008). Indeed, in the present study, the LSI-R assessments classified more participants as moderate risk relative to the START assessments, whereas START assessments classified more participants as low risk than did LSI-R assessments.

Taken together, convergent validity findings suggest there is moderate overlap between the instruments, even though the START was not originally designed to estimate risk for general offending. Moreover, these findings support the START's comprehensive approach to risk assessment and management that focuses on shared predictors of multiple adverse outcomes (Webster, Nicholls, Martin, Desmarais, & Brink, 2006). Findings are also consistent with recent assertions that risk assessment instruments are more or less comparable to one another and, thus, that tool selection should be informed by other issues, such as the purpose of the evaluation (i.e., classification, treatment planning, etc.), the population of interest, and resources available (Desmarais et al., 2016; Skeem & Monahan, 2011).

Our third aim was to evaluate the validity of START and LSI-R assessments in predicting general offending across multiple time frames. Overall, findings support for the predictive validity of START and LSI-R assessments across both shorter and longer follow-up periods. Interestingly, although START is designed to predict outcomes within 3 months of assessment, START vulnerability total scores predicted recidivism across all follow-up periods, *except* 3 months. In previous studies, START assessments have demonstrated validity in predicting institutional aggression in psychiatric inpatients over 1-month to 12-month follow-up periods (Chu et al., 2011b; Desmarais, Nicholls, et al., 2012). Because this is the first study to investigate the validity of START assessments in predicting general offending, further research is needed to explore the time frame over which START vulnerability total scores predict recidivism in general and in mental health diversion clients specifically.

Similarly, the predictive validity of the LSI-R is typically tested with respect to longer follow-up periods (i.e., 12 or more months; Fass et al., 2008; Kelly & Welsh, 2008; Manchak et al., 2008); however, LSI-R total scores were most predictive at 3- and 6-month follow-ups. In fact, LSI-R total scores did not significantly predict recidivism over the 18-month follow-up, though LSI-R total scores have predicted recidivism as many as 3.79 years following assessment in other research (Vose, Smith, & Cullen, 2013). Discrepancies between prior and current findings may be attributable to the unique nature of the current sample (i.e., mental health jail diversion clients), but also may reflect measurement and analytic issues. In particular, we operationalized recidivism as number of arrests and jail days, and not reconviction (e.g., Manchak et al., 2008). Additionally, we used negative binomial regression models to analyze count recidivism data, whereas prior studies typically have relied on analytic strategies, such as receiver operating curve analyses, that require researchers to dichotomize recidivism measures (Fass et al., 2008; Manchak et al., 2008). Our strategy is appropriate given the skewed distribution of recidivism data (Walters, 2007) and allowed us to more accurately account for the frequency of recidivism across participants. Nonetheless, as with the START vulnerability total scores, findings suggest the need for more research examining the time frame during which LSI-R total scores predict recidivism in mental health jail diversion clients.

START strength total scores emerged as the most consistent and robust predictor of arrests and jail days across follow-up periods. These results are consistent with prior research demonstrating the validity of START strength total scores in predicting institutional aggression in forensic psychiatric patients (Desmarais, Nicholls, et al., 2012) and suggest that protective factors may be especially relevant to the assessment and management of recidivism risk among community-based, justice-involved adults with mental illnesses. Consideration of protective factors may help differentiate among individuals in this population who present with many similar risk factors (e.g., chronic homelessness, substance abuse problems, mental health problems) and may guide risk management strategies by acting as the foundation or cornerstone for treatment (de Ruiter & Nicholls, 2011). Indeed, consideration of protective factors have demonstrated validity in predicting adverse outcomes, such as violence, above and beyond risk factors (de Vries Robbé, de Vogel, & Douglas, 2013; Desmarais, Nicholls, et al., 2012; Lodewijks, de Ruiter, & Doreleijers, 2010). However, questions remain regarding how risk and protective factors are related both to each other and to risk for adverse outcomes.

Although the need for a START general offending risk estimate has been articulated in practice and in previous research (Desmarais, Van Dorn, et al., 2012), this is the first study to provide evidence supporting its predictive validity. Specifically, START general offending risk estimates demonstrated validity in predicting jail days across all follow-up periods, but were somewhat less accurate in predicting arrests. LSI-R risk estimates similarly demonstrated greater validity in predicting jail days compared with arrests, but nonetheless were able to discriminate between participants at low compared with high risk of arrests across all but the 18-month follow-up period. Overall, both START and LSI-R risk estimates produced more robust effect sizes when significant compared with their total scores, suggesting they were more discriminating than 1-point gains in total scores, which is consistent with prior research demonstrating the superiority of risk estimates over total scores (Desmarais et al., 2016; Desmarais, Nicholls, et al., 2012; Douglas, Ogloff, & Hart, 2003; Douglas, Yeomans, & Boer, 2005). Though findings of the current study are encouraging, further research is needed establish the validity of a START general offending risk estimate.

Finally, our fourth aim was to explore differences in LSI-R and START assessments, as well as their predictive validity, as a function of client race. Our findings showed no differences between Caucasian and African American mental health jail diversion clients in the total scores or risk estimates for either the LSI-R or START. However, we did find some evidence of differences in the predictive validity of START and LSI-R assessments for jail days, mainly over a shorter follow-up period (i.e., 3-month). Specifically, both LSI-R and START general offending risk estimates underclassified African American clients for moderate risk and overclassified African American clients for high risk relative to their time incarcerated over a 3-month follow-up period. Similarly, both START vulnerability total scores and LSI-R total scores demonstrated weaker predictive validity for African American clients relative to Caucasian clients. These findings are consistent with previous studies that similarly found LSI-R assessments overclassified African American offenders in a community corrections sample (Whiteacre, 2006), and demonstrated racial biases in the prediction of institutional misconduct (Chenane et al., 2015) and rearrest or parole revocation (Ostermann & Salerno, 2016) in prison inmates. In contrast, we found no moderating effects of race on the validity of START and LSI-R assessments in predicting arrests. Additionally, there was no evidence of racial bias in the predictive validity of START strength total scores across any follow-up period or recidivism measure. Because this is the first study to explore racial differences in the predictive validity of START assessments; future research is needed to replicate these findings.

Although we investigated racial bias across all follow-up periods, evidence of disparity in the predictive accuracy of START and LSI-R assessments emerged primarily during 3-month follow-up period. This may reflect that participants were recently enrolled diversion program clients who were referred to community-based treatment and supportive services. As a result, risk and need factors contributing to higher risk classifications among African American participants on admission to the diversion program may have been effectively targeted throughout their participation, resulting in lower risk ratings. However, this is only speculation, and whether diversion programs can effectively target risks and needs has received little attention in the existing literature (Campbell et al., 2015) and is an important direction for future research (Skeem et al., 2014).

Together, these findings add to the national dialogue on whether application of risk assessment instruments may result in discriminatory practices and contribute to harsher sentences for offenders of minority racial status (Harcourt, 2015; Holder, 2014; Starr, 2014). The lack of differences in the assessment scores and risk estimates, as well as in their validity in predicting arrests, provides little evidence of racial bias in the instruments themselves. Whether application of risk assessment instruments in sentencing and corrections results in racially biased decisions regarding level or length of supervision remains to be seen. There is evidence to suggest that White individuals are charged with lesser crimes and receive shorter sentences than non-White individuals, even though they are arrested as frequently and, potentially, for the same behavior(s) (Mauer, 2006; Pettit & Western, 2004). However, our findings suggest that protective factors are less susceptible to racial biases and

more accurate than risk factors in predicting recidivism across groups of racially diverse offenders. Indeed, protective factors by definition do not include variables that have been cited as acting as proxies for race in risk assessment, such as history of criminal behavior (Harcourt, 2015). Thus, the consideration of protective factors during the process of assessing risk for reoffending may help reduce (rather than increase) discriminatory criminal justice practices, including decisions regarding supervision, treatment referral, and case management (Grove & Meehl, 1996).

Limitations and Future Directions

Our findings should be considered in light of limitations to the study design, which may inform directions of future research. First, participants represented a relatively small and unique sample of justice-involved adults with mental illnesses who were participating in mental health jail diversion programs in one U.S. county. Replication in other, larger samples of justice-involved adults with mental illnesses is needed. Second, recidivism data were limited to self-report and county-wide official records; we were unable to access state-wide records. As a result, our records may have underestimated true rates of recidivism; that is, participants may have been arrested and incarcerated outside of the county (or even the state). Relatedly, our recidivism measures were limited to arrests and jail days, both of which have the potential to be imperfect measures of criminal offending. In particular, the amount of time served in jail may be affected by a detainee's financial resources, including ability to post bond and access a private attorney. Third, risk assessments in this study were completed by research assistants trained by the research team. Future research is needed to establish the field reliability and validity of START and LSI-R assessments completed by practitioners in mental health jail diversion programs. Fourth, the current investigation was part of a broader study evaluating the provision of behavioral health treatment services in the context of a successful mental health jail diversion program and, as such, lower predictive validity estimates could reflect true changes in recidivism risk attributable to intervention. Future research should examine associations between changes in START and LSI-R assessment scores attributable to treatment and recidivism risk in mental health jail diversion clients (Labrecque, Smith, Lovins, & Latessa, 2014; Vose et al., 2013). Findings of such a study may be particularly relevant to START with its inclusion of protective factors which may play a role in improving community supervision outcomes for offenders (Woldgabreal, Day, & Ward, 2014).

Conclusions

This study provides the first evidence supporting the reliability and validity of START and LSI-R assessments in predicting recidivism across multiple time frames among mental health jail diversion clients. It also provides some evidence supporting the use of START and LSI-R assessments—and of START strength total scores, in particular—with mental health jail diversion clients of diverse racial backgrounds, though questions remain regarding differences in the validity of START vulnerability and LSI-R total scores and risk estimates in predicting jail days in the short-term. Consistent with the Risk-Need-Responsivity model (Andrews et al., 1990), findings suggest that implementation of the START and LSI-R mental health jail diversion programs should assist in the identification of those at higher and lower risk of recidivism and, consequently, inform the development of case plans and risk management strategies. However, whether mental health jail diversion programs can employ the results of risk assessments using these (or other) instruments to successfully target risk and protective factors, improve rehabilitation, and, ultimately, decrease recidivism remains a critical avenue for future research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Bristol-Myers Squibb Foundation.

Note

1. Due to the limited number of female participants (n = 18), sex differences in the predictive validity of START and LSI-R assessments were not a specific aim of the present study. However, we conducted exploratory analyses examining differences in risk assessment scores and 18-month recidivism measures by sex, which showed little evidence of sex differences in risk assessment scores themselves (ps ≥ .054) or arrests (p = .340). However, jail days differed significantly between male and female participants, $t(91.49) = 3.27$, $p = .002$. Given limited research investigating sex differences in the predictive validity of risk assessments and the growing use of jail days as an outcome of interest in diversion program research, future work in this area is needed.

References

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Andrews, D. A., & Bonta, J. (1998). *Level of Service Inventory–Revised: Screening Version (LSI-R: SV)*. Toronto, Ontario, Canada: Multi-Health Systems.

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

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law*, *16*, 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, *17*, 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *Level of Service/Case Management Inventory (LS/CMI)*. Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2008). The Level of Service/Risk-Need-Responsivity (LS/RNR). Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (1st ed., pp. 199-225). New York, NY: Routledge.

Baillargeon, J., Binswanger, I. A., Penn, J. V., Williams, B. A., & Murray, O. J. (2009). Psychiatric disorders and repeat incarcerations: The revolving prison door. *American Journal of Psychiatry*, *166*, 103-109. doi:10.1176/appi.ajp.2008.08030416

Barber-Rioja, V., Dewey, L., Kopelovich, S., & Kucharski, L. T. (2012). The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Criminal Justice and Behavior*, *39*, 475-492. doi:10.1177/0093854811432609

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

Callis, R. R., & Kresin, M. (2016). *Residential vacancies and homeownership in the second quarter*. Washington, DC: U.S. Census Bureau, U.S. Department of Commerce.

Campbell, M. A., Canales, D. D., Wei, R., Totten, A. E., Alex, W., & Wershler, J. L. (2015). Multidimensional evaluation of a mental health court: Adherence to the risk-need-responsivity model. *Law and Human Behavior*, *39*, 489-502. doi:10.1037/lhb0000135

Canales, D. D., Campbell, M. A., Wei, R., & Totten, A. E. (2014). Prediction of general and violent recidivism among mentally disordered adult offenders: Test of the Level of Service/Risk–Need–Responsivity (LS/RNR) instrument. *Criminal Justice and Behavior*, *41*, 971-991. doi:10.1177/0093854814523003

Case, B., Steadman, H. J., Dupuis, S. A., & Morris, L. S. (2009). Who succeeds in jail diversion programs for persons with mental illness? A multi-site study. *Behavioral Sciences & the Law*, *27*, 661-674. doi:10.1002/bsl.883

Chenane, J. L., Brennan, P. K., Steiner, B., & Ellison, J. M. (2015). Racial and ethnic differences in the predictive validity of the Level of Service Inventory–Revised among prison inmates. *Criminal Justice and Behavior*, *42*, 286-303. doi:10.1177/0093854814548195

Chu, C. M., Thomas, S. D. M., Ogloff, J. R. P., & Daffern, M. (2011a). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) in a secure forensic hospital: Risk factors and strengths. *International Journal of Forensic Mental Health*, *10*, 337-345. doi:10.1080/14999013.2011.629715

Chu, C. M., Thomas, S. D. M., Ogloff, J. R. P., & Daffern, M. (2011b). The short- to medium-term predictive accuracy of static and dynamic risk assessment measures in a secure forensic hospital. *Assessment*, *20*, 230-241. doi:10.1177/1073191111418298

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., & Tyrer, P. (2006). Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. *Journal of Nervous and Mental Disease*, *194*, 557-564. doi:10.1097/01.nmd.0000230392.83607.c5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge.

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *International Journal of Forensic Mental Health*, *10*, 160-170. doi:10.1080/14999013.2011.600602

de Vries Robbé, M., de Vogel, V., & Douglas, K. S. (2013). Risk factors and protective factors: A two-sided dynamic approach to violence risk assessment. *Journal of Forensic Psychiatry & Psychology*, *24*, 440-457. doi:10.1080/14789949.2013.818162

Desmarais, S. L. (2009). START research summary. In C. D. Webster, M-L. Martin, J. Brink, T. L. Nicholls, & S. L. Desmarais (Eds.), *Manual for the Short-Term Assessment of Risk and Treatability (START)* (Version 1.1, pp. 89-104). Coquitlam, British Columbia, Canada: British Columbia Mental Health & Addiction Services.

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments completed in U.S. correctional settings. *Psychological Services*, *13*, 206-222.

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

Desmarais, S. L., Van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences & the Law*, *30*, 448-469. doi:10.1002/bsl.2022

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services*, *54*, 1372-1379. doi:10.1176/appi.ps.54.10.1372

Douglas, K. S., Yeomans, M., & Boer, D. P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior*, *32*, 479-510. doi:10.1177/0093854805278411

Doyle, M., Lewis, G., & Brisbane, M. (2008). Implementing the Short-Term Assessment of Risk and Treatability (START) in a forensic mental health service. *BJPsych Bulletin*, *32*, 406-408. doi:10.1192/pb.bp.108.019794

Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the Compas: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, *35*, 1095-1108. doi:10.1177/0093854808320497

Ferguson, A. M., Ogloff, J. R. P., & Thomson, L. (2009). Predicting recidivism by mentally disordered offenders using the LSI-R:SV. *Criminal Justice and Behavior*, *36*, 5-20. doi:10.1177/0093854808326525

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, *102*, 813-823. doi:10.1198/016214506000001040

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293-323. doi:10.1037/1076-8971.2.2.293

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, *27*, 237-243. doi:10.1525/fsr.2015.27.4.237

Holder, E. (2014). *Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting*. Retrieved from https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Inett, A., Wright, G., Roberts, L., & Sheeran, A. (2014). Predictive validity of the START with intellectually disabled offenders. *Journal of Forensic Practice*, *16*, 78-88. doi:10.1108/JFP-12-2012-0029

Kelly, C. E., & Welsh, W. N. (2008). The predictive validity of the Level of Service Inventory—Revised for drug-involved offenders. *Criminal Justice and Behavior*, *35*, 819-831. doi:10.1177/0093854808316642

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*, 116-128. doi:10.1080/10509674.2013.868389

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. doi:10.2307/2529310

Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2010). The impact of protective factors in desistance from violent reoffending: A study in three samples of adolescent offenders. *Journal of Interpersonal Violence*, *25*, 568-587. doi:10.1177/0886260509334403

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E.J. (2004). Assessing the inter-rater agreement of the Level of Service Inventory Revised. *Federal Probation*, *68*, 34-38.

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA revisited: Testing the validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services*, *12*, 149-157. doi:10.1037/ser0000024

Lowenkamp, C. T., Lovins, B., & Latessa, E. J. (2009). Validating the Level of Service Inventory—Revised and the Level of Service Inventory: Screening Version with a sample of probationers. *Prison Journal*, *89*, 192-204. doi:10.1177/0032885509334755

Manchak, S. M., Skeem, J. L., & Douglas, K. S. (2008). Utility of the Revised Level of Service Inventory (LSI-R) in predicting recidivism after long-term incarceration. *Law and Human Behavior*, *32*, 477-488. doi:10.2307/30219000

Mauer, M. (2006). The crisis of the young African American male and the criminal justice system. In O. Harris & R. R. Miller (Eds.), *Impacts of incarceration on the African American family* (pp. 199-218). New Brunswick, NJ: Transaction.

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

Nicholls, T. L., Desmarais, S. L., Martin, M.-L., Brink, J., & Webster, C. D. (2006). Short-Term Assessment of Risk and Treatability (START). In R. D. Morgan (Ed.), *The SAGE encyclopedia of criminal psychology*. Thousand Oaks, CA: Sage.

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. doi:10.1037/a0035080

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

Ostermann, M., & Salerno, L. M. (2016). The validity of the Level of Service Inventory–Revised at the intersection of race and gender. *Prison Journal*, *96*, 554-575. doi:10.1177/0032885516650878

Pettit, B., & Western, B. (2004). Mass imprisonment and the life course: Race and class inequality in U.S. incarceration. *American Sociological Review*, *69*, 151-169. doi:10.1177/000312240406900201

Proctor, B. D., Semega, J. L., & Kollar, M. A. (2016). *Income and poverty in the United States: 2015*. Washington, DC: U.S. Census Bureau, U.S. Department of Commerce. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-256.pdf

Richardson, E., & McSherry, B. (2010). Diversion down under: Programs for offenders with mental illnesses in Australia. *International Journal of Law and Psychiatry*, *33*, 249-257. doi:10.1016/j.ijlp.2010.06.007

Rocque, M., & Plummer-Beale, J. (2014). In the eye of the beholder? An examination of the inter-rater reliability of the LSI-R and YLS/CMI in a correctional agency. *Journal of Criminal Justice*, *42*, 568-578. doi:10.1016/j.jcrimjus.2014.09.011

Schlager, M. D., & Simourd, D. J. (2007). Validity of the Level of Service Inventory—Revised (LSI-R) among African American and Hispanic male offenders. *Criminal Justice and Behavior*, *34*, 545-554. doi:10.1177/0093854806296039

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

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

Sirotich, F. (2009). The criminal justice outcomes of jail diversion programs for persons with mental illness: A review of the evidence. *Journal of the American Academy of Psychiatry and the Law*, *37*, 461-472.

Skeem, J. L., & Lowenkamp, C. T. (2015). *Risk, race, & recidivism: Predictive bias and disparate impact* (SSRN Scholarly Paper No. ID 2687339). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2687339

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

Skeem, J. L., Winter, E., Kennealy, P. J., Louden, J. E., & Tatar, J. R. (2014). Offenders with mental illness have criminogenic needs, too: Toward recidivism reduction. *Law and Human Behavior*, *38*, 212-224.

Slinger, E., & Roesch, R. (2010). Problem-solving courts in Canada: A review and a call for empirically-based evaluation methods. *International Journal of Law and Psychiatry*, *33*, 258-264. doi:10.1016/j.ijlp.2010.06.008

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803. Retrieved from https://www.stanfordlawreview.org/print/article/evidence-based-sentencing-and-the-scientific-rationalization-of-discrimination/

Steadman, H. J., Barbera, S. S., & Dennis, D. L. (1994). A national survey of jail diversion programs for mentally ill detainees. *Hospital & Community Psychiatry*, *45*, 1109-1113.

Steadman, H. J., & Naples, M. (2005). Assessing the effectiveness of jail diversion programs for persons with serious mental illness and co-occurring substance use disorders. *Behavioral Sciences & the Law*, *23*, 163-170. doi:10.1002/bsl.640

Steadman, H. J., Osher, F., Robbins, P. C., Case, B., & Samuels, S. (2009). Prevalence of serious mental illness among jail inmates. *Psychiatric Services*, *60*, 761-765. doi:10.1176/appi.ps.60.6.761

Steadman, H. J., Redlich, A., Callahan, L., Robbins, P. C., & Vesselinov, R. (2011). Effect of mental health courts on arrests and jail days: A multisite study. *Archives of General Psychiatry*, *68*, 167-172. doi:10.1001/archgenpsychiatry.2010.134

Sweeney, L. T., & Haney, C. (1992). The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, *10*, 179-195. doi:10.1002/bsl.2370100204

Torrey, E. F., Stieber, J., & Ezekiel, J. (1998). *Criminalizing the seriously mentally ill: The abuse of jails as mental hospitals*. Collingdale, PA: Diane.

Urban Institute. (2016). *Measuring recidivism at the local level: A quick guide*. Retrieved from http://www.urban.org/sites/default/files/recidivism-measures_final-for-website.pdf

U.S. Census Bureau. (2014). *Wealth and asset ownership*. Washington, DC: Author.

U.S. Census Bureau. (2016). *Educational attainment in the United States: 2015*. Washington, DC: Author.

Varela, J. G., Boccaccini, M. T., Murrie, D. C., Caperton, J. D., & Gonzalez, E. J. (2013). Do the Static-99 and Static-99R perform similarly for White, Black, and Latino sexual offenders? *International Journal of Forensic Mental Health*, *12*, 231-243. doi:10.1080/14999013.2013.846950

Vornovitsky, M., Gottschalck, A., & Smith, A. (2011). *Distribution of household wealth in the U.S.: 2000 to 2011*. Washington, DC: U.S. Census Bureau.

Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation*, *72*, 22-29.

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. doi:10.1177/0093854813508916

Walters, G. D. (2007). Using Poisson class regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem. *Criminal Justice and Behavior*, *34*, 1659-1674. doi:10.1177/0093854807307030

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

Webster, C. D., Nicholls, T. L., Martin, M.-L., Desmarais, S. L., & Brink, J. (2006). Short-Term Assessment of Risk and Treatability (START): The case for a new structured professional judgment scheme. *Behavioral Sciences & the Law*, *24*, 747-766. doi:10.1002/bsl.737

Whiteacre, K. W. (2006). Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review*, *17*, 330-342. doi:10.1177/0887403405284766

Wilper, A. P., Woolhandler, S., Boyd, J. W., Lasser, K. E., McCormick, D., Bor, D. H., & Himmelstein, D. U. (2009). The health and health care of US prisoners: Results of a nationwide survey. *American Journal of Public Health*, *99*, 666-672. doi:10.2105/AJPH.2008.144279

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., & Brink, J. (2010). The role of client strengths in assessments of violence risk using the Short-Term Assessment of Risk and Treatability (START). *International Journal of Forensic Mental Health*, *9*, 282-293. doi:10.1080/14999013.2010.534694

Woldgabreal, Y., Day, A., & Ward, T. (2014). The community-based supervision of offenders from a positive psychology perspective. *Aggression and Violent Behavior*, *19*, 32-41. doi:10.1016/j.avb.2013.12.001