Diagnostic and Risk Assessment Characteristics of Offenders Reevaluated for SVP Civil Commitment

Marcus T. Boccaccini, Paige B. Harris, Gabriele F. Trupp, & Jorge G. Varela
Sam Houston State University

[Sexual Offender Treatment, Volume 14 (2019), Issue 1]

Abstract

This study examined risk measure scores, diagnoses, and evaluator opinions for 51 offenders who were evaluated for civil commitment as sexually violent predators (SVP), released, returned to custody, and reevaluated for SVP commitment (M = 5.45 years between evaluations). Most of the offenders had been released on parole or mandatory supervision and had returned to custody for violating the terms of their release. Although there was no evidence of a consistent increase or decrease in risk measure scores or diagnoses over time and only three of the offenders had been arrested for a new sexual offense, evaluators were somewhat more likely to conclude that offenders met criteria for commitment upon reevaluation (81.8%) than upon initial evaluation (60.7%). Despite the average of more than 5-years between evaluations, test-retest reliability estimates for instrument scores and diagnoses were similar to those from field studies of SVP evaluations conducted much closer together in time.

Keywords: sexually violent predator, risk assessment, civil commitment, PCL-R, psychopathy

Sexually Violent Predator (SVP) laws in the United States allow states to civilly commit certain sex offenders for an indefinite period of time, even after they complete their criminal sentences. Most SVP commitment laws follow the criteria that the Supreme Court set forth in Kansas v. Hendricks (1997) and require four elements for commitment: (a) a history of sexual offending, (b) a mental abnormality (sometimes defined as a mental disorder, personality disorder, or "behavioral abnormality"), (c) a volitional impairment rendering him less able to control his sexual behavior (Kansas v. Crane, 2002), and (d) a likelihood of future sexual offending (Miller, Amenta, & Conroy, 2005). SVP referral and commitment procedures routinely involve evaluation results from forensic psychologists and psychiatrists, who conduct risk assessments and evaluate offenders' eligibility for commitment.

Many offenders evaluated for civil commitment as SVPs are not committed (see e.g., DeClue & Rice, 2016; Harris, Boccaccini, & Rice, 2017; Mercado, Jeglic, Markus, Hanson, & Levenson, 2011). These offenders may be released into the community after the SVP evaluator concludes that the offender did not meet criteria for commitment, the prosecutor/petitioner decides against pursuing commitment, or a court ruling leads to the offender's release. For example, one Texas study found that 687 of the 898 (77%) offenders who underwent SVP evaluations between 1999 and 2011 were not committed (Harris et al., 2017). Some of these evaluated but released offenders end up violating the terms of their release (e.g., parole, mandatory supervision) or are arrested for a new sexual offense. In these instances, the offenders may be reevaluated for SVP commitment if they once again become eligible for release.

In this study, we compare findings from initial SVP evaluations and subsequent SVP reevaluations

for 51 sexual offenders from Texas who were evaluated for SVP commitment, released, returned to custody, became eligible for release again, and reevaluated for SVP commitment. Our goals were to provide information about why these offenders returned for reevaluation, whether they appeared to be a unique subset of especially pathological or high-risk offenders, whether their reevaluation results suggested substantial change from their initial evaluation results, and whether the case characteristics associated with evaluators' commitment recommendations changed over time.

The Texas SVP statute defines an SVP as a "repeat sexually violent offender" who "suffers from a behavioral abnormality that makes the person likely to engage in a predatory act of sexual violence" (Texas Health and Safety Code § 841.003). Texas SVP evaluators are asked to make a clinical assessment of behavioral abnormality "based on testing for psychopathy, a clinical interview, and other appropriate assessments and techniques to aid the department in its assessment" (Texas Health & Safety Code §841.023). Approximately one out of every five offenders screened for commitment in Texas gets a full SVP evaluation, and approximately one out of four offenders who gets a full SVP evaluation is ultimately committed (Harris et al., 2017; Murrie, Boccaccini, Caperton, & Rufino, 2012).

There are at least two plausible explanations for why some offenders eventually come to be reevaluated for commitment. First, it is possible that many of the released offenders who were reevaluated were at a high risk for reoffending at the time of their initial release and committed a new sexual offense after release. In Texas, as in other states, some of the offenders that evaluators conclude are eligible for commitment are not committed (DeClue & Rice, 2016; Harris et al., 2017; Mercado et al., 2011). Perhaps these offenders, whom evaluators concluded (per statute) were "likely to engage in a predatory act of sexual violence" committed new sexual offenses after release. Although plausible, this explanation seems unlikely given recent findings of low recidivism rates among sexual offenders released after being deemed eligible for commitment (DeClue & Rice, 2016; Harris et al., 2017).

The second and more plausible explanation is that many of the evaluated offenders were released under some form or supervision (e.g., mandatory supervision, parole) and returned to custody after violating the terms of their supervised release. In other words, as opposed to being (in hindsight) at an especially high risk for reoffending, many of these offenders may have committed a technical violation that led to their return to custody to finish (or partially finish) serving a sentence for one or more of the sexual offenses that led to their initial eligibility for commitment. In a large sample of Texas offenders released after being screened (but not necessarily evaluated) for SVP commitment, about half were discharged under parole or mandatory supervision and could have returned to custody (and been referred for an SVP evaluation) after a supervision violation (Boccaccini, Murrie, Caperton, & Hawes, 2009). Simply put, many released offenders evaluated for SVP are released in such a way that they might return to custody for a technical violation as opposed to a new sexual offense.

If many of those who were reevaluated returned to custody after being charged with a new violent or sexual offense, we might expect to see - in hindsight - that this was an especially high-risk subgroup of offenders that should have stood out from other sexual offenders with respect to risk and psychopathology at the time of their initial evaluations. For example, we might expect to see that they received higher scores on the Static-99 (Hanson & Thornton, 2000) and Psychopathy Checklist-Revised (PCL-R; Hare, 2003), as well as diagnoses associated with SVP commitment recommendations, such as paraphilia and antisocial personality disorder (Levenson & Morin, 2006). If we find high rates of sexual recidivism among the reevaluated offenders but unremarkable instrument scores and diagnoses, it might suggest that factors contributing to these offenders' previously undetected high risk of reoffending were outside the Static-99 and PCL-R (e.g.,

constructs on the STABLE-2007) or that the reevaluated offenders were a subset of truly low risk offenders who happened to reoffend. Regardless of whether these offenders were identifiable as high risk at the time of their initial evaluation, we would expect to see higher scores and diagnosis rates at reevaluation, which would lead to low levels of agreement for instrument scores and diagnoses across evaluations.

If those who were reevaluated returned to custody after violating the terms of their supervised release, we might expect to see changes in some risk measure scores and diagnosis rates, but these changes would not be as notable as they would if offenders had committed new violent or sexual offenses. For example, offenders who violated the terms of their community supervision might receive slightly higher scores on the PCL-R, but the increases would likely not be as large as they would have been if they offenders committed new violent offenses. On the Static-99, offenders scores should only change if they have been charged with or convicted of a new sexual offense. Offenders reevaluated after a return to custody for a technical violation or non-sexual offense should receive exactly the same score at reevaluation.

The data for this study also allow for the first, to our knowledge, examination of the long-term test-retest reliability of instrument scores, diagnoses, and opinions offered by SVP evaluators. Several studies provide information about the reliability of risk measure scores and SVP evaluators' opinions when both evaluations were conducted for the same legal proceeding (Boccaccini, Turner, Murrie, & Rufino, 2012; DeMatteo et al., 2014; Hanson, 2001; Levenson, 2004; Murrie et al., 2009; Miller, Kimmonis, Otto, Kline, & Wasserman, 2012; Perillo, Spada, Calkins, & Jeglic, 2014). Findings from these studies suggest that evaluator agreement for scores on measures such as the Static-99 and PCL-R are often, but not always, weaker in this type of field setting than they are in more controlled research settings. For example, agreement values (intraclass correlation coefficients: ICC) for PCL-R total scores range from .40 to .84 in SVP cases, while those for Static-99 total scores range from .58 to .98. For diagnoses, researchers have found that agreement is generally modest to poor across most diagnostic categories, but stronger for antisocial personality disorder (.51 to .54) and pedophilia (.55 to .65) diagnoses than other paraphilia diagnoses (.01 to .35; Levenson, 2004; Perillo et al., 2014). The one study examining evaluator agreement for civil commitment recommendations reported an agreement value of =.54 (Levenson, 2004).

Despite this wealth of research, however, the extent to which these reliability findings may generalize to our sample is unclear. The average amount of time between initial evaluations and reevaluations in our study is 5.45 (SD = 2.70) years, and some changes in instrument scores and diagnoses may be due to changes in evidence supporting risk item scores, diagnoses, and evaluator opinions. For example, a previously undiagnosed offender may be diagnosed with a paraphilic disorder if evidence of ongoing or renewed deviant sexual interest and behavior is revealed during reassessment. Similarly, a previously diagnosed offender may not be diagnosed at reevaluation if assessment data reveal that deviant sexual interest and/or behavior has abated or reduced to subclinical levels. In the case of PCL-R scores, behavior between assessments may necessitate changes to item scores and, thus, associated facet, factor, and total scores. For example, an offender may be involved in additional marital-like relationships, engage in new interpersonal violence, engage in acts of fraud or exploitation, or other behaviors that may lead to increased scores on corresponding PCL-R items.

It is also possible, however, that test-retest reliability in our sample may be similar to rater-agreement findings in other SVP samples. Two examinations of the stability of PCL-R total scores (from non-SVP samples) over longer periods of time (M > 2.00 years) reported ICC values that were consistent with other field reliability studies (ICC = .60 to 70; Rutherford, Cacciola, Alterman, McKay, & Cook, 1999; Sturup et al.,, 2014). The well-known measurement phenomenon

of regression to the mean explained much of the disagreement between initial and reevaluation scores in one of these studies (Sturup et al., 2014). Specifically, offenders with especially high or low scores at the time of the initial evaluation tended to have scores that were closer to the sample mean (i.e., less extreme) at reevaluation.

Finally, data for this study also allow us to examine whether the offender characteristics (e.g., diagnoses, instrument scores) that predicted evaluator opinions of perceived risk for reoffending and eligibility for commitment were the same for initial evaluations and reevaluations. One possible explanation for offenders returning for reevaluation is that the initial evaluators failed to consider the best available risk factors (e.g., Static-99 scores) when formulating their opinions. Existing studies suggest that evaluators' opinions of sexual offender risk and eligibility for commitment will be associated with Static-99 scores, PCL-R scores, and diagnoses of paraphilia and antisocial personality disorder (Gardner, Boccaccini, & Murrie, 2018; Levenson & Morin, 2006; McCallum, Boccaccini, & Bryson, 2017).

Method

Participants

Data for this study come from a sample of 687 male offenders who were evaluated for SVP commitment in Texas between October of 1999 and November of 2011, but released after the state decided against pursuing commitment (Harris et al., 2017). During this timeframe, 51 (10.8%) of these released offenders returned to custody, became eligible for release, and were reevaluated for SVP commitment. This study focuses on these 51 offenders.

Most of the reevaluated offenders had been released on mandatory supervision or parole and returned to custody because they violated the terms of their release (n = 43, 84.3%). Four of the offenders (7.8%) returned to custody after being charged with a new offense. Two of the offenders had been charged with one new sexual assault, one had been charged with two new sexual assaults, and one had been charged with stalking. For four offenders (7.8%), the reason why the offender returned to custody and was reevaluated was unclear from the information we coded for the larger recidivism study (Harris et al., 2017), although we do know that these four offenders had not been charged with a new sexual offense or other type of violent offense during their release.

The average amount of time between SVP evaluations was 5.45 (*SD* = 2.70) years (range = 0.36 to 10.41). The offenders' mean age was 42.35 (*SD* = 7.80) years at the time of their initial evaluations and 47.86 (*SD* = 8.44) years at the time of their reevaluations. Only one offender was below the age of 25 years at the time of the initial evaluation. Evaluators identified offenders as either White/Caucasian (*n* = 24, 47.1%), Black/African-American (*n* = 20, 39.2%), or Hispanic/Latino (*n* = 7, 13.7%).

Evaluators and Evaluator Reports

We collected information from behavioral abnormality evaluation reports submitted to the state's department of corrections by the SVP evaluators. Evaluators typically meet with and interview the offender as part of the evaluation process, although some offenders refuse to participate in the interview. The evaluator submits a written evaluation report describing their findings to the department of corrections. Although the structure and format of these reports vary from evaluator to evaluator, they typically provide information about diagnoses, psychopathy, risk, and an opinion regarding whether or not they believe the offender has a behavioral abnormality (i.e., meets criteria

for commitment).

Seventeen evaluators conducted at least one behavioral abnormality evaluation for an offender included in this study (M = 6.00 evaluations, SD = 5.22, range = 1 to 19). There was no case in which the same evaluator conducted the initial evaluation and the reevaluation (i.e., there were two different evaluators for each offender).

Based on the information available for this study, we do not know how often the second evaluator (reevaluation) had access to the initial evaluator's report. One of the study authors (JGV) conducts SVP evaluations in Texas and has been asked to perform a small number of reevaluations (for offenders reevaluated after the timeframe of this study). He has had cases in which the initial evaluation report was in the offender's records and cases in which the initial evaluation report was not in the offender's records.

Measures

Psychopathy Checklist-Revised (PCL-R). The PCL-R is a 20-item, clinician-scored measure of interpersonal, affective, lifestyle, and behavioral traits associated with psychopathy (Hare, 2003). The clinician assigns a score of 0 (*not present*), 1 (*possibly present*), or 2 (*definitely present*) to quantify the degree to which the examinee manifests particular psychopathy criteria. In Texas SVP cases, evaluators typically score the PCL-R on the basis of a clinical interview and collateral record review. When offenders refuse to be interviewed, evaluators score the PCL-R if the collateral records provide sufficient information for scoring; however, evaluators sometimes conclude that they cannot score the PCL-R from records alone.

The 20 PCL-R items can be combined to provide a total score, two factor scores (Factor 1 = Interpersonal/Affective, Factor 2 = Social Deviance), and four facet scores (Interpersonal, Affective, Lifestyle, Antisocial). In our sample, evaluators reported PCL-R total scores in 92.1% (n = 47) of their initial evaluation and reevaluation reports, factor scores in 62.7% (n = 32) of their initial evaluations and 45.1% (n = 23) of their reevaluations, and facet scores in no initial evaluations and only 5.9% (n = 3) of their reevaluations.

The PCL-R is the second most commonly used measure in sex offender risk assessment (Boccaccini, Chevalier, Murrie, & Varela, 2017; Neal & Grisso, 2014). Meta-analyses show a small, but consistent association between PCL-R scores and sexual recidivism (total score d = 0.40; Hawes, Boccaccini, & Murrie, 2013). The PCL-R manual (Hare, 2003) reports total score rater agreement values (ICC_1) from .86 for male inmates to .88 for male forensic psychiatric patients. However, as emphasized above, agreement values appear poorer in the field.

Static-99. The Static-99 (Hanson & Thornton, 2000) is a 10-item actuarial risk assessment instrument developed to predict sexual recidivism among adult male sex offenders (see www.static99.org). The Static-99 items broadly assess age, criminal history, victim characteristics, and relationship history. Static-99 total scores are computed by summing scores on the 10 items, with possible scores ranging from 0 to 12. Based on the total score, offenders can be assigned to one of the following risk categories: low (0 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6+). The estimated median score for the Static-99 is 2 (Hanson, Lloyd, Helmus, & Thornton, 2012). Although the Static-99 developers released a revised version of the measure in 2012 (Static-99R; Helmus, Thornton, Hanson, & Babchishin, 2012) and it is possible to transform Static-99 to Static-99R scores, we conducted analyses with the original Static-99 scores because they are the scores that the evaluators used for decision-making. Meta-analytic findings indicate that Static-99 (d = .67; Hanson & Morton-Bourgon, 2009) and Static-99R (AUC = .69; Helmus,

Hanson, Thornton, Babchishin, & Harris, 2012) scores have moderate and similar predictive accuracy for sexual recidivism.

It was more common for evaluators to have used the Static-99 for reevaluations (n = 42) than initial evaluations (n = 20), likely due to the Static-99 being relatively new at the time many of the offenders were initially evaluated. For example, 20 of the offenders were initially evaluated in 1999 or 2000, and none of these offenders were scored on the Static-99 by the initial behavioral abnormality evaluator. The first Static-99 score from an initial evaluator was assigned in March of 2001.

Diagnosis. Evaluators reported diagnoses in their behavioral abnormality evaluation reports. We coded reports for the assignment of paraphilia and antisocial personality disorder diagnoses because they were the most common disorders reported by evaluators and because they are associated with evaluators' civil commitment recommendations in other states (Levenson & Morin, 2006). Due to the small sample and the low base rate of some paraphilia diagnoses, we created three paraphilia diagnosis categories for analysis: a) pedophilia, b) any paraphilia other than pedophilia, and c) any paraphilia (i.e., either pedophilia or any other paraphilia diagnosis). We also used an additional diagnostic category of either paraphilia or antisocial personality disorder for examining the association between diagnoses and evaluator decision-making.

Behavioral abnormality opinions. Although reporting practices vary from evaluator to evaluator, it is common for evaluators to provide a statement in their reports about the offender's overall level of risk for sexual reoffending and an opinion about whether or not the offender has a behavioral abnormality (i.e., meets statutory criteria for commitment). For each evaluation, we coded the risk statements into one of seven risk-level categories (1 = *Low*, 2 = *Low to moderate*, 3 = *Moderate*, 4 = *Moderate to high*, 5 = *High*, 6 = *High to very high*, 7 = *Very high*). Because no evaluator ever concluded that an offender was at a low risk for reoffending, the scores on this rating range from 2 to 7.

We used a single dichotomous variable to indicate whether the evaluator stated that the offender did or did not have a behavioral abnormality (1 = *Behavioral abnormality*, 0 = *No behavioral abnormality*). Texas SVP evaluators are asked to determine whether or not the offender has a behavioral abnormality that makes him or her likely to engage in repeated predatory acts of sexual violence. During the timeframe of this study, offenders who were determined to meet statutory requirements for a behavioral abnormality were referred to the state's Special Prosecution Unit, which ultimately decided against pursuing commitment.

Procedure

We obtained PCL-R scores, Static-99 scores, diagnoses, and evaluator opinions from evaluator reports. All files and reports were coded by the second author at a department of corrections administrative office. We double coded 10 reports as a reliability check and found 100% agreement across all variables.

Results and Discussion

Table 1 provides descriptive statistics for all of the study variables, including values based on all available data (labeled as "all" in the table) and values based on cases with data for the variable at both the initial evaluation and reevaluation (labeled as "both" in the table). Offenders' initial evaluation Static-99 (M = 4.60 to 4.72) and PCL-R scores (M = 20.58 to 20.94) appeared to be

more similar to the mean scores from Texas offenders who have been evaluated, but not committed (Static-99 = 4.25, PCL-R = 19.29), than to those who have been evaluated and committed (Static-99 = 5.91, PCL-R = 25.03; Boccaccini et al., 2009; Murrie et al., 2012, Harris et al., 2017). Both the initial and subsequent evaluation PCL-R scores were somewhat lower than the average total PCL-R score of 24.2 for male sex offenders described in the PCL-R Manual (Hare, 2003). In other words, these initial evaluation scores did not appear to suggest that this was a subgroup of especially high risk or psychopathic offenders within the larger SVP evaluation sample. This conclusion is also consistent with the finding that only three (5.9%) offenders were rearrested for a new sexual offense.

Table 1: Comparison of Diagnoses, Evaluator Opinions, and Instrument Scores from Initial SVP Evaluations and SVP Reevaluations

	Initial evaluation			**Reevaluation**				
Diagnosis/Opinion/Score	***N***	***n* yes**	**% yes**	***N***	***n* yes**	**% yes**	**Comparison effect size**	**Agreement**
Pedophilia diagnosis (both)	51	12	23.5	51	15	29.4	OR = 2.50 [0.49, 12.89], p = .45	86%, = .65
Other paraphilia diagnosis (both)	51	8	15.7	51	10	19.6	OR = 1.33 [0.46, 3.84], p = .97	73%, = .06
Any paraphilia diagnosis (both)	51	18	35.3	51	23	45.1	OR = 1.83 [0.68, 4.96], p = .33	67%, = .31
Antisocial PD diagnosis (both)	51	23	45.1	51	21	41.2	OR = 1.29 [0.49, 3.45], p = .80	69%, = .36
Behavioral abnormality (all)	28	17	60.7	44	36	81.8		
Behavioral abnormality (both)	26	15	57.7	26	20	76.9	OR = 3.50 [0.72, 16.85], p = .18	65%, = .25
	N	***M***	***SD***	***N***	***M***	***SD***	**Comparison**	**Agreement**
PCL-R Total (all)	47	20.94	8.26	47	21.96	6.61		
PCL-R Factor 1 (all)	32	9.74	3.87	23	9.78	3.85		
PCL-R Factor 2 (all)	31	10.20	4.09	23	9.52	4.10		
Static-99 (all)	20	4.60	1.43	42	4.64	1.61		
Overall risk opinion (all)	32	4.50	0.92	34	4.50	0.99		
PCL-R Total (both)	45	20.58	8.01	45	21.84	6.64	d = .17, p = .24	$ICC_{A,1}$ = .53
PCL-R Factor 1 (both)	15	9.78	3.31	15	9.80	3.63	d = .003, p = .99	$ICC_{A,1}$ = .34
PCL-R Factor 2 (both)	14	9.96	4.02	14	9.93	3.59	d = .-.01, p = .98	$ICC_{A,1}$ = .41
Static-99 (both)	18	4.72	1.36	18	5.06	1.51	d = .24, p = .25	$ICC_{A,1}$ = .66
Overall risk opinion (both)	24	4.67	0.82	24	4.58	0.83	d = -.11, p = .71	$ICC_{A,1}$ = .11

Note. all = data reported for all offenders with information at that time point. Both = data reported for offenders with information at both initial evaluation and reevaluation All agreement analyses and inferential statistical comparisons between initial evaluation and reevaluation are limited to cases in which information was available at both time points. N = number of cases with data. For categorical variables, n yes = number of cases in which the diagnosis or opinion was given. OR and p values

are based on McNemar's test for correlated proportions. = Cohen's kappa. For numerical ratings, *p* values are based on paired-samples t-tests. *d* = Cohen's *d*. $ICC_{A,1}$ = Single rater, absolute agreement intraclass correlation coefficient. $ICC_{A,2}$ = Multiple rater, absolute agreement intraclass correlation coefficient.

The initial evaluation rates of pedophilia and any paraphilia (23.5% & 35.4%; see Table 1) were also consistent with the larger sample of evaluated but released Texas offenders (27.9% & 35.4%; see Harris et al., 2017), although the rate of antisocial personality disorder diagnoses was somewhat higher (45.1% vs. 35.7%). The rate of concluding that offenders had a behavioral abnormality at the time of the initial evaluation (60.7%) was slightly lower than the overall SVP evaluation sample rate of 70%.[1]

One notable pattern of findings in Table 1 is that (using all available data) the rate of concluding that the offender had a behavioral abnormality increased from initial evaluations (60.7%) to reevaluations (81.8%), even though most of the offenders had not committed a new sexual offense while released. We would expect that elapsed time without being charged with a new sexual offense (even if under supervision) would serve as evidence of decreased behavioral abnormality. One possible explanation for the increase is that evaluators viewed violating the terms of supervised release (i.e., the offenders' inability to meet the expectations of their supervision) as an indicator of risk and justification for a behavioral abnormality conclusion. Considering that the Static-99 and PCL-R scores were on average largely unchanged, it may be that evaluators place significant weight on violations while under supervision when coming to conclusions about behavioral abnormality.

Although technical violations may seem unrelated to a mental disorder diagnosis - other than, perhaps, antisocial personality disorder - it is possible that evaluators view these violations as related to behavioral abnormality in light of Texas case law. Specifically, the Supreme Court of Texas has concluded that the predisposition to commit a sexually violent offense - that is, elevated risk - defines behavioral abnormality in SVP cases and that a formal diagnosis is unnecessary (see *In Re Commitment of Michael Bohannan*, 2012). In light of this definition, evaluators who perceive a technical violation to be an indicator of increased risk may deliver an opinion favoring a behavioral abnormality, even if the technical violation does not lead to a new mental disorder diagnosis.

Test-Retest Reliability: Initial Evaluation to Reevaluation

Table 1 also provides information about the consistency of instrument scores, diagnoses, and evaluator opinions across evaluations (i.e., across time). These analyses focus on cases for which both evaluators provided scores, diagnoses, or opinions. Because different evaluators reported different information, the number of cases with the same information from both evaluators ranges from 14 cases (Static-99) to 51 cases (diagnoses; see Table 1).

Because all of the offenders had been released and returned to custody, and because there was an average of 5.45 (*SD* = 2.70) years between evaluations, events that occurred between the evaluations may have led to legitimate changes in PCL-R scores, diagnoses, and opinions. Thus, we did not necessarily expect scores and opinions to be as consistent as they would be in a more typical examination of test-retest reliability, with less time between evaluations. In our sample, test-retest analyses provide information regarding whether the level of consistency in this long-term reevaluation context is notably different from that found in more typical reevaluation contexts (i.e., two evaluators for the same legal proceeding) and whether there is any evidence for a pattern of

increasing or decreasing risk and psychopathology over time.

Diagnoses. Despite the extended length of time between evaluations, our reliability findings for paraphilia diagnoses were generally consistent with those from other SVP evaluation samples (Levenson, 2004; Perillo et al., 2014). Reliability was strongest for pedophilia diagnoses (= .65), falling within the =.55 to .65 range reported in other samples. Similarly, reliability was weaker for other paraphilia diagnoses (= .06) and any paraphilia diagnosis (= .31), just as it has been for other paraphilia diagnoses in other SVP samples (.01 to .35). However, reliability for antisocial personality disorder diagnoses was somewhat weaker in our sample (= .36) than other SVP evaluation samples (.51 to .54). Although all diagnoses, with the exception for antisocial personality disorder, were somewhat more common for reevaluations than initial evaluations, none of these increases were large enough to reach statistical significance (see Table 1).

Static-99 scores. The single evaluator absolute agreement intraclass correlation ($ICC_{A,1}$) for Static-99 scores was .66, which is generally consistent with the $ICC_{A,1}$ values for scores assigned by evaluators in Texas SVP cases ($ICC_{A,1}$ .58 to .64; Boccaccini et al., 2009; Murrie et al., 2009). Researchers have reported much stronger agreement values (ICC .80 to .90) among SVP evaluators in other states (e.g., California, Florida, New Jersey; Boccaccini, Murrie et al., 2012; Hanson, 2001; Levenson, 2004; Miller et al., 2012), and among department of corrections employees in Texas ($ICC_{A,1}$ = .81, Rice, Boccaccini, Harris, & Hawes, 2014).

With only 18 offenders having pairs of Static-99 scores in the current study, the reasons for the lower agreement values among Texas SVP evaluators are unclear, and they may be different from the reasons for disagreement in prior studies (e.g., adversarial allegiance; Murrie et al., 2009). There was a slight increase in Static-99 scores over time (d = .24), although this difference was not large enough to reach statistical significance in this small sample (see Table 1). Scores for each of the three offenders with new sexual assault charges increased from initial evaluation to reevaluation (two increased by one point and one increased by two points), but excluding these three offenders did not lead to a notable improvement in agreement ($ICC_{A,1}$ = .67, n = 15). For the 15 offenders with no new sexual offense, there should have been no substantive reason for the Static-99 score to change. It is possible that the two evaluators were working with different sets of records, which could have influenced their judgments about specific items, or that subsequent evaluators failed to recognize that post-release behaviors should lead to a change in scoring for only those offenders with new sexual offenses.

PCL-R scores. Agreement for PCL-R Total scores ($ICC_{A,1}$ = .53) was consistent with mid-range to lower-range estimates of ICC values from other test-retest reliability studies (ICC .40 to .84), including those of scores assigned by SVP evaluators in Texas (Boccaccini, Turner et al., 2012), SVP evaluators in other states (DeMatteo et al., 2014; Levenson, 2004; Miller et al., 2012), and scores assigned several years apart by evaluators in other clinical-forensic settings (Rutherford et al., 1999; Sturup et al., 2014). Our agreement findings for Factor 1 scores (ICC = .31) fall squarely within the range reported in other sex offender samples (ICC .15 to.48), whereas our Factor 2 findings (ICC = .41) are somewhat lower than those from other samples (ICC .55 to .75; Edens, Boccaccini, Johnson, & Johnson, 2010; Miller et al., 2012). There was, however, no evidence that disagreement for PCL-R total or factor scores was due to a systematic increase or decrease in scores over time, with Cohen's d values ranging from only -.01 to .17 for PCL-R scores (see Table 1).

Following Sturup et al. (2014), we used a Galton Squeeze Diagram (see Figure 1) as a visual aid to examine whether regression toward the mean might help explain the modest level of test-retest reliability for PCL-R scores (see Figure 1). Using the mean PCL-R score from the initial evaluation

(20.58) as a reference point, there was some evidence that the most extreme scores (very low, very high) came back toward the mean at reevaluation. There were 19 cases with score differences of six or more points, which is about twice the standard error of measurement for PCL-R total scores (Hare, 2003). Of these 19 cases, 12 (63.2%) had scores that came back toward the mean at reevaluation, whereas seven (36.8%) had scores that moved further away from the mean at reevaluation. None of these seven increasing score cases included the four offenders who had been arrested for a new violent offense. In fact, the score for the one offender with two new sexual assault charges actually decreased by 8 points, from 30 to 22.

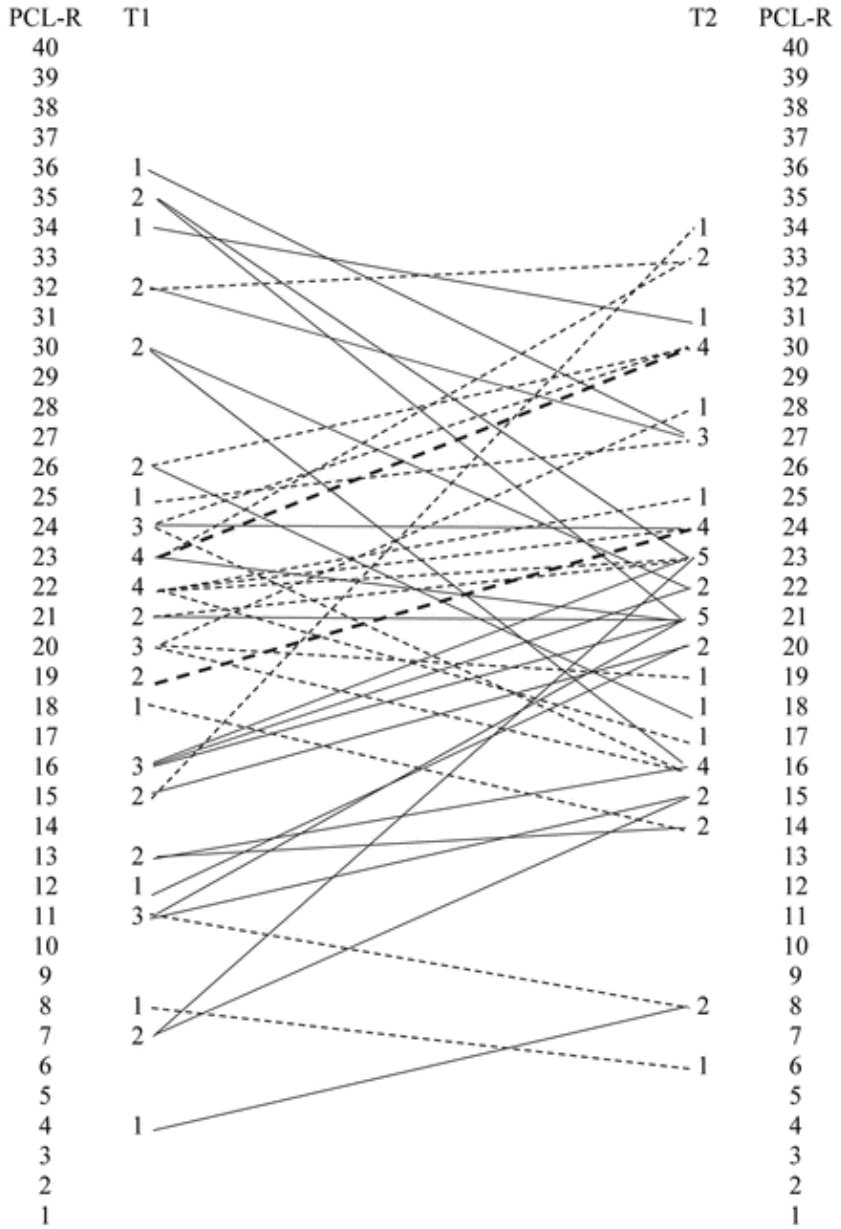

Figure 1: Galton squeeze diagram of PCL-R total scores at initial evaluation (T1) and reevaluation (T2). A solid line indicates a pair of scores for which the second score moved back toward the initial sample mean of 20.58. A dashed line indicates a pair of

scores for which the second score moved further away from the mean.

Again, following Sturup et al. (2014), we used a bivariate correlation to examine whether there were larger PCL-R difference scores (initial minus reevaluation) for cases with initial evaluation scores that were furthest from the initial evaluation group mean (initial score minus group mean). The correlation between the absolute value of these difference scores was .35 (p = .02). Thus, the further the initial score was from the initial group mean (in either direction), the larger the subsequent difference between initial and reevaluation scores.

Perceived risk and behavioral abnormality. In terms of effect size, the largest increase from initial evaluation to reevaluation (among those with data for both evaluations) was for opinions relating to behavioral abnormality (OR = 3.50, p = .18), with evaluators identifying 57.7% of offenders as having a behavioral abnormality at the initial evaluation and 76.9% as having a behavioral abnormality at reevaluation. Although not large enough to reach statistical significance, this pattern helps to explain the low level of agreement for behavioral abnormality opinions over time (= .25, see Table 1). Agreement for this type of ultimate SVP opinion is higher in samples of offenders evaluated for the same legal proceeding (= .54, Levenson, 2004). Evaluators concluded that all three of the offenders who had been charged with a new sexual offense had a behavioral abnormality at reevaluation, but this did not help to explain the low level of agreement across evaluations. Only two of these offenders had a behavioral abnormality opinion from the initial evaluator, and both of these initial opinions stated that the offender met criteria for having a behavioral abnormality.

Agreement was lower for perceived risk for future sexual violence than for any other study variable ($ICC_{A,1}$ = .11). There was no evidence that this level of disagreement was attributable to a systematic increase in perceived risk over time, as the mean Static-99 risk ratings were very similar for initial evaluations (M = 4.67, SD = .82) and reevaluations (M = 4.58, SD = .83; see Table 1).

Predictors of Risk and Behavioral Abnormality Opinions

The low levels of agreement for evaluators' risk and behavioral abnormality opinions raise questions about whether these opinions might be associated with different offender characteristics (e.g., diagnoses, instrument scores) for initial evaluations and reevaluations. We used bivariate correlations[2] to examine the extent to which PCL-R scores, Static-99 scores, and diagnoses were associated with risk and behavioral abnormality opinions (see Table 2).

Table 2: Instrument Scores and Diagnoses as Predictors of Evaluators' Risk and Behavioral Abnormality Opinions

	Risk for recidivism				Behavioral abnormality opinion			
	Initial evaluation		Reevaluation		Initial evaluation		Reevaluation	
Score/Diagnosis	***r***	***n***	***r***	***n***	***r***	***n***	***r***	***n***

PCL-R Total	.28	30	.36*	31	.54**	24	.28	41
PCL-R Factor 1	.44^	19	.17	11	.55*	16	.59**	22
PCL-R Factor 2	.37	19	-.09	11	.67**	16	.27	22
Static-99	.44	15	.60***	29	-.22	14	.58***	40
Antisocial PD	.25	33	.21	34	.28	26	.15	44
Pedophilia	.17	33	-.02	34	.11	26	.18	44
Other paraphilia	.10	33	.34*	34	.33	26	.24	44
Any paraphilia	.19	33	.15	34	.24	26	.33*	44
Any para or antisocial	.33^	33	.28	34	.32	28	.41**	44

Note. ***$p < .001$. **$p < .01$. *$p < .05$. ^$p = .06$. Risk for recidivism: 1 = *low risk*, 7 = *very high risk*. Behavioral abnormality opinion: 1 = behavioral abnormality, 0 = no behavioral abnormality.

Although these correlations are difficult to compare directly (due to missing data for many offenders) there are several overall trends that are apparent in Table 2. For both initial evaluations and reevaluations, all but two of the correlations were positive in direction indicating that higher instrument scores (PCL-R, Static-99) and the presence of paraphilia and/or antisocial personality disorder diagnoses were generally associated with higher ratings of perceived risk and an increased likelihood of concluding that the offender had a behavioral abnormality. No single diagnosis stood out as a predictor of behavioral abnormality opinions or perceived risk, but having either a paraphilia or antisocial personality disorder diagnosis[3] was most consistently associated with risk and behavioral abnormality opinions (r = .28 to r = .41, see Table 2).

Static-99 total scores were the most consistent predictors of perceived risk for sexual reoffending ($r = .44$, $p = .10$ for initial evaluations, $r = .60$, $p = .001$ for reevaluations), but they were only predictive of behavioral abnormality opinions for reevaluations ($r = .58$, $p < .001$). For the PCL-R, Factor 1 scores were more consistently associated with perceived risk ($r = .44$, $p = .06$ for initial evaluations) and behavioral abnormality opinions ($r = .55$, $p = .02$ for initial evaluations, $r = .59$, $p = .004$ for reevaluation) than Factor 2 or Total scores, although total scores were also small to moderate predictors of evaluator opinions ($r = .28$ to .54, see Table 2).

These findings for the Static-99 add to the growing body of field and experimental studies showing that evaluators' opinions in sex offender risk assessment cases are associated with the scores they assign on the Static-99 and other actuarial risk assessment instruments (Gardner et al., 2018; McCallum et al., 2017). They also suggest that the size of this association may have increased over time, with evaluators placing more weight on these scores as the volume of research supporting their use has increased over time. These findings are also generally consistent with what we know about the predictive properties of scores on these measures (Hanson & Morton-Bourgon, 2009; Helmus et al., 2012). The PCL-R factor score findings add to those from a recent experimental study suggesting that clinicians are more strongly influenced by sexual offenders' Factor 1 scores than Factor 2 scores (see Gardner et al., 2018), despite the stronger empirical support for Factor 2 scores as predictors of future violence among sexual offenders (Hawes et al., 2013).

Results Summary

In sum, our results suggest that those who returned for a second SVP evaluation were generally similar to other offenders who had been evaluated and released, as opposed to a select group of high-risk offenders. Although most of the offenders had returned to custody for a technical violation, evaluators appeared to be somewhat more likely to conclude that the offenders met criteria for commitment at reevaluation, for reasons that are not entirely clear from our findings. Reliability analyses revealed levels of agreement for instrument scores and diagnoses that were generally consistent with other field studies, even though the amount of time between evaluations was much longer in this study than in previous studies. At reevaluation, Static-99 scores were the strongest predictors of both perceived risk and eligibility for commitment.

Limitations and Conclusions

There are several methodological limitations that necessarily limit the conclusions that can be drawn from this study. Because of variations in evaluation and reporting practices, especially among early SVP evaluations, many offenders did not have the same types of data in their reports for both their initial evaluations and reevaluations. Thus, many analyses included data from only a subset of offenders and many analyses included only partially overlapping subsets of offenders. These aspects of this study make it difficult to make direct comparisons between variables and limit the statistical power of inferential statistical tests. The findings are also limited in that we did not collect information about the behaviors that led to the offenders' parole and mandatory supervision violations, which would have helped us to better explain why the offenders returned to custody. We coded nearly 900 SVP reports for a larger study examining sexual recidivism (Harris et al., 2017) and did not anticipate the potential value of detailed information about these technical violations.

Despite these limitations, the 51 cases in this study allowed for a unique examination of the risk and diagnostic characteristics of the relatively small group of convicted sexual offenders who were reevaluated for SVP commitment over a 12-year period. Our findings show that these reevaluations typically occurred when offenders returned to custody for a violation of their supervised released. Although most of these offenders had not been arrested for a new sexual offense during the timeframe of their release and there was no clear evidence of increased risk on the Static-99 or PCL-R scores, most evaluators concluded that the offenders met criteria for commitment at the time of reevaluation, perhaps because many of the offenders had violated the terms of their conditional release. Or perhaps the evaluators tended to conclude that these offenders met criteria for commitment because these cases are unusual and uncommon. The fact that the offender was referred for a second evaluation in and of itself may have in a way primed evaluators into seeing these offenders as requiring commitment, even though their risk scores and diagnoses were not notably different at reevaluation.

The current findings also have several implications for practitioners. The increased rate of behavioral abnormality opinions upon reevaluation in the absence of new offenses raises the possibility that clinicians may be misled into judging an offender as having elevated risk for sexual reoffending when only new technical violations of supervision are present. Although technical violations related to inappropriate sexually-related behavior, such as contact with minors, may understandably raise concerns about risk, most technical violations are arguably unrelated to risk for sexual reoffense.

The increase in Static-99 scores over time also reveals important considerations for practitioners. As indicated in the instrument manual (Phenix, Fernandez, Harris, Helmus, Hanson, & Thornton,

2016), an offender's score should be based on the date of initial release for the index sexual offense and remain the same unless the offender commits a new sexual offense. Thus, the small increase we observed in Static-99 scores in the absence of new offenses was unexpected and may be due to some practitioners using post-release behavior to assign item scores. Finally, recent research has revealed that time in the community without committing an offense is associated with reduced risk of recidivism (Hanson, Harris, Helmus, & Thornton, 2014). If practitioners' opinions were consistent with this research, we should have seen a reduction in estimates of risk and opinions of behavioral abnormality in this study, as many offenders had been in the community for a significant amount of time without being charged with a new sexual offense. Thus, our findings suggest that offense-free time in the community may be underweighted by evaluators when coming to conclusions about risk for reoffense.

Author Note

The research contained in this document was coordinated in part by the Texas Department of Criminal Justice (587-AR09). The contents of this document reflect the views of the authors and do not necessarily reflect the views or policies of the Texas Department of Criminal Justice.

References

1. Boccaccini, M. T., Chevalier, C., Murrie, D. C., & Varela, J. G. (2017). Psychopathy Checklist-Revised use and reporting practices in sexually violent predator evaluations. Sexual Abuse: A Journal of Research and Treatment, 29, 592-614. doi:10.1177/1079063215612443
2. Boccaccini, M. T., Murrie, D. C., Caperton, J. D., & Hawes, S. W. (2009). Field validity of the STATIC-99, and MnSOST-R among sex offenders evaluated for commitment as sexually violent predators. Psychology, Public Policy, and Law, 15, 278-314. doi: 10.1037/a0017232
3. Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A., & Jeglic, E. (2012). Implications of Static-99 field reliability findings for score use and interpretation. Criminal Justice and Behavior, 39, 42-58. doi: 10.1177/0093854811427131
4. Boccaccini, M. T., Turner, D., Murrie, D. C., & Rufino, K. A. (2012). Do PCL-R scores from state or defense experts best predict future misconduct among civilly committed sex offenders? Law and Human Behavior, 36, 159-169. doi: 10.1037/h0093949
5. DeClue, G., & Rice, A. (2016). Florida's released "sexually violent predators" are not "high risk." Open Access Journal of Forensic Psychology, 8, 22-51.
6. DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., & Formon, D. (2014). The role and reliability of the Psychopathy Checklist-Revised in U.S. Sexually Violent Predator evaluations: A case law survey. Law and Human Behavior, 38, 248-255.
7. Edens, J., Boccaccini, M. T., Johnson, D., & Johnson, J. (2010). Inter-rater reliability of the PCL-R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioral features? Behavioral Sciences and the Law, 28, 106-119. doi: 10.1002/bsl.918
8. Gardner, B. O., Boccaccini, M. T., & Murrie, D. C. (2018). Which PCL-R scores best predict forensic clinicians' opinions of offender risk? Criminal Justice and Behavior, 45, 1404-1419. doi: 10.1177/0093854818789974
9. Hanson, R. K. (2001). Note on the reliability of Static-99 as used by California DMH evaluators (Unpublished Report). Sacramento, CA: California Department of Mental Health.
10. Hanson, R. K., Harris, A. J. R., Helmus, L., & Thornton, D. (2014). High risk sex offenders may not be high risk forever. Journal of Interpersonal Violence, 29, 2792-2813. doi: 10.1177/0886260514526062

11. Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. International Journal of Forensic Mental Health, 9, 11-23. doi:10.1080/14999013.2012.667511
12. Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. Psychological Assessment, 21, 1-21. doi: 10.1037/a0014421
13. Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. Law and Human Behavior, 24, 119-136. doi:10.1023/A:1005482921333
14. Hare, R. D. (2003). Hare PCL-R, 2nd edition. New York: Multi-Health Systems.
15. Harris, P. B., Boccaccini, M. T., & Rice, A. K. (2017). Field measures of psychopathy and sexual deviance as predictors of recidivism among sexual offenders. Psychological Assessment, 29, 639-651. doi:10.1037/pas0000394
16. Hawes, S. M., Boccaccini, M. T., & Murrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist-Revised. Psychological Assessment, 25, 233-243. doi:10.1037/a0030391
17. Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. Criminal Justice and Behavior, 39, 1148-1171. doi:10.1177/0093854812443648
18. Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. Sexual Abuse: A Journal of Research and Treatment, 24, 64-101. doi: 10.1177/1079063211409951
19. In Re Commitment of Michael Bohannan, 388 S.W.3d 296 (Tex. 2012)
20. Kansas v. Crane, 534 U.S. 407 (2002).
21. Kansas v. Hendricks, 521 U.S. 346 (1997).
22. Levenson, J. S. (2004). Reliability of sexually violent predator civil commitment criteria in Florida. Law and Human Behavior, 28, 357-368.
23. Levenson, J. S., & Morin, J. W. (2006). Factors predicting selection of sexually violent predators for civil commitment. International Journal of Offender Therapy and Comparative Criminology, 50, 609-629. doi: 10.1177/0306624X06287644
24. McCallum, K. E., Boccaccini, M. T., & Bryson, C. N. (2017). The influence of risk assessment instrument scores on evaluators' risk opinions and containment recommendations. Criminal Justice and Behavior, 44, 1213-1235. doi: 10.1177/0093854817707232
25. Mercado, C, C., Jeglic, E., Markus, K., Hanson, R. K., & Levenson, J. (2011, January). Sex offender management, treatment, and civil commitment: An evidence based analysis aimed at reducing sexual violence. Research report submitted to the National Institute of Justice. Available at https://www.ncjrs.gov/pdffiles1/nij/grants/243551.pdf
26. Miller, H. A., Amenta, A. E., & Conroy, M. A. (2005). Sexually Violent Predator Evaluations: Empirical Evidence, Strategies for Professionals, and Research Directions. Law and Human Behavior, 29, 29-54.
27. Miller, C. S., Kimmonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. Psychological Assessment, 24, 944-953. doi: 10.1037/a0028411
28. Murrie, D. C., Boccaccini, M. T., Caperton, J., & Rufino, K. (2012). Field validity of the Psychopathy Checklist-Revised in sex offender risk assessment. Psychological Assessment, 24, 524-229. doi: 10.1037/a0026015

29. Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C. & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? Psychology, Public Policy, and Law, 15, 19-53. doi: 10.1037/a0014897
30. Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. Criminal Justice and Behavior, 41, 1406-1421. doi: 10.1177/0093854814548449
31. Perillo, A. D., Spada, A. H., Calkins, C., & Jeglic, E. (2014). Examining the scope of questionable diagnostic reliability in Sexually Violent Predator (SVP) evaluations. International Journal of Law and Psychiatry, 37, 190-197. doi: 10.1016.j.iljp.2013.11.005
32. Phenix, A., Fernandez, Y., Harris, A. J. R., Helmus, M., Hanson, R. K., & Thornton, D. (2016). STATIC-99R coding rules revised - 2016. Ottawa, Canada: Corrections, Directorate, Solicitor General Canada. Retrieved from www.static99.org/pdfdocs/Coding_manual_2016_v2.pdf
33. Rice, A. K., Boccaccini, M. T., Harris, P. B., & Hawes, S. W. (2014). Does field reliability for Static-99 scores decrease as scores increase? Psychological Assessment, 26, 1085-1094. doi.10.1037/pas0000009
34. Rutherford, M., Cacciola, J. S., Alterman, A. I., McKay, J. R., & Cook, T. G. (1999). The 2-year test-retest reliability of the Psychopathy Checklist-Revised in methadone patients. Assessment, 6, 285-291.
35. Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist-Revised among life sentenced prisoners in Sweden. Law and Human Behavior, 38, 315-324. doi: 10.1037/lhb0000063
36. Texas Health & Safety Code § 841.000-841.150 (2000).

Footnotes

[1]We calculated this 70% value from the larger sample of 898 committed and non-committed offenders evaluated between 1999 and 2011 (Harris et al., 2017). In these 898 cases, evaluators concluded that the offender had a behavioral abnormality in 548 cases, concluded that the offender did not have a behavioral abnormality in 231 cases, and made no statement about behavioral abnormality in 260 cases.
[2]We acknowledge that bivariate correlations for dichotomous variables (i.e., point-biserial, phi) become attenuated as the base rate for the variable departs from .50. We also ran a parallel set of AUC and odds-ratio analyses, but they led to the same substantive conclusions (results available from the corresponding author). We chose to present the bivariate correlations to facilitate interpretation across variables.
[3]The base rate for this combined category of paraphilia or antisocial personality disorder was 70.6% for initial evaluations and 74.5% for reevaluations.

Author address

Marcus T. Boccaccini
Department of Psychology and Philosophy
Sam Houston State University
Huntsville, TX 77341
Boccaccini@shsu.edu