Journal of Threat Assessment and Management
2014, Vol. 1, No. 2, 102–117

2169-4842/14/$12.00 http://dx.doi.org/10.1037/tam0000014

The Field Validity of Static-99/R Sex Offender Risk Assessment Tool in California

R. Karl Hanson
Public Safety Canada, Ottawa, Ontario, Canada

Alyson Lunetta
California Department of Justice, Sacramento, California

Amy Phenix
Morro Bay, California

Janet Neeley
California Attorney General's Office, Sacramento, California

Doug Epperson
College of Liberal Arts, California Polytechnic State University–San Luis Obispo

Policies that differentially apply to sexual offenders at different risk levels require defensible procedures for classifying offenders into risk categories. The current study examines the reliability and validity of Static-99 and Static-99R sexual offender risk assessment tools as implemented in the State of California. California is a valuable case study because it is a large jurisdiction that has devoted considerable resources to the implementation of risk tools. In Study 1, 55 corrections and probation officers scored Static-99R for 14 common cases. Overall, rater reliability was acceptable (intraclass correlation [ICC] = .78), with higher reliability found for experienced scorers (ICC = .85) than less experienced scorers (ICC = .71). In the second study, the predictive ability of Static-99 and Static-99R was examined in a prospective study of 475 randomly selected adult males released in 2006–2007 and followed for 5 years. Static-99/R scores were strongly related to subsequent sexual recidivism (areas under the curve $\geq$ .80). As well, there was acceptable fit between the expected and observed recidivism rates. These results suggest that it is possible to implement empirically derived risk assessment tools in large jurisdictions without degrading predictive accuracy.

Keywords: sexual offenders, risk assessment, Static-99R, field validity

Public concern about the risk of sexual victimization has resulted in specialized policies that apply uniquely to sexual offenders. In the United States, laws require sexual offenders to register with police, for states to post sexual offenders' identities on publicly available Web sites and for certain high-risk offenders to be civilly committed (Mancini, Barnes, & Mears, 2013). All of these measures are predicated on the assumption that all individuals with a history of sexual offenses are at increased risk for committing a new sexual offense. These mea-

R. Karl Hanson, Public Safety Canada, Ottawa, Ontario, Canada; Alyson Lunetta, California Department of Justice, Sacramento, California; Amy Phenix, Morro Bay, California; Janet Neeley, California Attorney General's Office, Sacramento, California; Doug Epperson, College of Liberal Arts, California Polytechnic State University–San Luis Obispo.

The views expressed are those of the authors and not necessarily those of Public Safety Canada or the California Attorney General. R. Karl Hanson is coauthor of the Static-99/R. Amy Phenix and R. Karl Hanson are certified Static-99R trainers. The copyright for Static-99/R is held by the Government of Canada and none of the authors receive royalties from this measure. We thank Kelly Babchishin and Leslie Helmus for assistance with the data analysis and all the volunteers who participated in the reliability study.

Correspondence concerning this article should be addressed to R. Karl Hanson, Corrections Research, Public Safety Canada, 340 Laurier Avenue West, Ottawa, Ontario, Canada, K1A 0P8. E-mail: karl.hanson@ps-sp.gc.ca

sures also assume that not all sexual offenders are equally likely to reoffend.

There is considerable variation in the methods used to differentiate offenders in terms of risk. Whereas comprehensive risk assessments are used in clinical practice and high-stakes evaluations (e.g., civil commitment, Jackson & Hess, 2007), such evaluations are infeasible when all convicted sexual offenders in a large jurisdiction must be triaged. Consequently, in many jurisdictions, sexual offenders are placed into tiers based on the criminal code sections for which they have been sentenced. Although such offense-based classifications are legally defensible and easy to administer, they have typically been developed in the absence of empirical evidence. Follow-up studies have found only weak associations between offense-based classification systems and the risk for sexual recidivism (Freeman & Sandler, 2010; Zgoba et al., 2012).

Static-99 is by far the commonly used empirical-actuarial risk scale for sexual offenders (Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010; Viljoen, McLachlan, & Vincent, 2010). It was created by aggregating criminal history and demographic characteristics that were shown in previous meta-analyses to be robust predictors of sexual recidivism (Hanson & Thornton, 2000). Static-99 is as accurate as any other sexual offender risk tool, demonstrating moderate ability to discriminate between recidivists and nonrecidivists (Hanson & Morton-Bourgon, 2009). As an empirically derived risk tool, it is periodically revised as new and better evidence becomes available. Recently, the age weights were revised (creating Static-99R) to account for particularly low recidivism rates among older offenders (>60; Helmus, Thornton, Hanson, & Babchishin, 2012).

Even though a particular risk scale works well in research studies, it may or may not work in any specific applied context. Implementation of any risk assessment procedure requires adequate information, trained and motivated staff, and procedures for maintaining fidelity with the coding rules. For example, a follow-up study of 1,928 sexual offenders screened for civil commitment in the State of Texas found that Static-99 scores showed only weak discrimination between recidivists and nonrecidivists (Boccaccini, Murrie, Caperton, & Hawes, 2009). Other Static-99 field validity studies, however, have found discrimination effect sizes as large as, or larger than, those reported in the original development studies (Rettenberger, Haubner-MacLean, & Eher, 2013; Storey, Watt, Jackson, & Hart, 2012). The reasons for the variation in predictive accuracy are not fully known.

In addition to potential effects of implementation and training, the ethnic composition of the sample may also influence predictive accuracy, with lower effects found for individuals of non-European heritage (Babchishin, Blais, & Helmus, 2012; Långström, 2004). Långström (2004) found that Static-99 worked well for Swedes and other Europeans in Sweden; however, it was not significantly related to recidivism in Sweden for individuals of non-European heritage (African, Asian). Babchishin et al. (2012) found that Static-99R, but not Static-2002R, worked equally well for Canadians of Aboriginal and non-Aboriginal heritage. The reasons for the reduced accuracy of Static-2002R among Canadian Aboriginal is difficult to explain, but does suggest that race could be a relevant factor in sexual offender risk assessment.

Even if the scale is coded accurately, however, the interpretation of the scores may vary across settings. Helmus, Hanson, Thornton, Babchishin, and Harris (2012) found that the 5 year recidivism rates for a Static-99R score of 5 could vary from 7% to 27% across settings. Although some of this variation should be attributed to random error, it is likely that that there is real variation in absolute recidivism rates across jurisdictions for reasons that are not fully known. Consequently, the validity of any particular implementation of a risk scale is worthy of empirical investigation.

Static-99/R in California

The current study examines the reliability and validity of Static-99 and Static-99R as implemented in the State of California. California was selected for research because it is a large jurisdiction that has mandated the use of Static-99 for offenders on their sexual offender registry. California has had a sex offender registry since 1947, which as of January 2013 contained over 93,000 registrants. California's use of Static-99 is also worth studying because those involved in the scale's implementation

were keenly aware of the challenges faced by introducing actuarial risk assessment, and that special care was required to ensure that those responsible for scoring Static-99 had the motivation and skills to do so.

In 2006, the use of evidence-based sex offender risk assessment in California was mandated by law (California Penal Code, § 290.03). Following this policy, a state committee oversees the selection of risk assessment instruments and training on those instruments. Known as the SARATSO Committee (State Authorized Risk Assessment Tool for Sex Offenders), members represent the Attorney General of California, California Department of Corrections and Rehabilitation (CDCR), Department of State Hospitals (DSH; formerly Department of Mental Health), and Chief Probation Officers of California (see www.saratso.org). SARATSO is mandated to select risk assessment instruments that "reflect the most reliable, objective and well-established protocols for predicting sex offender risk of recidivism, [be] scientifically validated and cross-validated, and [are], or [are] reasonably likely to be, widely accepted by the courts." (California Penal Code, § 290.04. eff. Sept 20, 2006). In 2007, SARATSO chose the Static-99 as the first official static risk assessment instrument for adult males, meaning that it must be scored for mandated cases.

CDCR began using Static-99 in 2005 (prior to the legal mandate), and probation departments began to score Static-99 in 2008. For both CDCR and probation, only new cases were scored; offenders whose only conviction was prior to those years do not have a Static-99 score. In 2009, SARATSO updated from Static-99 to Static-99R. By the fall of 2012, about 30,000 sex offenders in the California registry had a Static-99 or Static-99R score.

Evaluating the implementation of Static-99R is important for California because it is widely used for decision-making with sexual offenders. Along with other risk assessment tools, judges must consider Static-99R risk scores when sentencing a sex offender (California Penal Code, 1203c(a)(2)) and law enforcement may consider risk scores when deciding whether a registered sex offender poses a sufficient risk to justify community notification. For probation and parole, risk assessment scores determine whether a registered sex offender is supervised as high risk. Parolee sex offenders are required by law to wear a GPS monitoring device regardless of risk level, but sex offenders on probation must wear a GPS monitoring device if they score high risk on Static-99R. All sexual offenders on community supervision must participate in treatment; however, risk assessment is used to determine the level and frequency of treatment. Starting in 2013, California law also requires that Static-99R scores be posted online for those offenders on the public Megan's Law Web site.

The implementation of Static-99 and Static-99R has always been directed by the SARATSO committee. Dr. Amy Phenix, one of the coauthors of the coding rules for the Static-99, helped craft the initial legislation establishing the use of risk assessment tools in California and remains an advisor to the committee.

Only persons who have taken the official SARATSO training are authorized to score Static-99R for official purposes. The official training is provided by either a certified Static-99/R trainer (see www.static99.org) or by qualified local trainers (called "supertrainers"). To qualify as a SARATSO supertrainer, it is necessary to attend a train-the-trainers workshop conducted by a certified Static-99R trainer and have at least 2 years' experience scoring the Static-99R. A uniform curriculum is used to train new scorers based on a standardized PowerPoint presentation. The training materials are periodically updated by a certified Static-99R trainer engaged by SARATSO. The Chief Probation Officers of California collaborate with SARATSO to provide training on Static-99/R to probation officers.

All users and supertrainers must demonstrate their competence by accurately scoring standard cases. New users are encouraged to work with a mentor, and, whenever possible, the supertrainers review their first 10–20 cases. Scorers are required, by law, to be retrained every 2 years.

Overview of Current Study

This article describes two studies examining the reliability and validity of Static-99 and Static-99R as implemented in the State of California. The first study focuses on rater reliability. In this study, 55 corrections and probation officers scored Static-99R for 14 common cases. This study provided information on how well practitioners understood the coding rules, and on possible sources of variations in rater reli-

ability. In the second study, the predictive ability of Static-99 and Static-99R was examined in a prospective study of 475 randomly selected adult males released in 2006–2007 and followed for 5 years. Predictive accuracy was quantified by both discrimination (the extent to which recidivists were different than nonrecidivists) and calibration (correspondence between expected and observed recidivism rates).

Study 1

Method

Participants. Fifty-seven (55 with usable data) state and county government employees volunteered to participate in an interrater reliability study in response to an announcement advertising the study. Thirty-one of the volunteers were parole officers from CDCR, and all had previously completed Static-99R training from a certified Static-99R trainer within the previous 2 years. Twenty-six of the volunteers were from California county probation offices, including 22 probation officers and 4 clinicians who worked in probation outpatient clinics. All volunteers from probation had previously completed Static-99R training within the past 2 years. The vast majority were trained by a certified Static-99 trainer, and the remainder from a probation master trainer, who had attended a train-the-trainer workshop provided by a certified trainer.

Although all participants had completed appropriate training, experience scoring Static-99R varied widely. The number of cases scored in the past year ranged from 0 to over 1,000, with a median of 40.

Measures. Static-99R (Hanson & Thornton, 2000; Helmus, Thornton et al., 2012) is an empirically derived actuarial risk assessment tool designed to predict sexual recidivism in adult male sex offenders based on commonly available demographic and criminal history information (see also www.static99.org). It has 10 items and the total score (ranging from −3 to 12) can be used to place offenders in one of four risk categories: low (−3 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6+). The Static-99R items are identical to Static-99, with the exception of updated age weights.

Materials. Fourteen redacted records from real cases were used for this reliability study. The 14 cases were selected quasi-randomly from CDCR files, with the provision that the resulting files represented a range of risk levels. There was no intent to select cases for which the scoring of any particular item was easy or difficult; instead, the goal was to select cases that were representative of the cases encountered in routine practice. Each case contained a record of chronological history while in the CDCR, the criminal legal records including charging and conviction documents, the Probation Officer's Report containing offense facts for the sexual offense, related Police Reports, and the official criminal history. Each case file averaged approximately 40 pages.

Procedure. Participants reported to the study site in Sacramento, CA, at 8:00 a.m. on the day of the study (June 2011). Each participant was given copies of the 14 case files and 14 score recording sheets and was expected to score all 14 cases by 5:00 p.m. the same day. Participants had access to a Static-99R scoring manual (Harris, Phenix, Hanson, & Thornton, 2003), as well as any notes that they routinely used. They were not permitted to talk or consult with each other at any time during the study. A certified Static-99R trainer was in attendance the entire day and scored the same 14 cases to provide a standard of comparison.

Because the study was done during normal work hours, the volunteers were able to claim the time as work. Those that had to travel a significant distance were also compensated for their travel expenses.

Subsequent to the study, participants were sent their own scoring sheets and the scores from the certified trainer for each case as feedback. Scoring issues identified in this study were used to update and enhance training in California.

Analyses. Interrater reliability for Static-99R total scores was calculated using the intraclass correlation coefficient (ICC) for absolute agreement using a two-way, random effects model. The value of the ICC for absolute agreement can be conceptualized as the proportion of variance in total scores because of true differences in offenders (case files), as opposed to variance because of evaluator differences, or random error. Absolute ICC values range from −1 to 1,

with positive values greater than .75 indicating excellent agreement (.60 to .75 = good; .40 to .59 = fair; <.40 = poor; Cicchetti, 1994).

Results

Two of the initial 57 participants abandoned the exercise halfway through. Neither of these participants had prior experience scoring real cases (a selection criteria), and they only realized that they did not qualify for the study after arrival at the testing site. Consequently, analyses were performed on the data from 55 participants. Overall, rater reliability for total scores was excellent (ICC = .78, 95% confidence interval [CI] = .64–.90). The Pearson correlations between the "correct" scores provided by the certified trainer and the average scores provided by the raters were also high (.90–.97; Table 1).

Generally, the dispersion of total scores was small (median *SD* = .82); however, some cases were harder to score than others, with Cases 2, 9, 11, and 14 being the most challenging for these participants (*SDs* ranging from 1.36 to 1.59; percent correct of <20%). On average, about half the scorers gave the exactly correct score, but most of the raters were within ±1 point of the correct score for most of the cases (median of 88.1%; Table 1).

To examine the difficulty of specific Static-99R items, Table 2 presents summary statistics for the percentage of the 55 officers who scored the correct response to each of the 10 Static-99R items (aggregated across the 14 cases). The average percent agreement ranged from a low of 76.8% for prior sex offenses to a high of 95.8% for noncontact sex offenses. Certain items within certain cases, however, were particularly difficult. For example, only one of the 55 raters obtained the correct score for prior sex offenses for Case 11. In contrast, all the raters obtained the correct score for 34 of the 140 possible item/case combinations. The highest rater reliability was obtained for noncontact sexual offenses item, where perfect agreement was obtained for 8 out of the 14 cases.

Rater reliability was strong for both parole officers (ICC = .81, n = 30) and for probation officers (ICC = .77; n = 25). There was a meaningful difference, however, between the experienced coders (ICC = .85, n = 33) and the less experienced coders (ICC = .71, n = 22). Experienced was defined as having coded 26 or more real cases prior to participating in the study. In general, parole officers were more experienced (83%, 25/30) than were the proba-

Table 1
Static-99R Total Score Descriptive Statistics

Case	Correct score	Evaluators' scores			*SD*	% Correct	% ±1 Correct
		Median	Mode	Mean			
1	1	1	1	1.20	0.70	72.7	94.5
2	7	4	6	4.38	1.58	1.8	30.9
3	3	3	3	3.20	0.56	76.4	96.4
4	2	2	2	2.29	0.81	70.9	92.8
5	4	4	4	3.91	0.59	70.9	98.2
6	3	3	3	3.04	0.51	80.0	98.2
7	3	3	3	2.62	0.73	58.2	94.6
8	6	6	6	5.93	1.02	52.7	89.0
9	7	5	6	5.13	1.59	12.7	49.1
10	4	5	5	4.78	0.98	32.7	80.0
11	6	7	7	6.93	1.36	9.1	63.7
12	6	6	6	5.22	1.08	49.1	58.2
13	6	6	6	5.95	0.83	70.9	87.2
14	10	9	9	8.53	1.48	18.2	61.8
					Median	55.5	88.1
					Mean	48.3	78.2

Note. $r_{(\text{correct, median})} = .90$, $r_{(\text{correct, mode})} = .97$, $r_{(\text{correct, mean})} = .91$.

Table 2
Item Agreement (% Correct) for Static-99R Items Averaged Across 14 Test Cases Scored by 55 Raters

Item	% Correct		
	Median	Mean	Range
1. Age at release	95.5	88.8	49.1–100
2. Ever lived with a lover for 2 years	93.2	87.5	47.3–98.2
3. Index nonsexual violence	94.6	86.4	36.4–100
4. Prior nonsexual violence	91.8	78.0	34.5–98.2
5. Prior sex offences	93.6	76.8	1.8–100
6. Prior sentencing dates	96.4	87.4	29.1–100
7. Noncontact sex offences	100.0	95.8	80.0–100
8. Any unrelated victims	97.3	94.0	60.0–100
9. Any stranger victims	97.2	93.9	59.3–100
10. Any male victims	98.2	94.4	47.3–100

tion officers, 32%, 8/25; $\chi^2(1) = 14.97$, $p < .001$.

Discussion

This study found good to excellent rater reliability in a sample of 55 California parole and probation officers (overall ICC = .78). The overall level of agreement was similar to that observed in other field reliability studies (Boccaccini et al., 2012; Miller, Kimonis, Otto, Kline, & Wasserman, 2012) and provides some support for the current training program implemented in this state. There were, however, meaningful differences between the raters based on scoring experience, with more experienced scorers (26+ cases) being more accurate (ICC = .85) than the less experienced scorers (ICC = .71; 25 or less cases). These findings suggest that extra supervision during the scoring of initial cases could be useful to ensure acceptable accuracy. The precise number of cases required to obtain proficiency in Static-99R scoring could not be established from the current study and would obviously depend on the complexity of the case files and the abilities of the scorers. Nevertheless, a threshold of 20–25 cases probably provides a reasonable balance between the costs of increased supervision and the benefits of increased reliability.

Currently, the CDCR parole division requires that each coder be mentored for his or her first 25 cases. Thereafter, coders are audited on 10%–50% of their cases by super trainers (i.e., individuals who have attended a train-the-trainers workshop conducted by a certified Static-99R trainer and have at least 2 years' experience scoring the Static-99R). This audit procedure provides a check on coder drift and assists in identifying coding problems on an ongoing basis. As an additional check on system-wide drift, all probation and parole officers are required to participate in a Static-99R training every 2 years.

Analysis of the errors in the four most challenging cases in the current study found that all of the large differences in total scores could be attributed to problems identifying the index offense (e.g., pseudorecidivism). Consequently, those responsible for implementation of Static-99R should ensure that the rules concerning prior and index offenses are communicated clearly to trainees, and reenforced in the refresher trainings. Given that the Static-99R age item is more complex than the Static-99 age item (4 categories compared with 2), it is worth noting that the raters had no particular difficulty scoring age at release. The number of errors in scoring Static-99R age item was similar to number of errors found for the other items (Table 2).

Although the overall rater reliability was acceptable, it was lower than the rater reliability typically found in research studies. McGrath, Lasher, and Cumming (2012), for example, found an ICC of .89, and Looman and Abracen (2013) reported $r = .90$ for Static-99R total scores based on file review. In the typical research study, a small number of individuals score a large number of cases. Consequently, research coders would all be considered "experienced" by the definition used in the current study (having coded 26 or more cases).

It is important for users of any assessment instrument to remember that rater reliability is never perfect (Boccaccini et al., 2012). Using the rater reliability in the current study (ICC = .78), and the standard deviation from published norms ($SD = 2.3$; Hanson, Lloyd, Helmus, & Thornton, 2012), the standard error of measurement (*SEM*) would be estimated at $1.08 = 2.3 (1 - .78)^{1/2}$ using the formula: $SEM = SD (1 - r_{xx})^{1/2}$ (Ley, 1972). This means that 68% of the time, the offender's "true score" would be within 1 point of his observed score (±1.08), and within 2 points of his observed score 95% of the time. For experienced raters, the *SEM* would be only marginally lower: $0.89 = (2.3[1 -$

.85]$^{1/2}$). In evaluations where single-point differences could have consequences, it would be possible to increase the reliability by using multiple raters. For example, rater reliability based on the average of two experienced raters would be .92, using the Spearman-Brown formula: ($k * r_{xx}$)/(1 + [k – 1]r_{xx}), where k is the number of raters (Ley, 1972) and r_{xx} = .85. The use of multiple raters and other checks on rater reliability would be particularly important when bright-line thresholds for risk scale scores are enshrined in policy or law (e.g., Joint Legislative Audit and Review Commission, 2011).

Study 2

Overview

Study 2 examined the predictive accuracy of Static-99 and Static-99R scores in a sample of 475 sexual offenders released from the California Department of Corrections and Rehabilitation (CDCR) in 2006–2007. This sample was selected because they were the first cohort who was routinely scored on Static-99 and who had been at risk long enough to evaluate predictive accuracy (5 years). Consistent with standard practice for prognostic tools, we examined both discrimination (the extent to which recidivists differ from nonrecidivists) and calibration (fit between expected and observed recidivism rates; Moons, Royston, Vergouwe, Grobbee, & Altman, 2009). The expected values were based on the norms provided for Static-99R for routine correctional samples (Phenix, Helmus, & Hanson, 2012).

Method

Sample. The sample consisted of adult, male sexual offenders released from CDCR, who had been convicted of a sexually motivated offense against an identifiable victim (Category A offenses; Harris et al., 2003). As per the Static-99 scoring rules, offenders were excluded if their only registerable sex offense was for child pornography (Cal. Pen. Code, §§ 311.1–311.11), statutory rape (§ 261.5, consensual intercourse with a minor ages 14–17), or a Hofsheier offense (other consensual sexual activity with a minor ages 14–17, e.g., §§ 286(b)(1) & (2); 288a(b)(1) & (2); 289(h) & (i)).

The initial sample comprised every fourth offender from those scored on Static-99 between June 2006 and June 2007. Of the initial 541 cases (from a total available sample of 2,164), 66 were eliminated because they had less than 5 years between release date and date of follow-up (March 2012): 14 were released after March 15, 2007; for 7 cases, the release date was unknown; 23 were lost to follow-up prior to 2012; and 22 died prior to 5 years after release (4 nonviolent recidivists, 1 sexual recidivist).

Of the remaining 475 offenders, 42.1% (n = 200) were Hispanic, 29.5% (n = 140) were White, 20.8% (n = 99) were Black, 4.4% (n = 21) were "other," and for 3.2% (n = 15) race was unknown. On average, the offenders were 41.9 years at release (SD = 10.7, range = 20.7–86.6). The most common index offense convictions were for lewd and lascivious acts against children (54.1%; Cal. Pen. Code § 288), rape/sexual battery (20.2%; § 243.4, § 261) and exhibitionism (6.9%,§ 314).

Measures. Static-99 (Hanson & Thornton, 2000) is a 10-item empirical actuarial risk tools designed to predict sexual recidivism among adult male offenders. Total scores range from 0 with 12, with the following risk category labels: Low (0,1), Low-Moderate (2,3), Moderate-High (4,5) and High (6+). Normative information is available at www.static99.org.

Static-99R (Helmus, Thornton et al., 2012) is identical to Static-99 with the exception of revised age weights (see Study 1).

Procedure. Offenders were scored on Static-99 by CDCR staff as part of routine practice. Although the exact number of individuals providing this set of scores was not recorded, CDCR had approximately 75 scorers on staff in 2007. During 2006–2007, CDCR policy required that all released sexual offenders were scored on Static-99, and the scoring prioritized cases nearing release. Static-99R was computed retrospectively based on Static-99 total scores and age at release. Rater reliability for this specific set of scores was not available.

Recidivism information was provided by the California Department of Justice as of March 2012. Recidivism was defined as an arrest for a sexual offense. Sexual offenses were divided into contact and noncontact based on offense descriptions in California's Penal Code. Violations of California's regis-

tration law were not included as sexual recidivism but were counted among the nonsexual offenses.

Plan of analysis.

Discrimination. The ability of the risk scales to discriminate between recidivists and nonrecidivists was indexed using odds ratios from logistic regression (e.g., Hosmer & Lemeshow, 2000), and by the area under the curve (AUC) from receiver operating characteristic (ROC) analysis (e.g., Rice & Harris, 1995). Odds ratios indicate the change in relative risk associated with a 1-unit change in Static-99/R scores. AUC values indicate the probability that a randomly selected recidivist would have a more deviant score than a randomly selected nonrecidivist. AUC values are expected to be smaller in prognostic studies than in diagnostic studies because the outcome of interest in prognostic studies does not exist at the time of assessment and may never happen (Royston, Moons, Altman, & Vergouwe, 2009). According to Rice and Harris (2005), AUCs in violence prediction studies of .56 would be considered small, .64 would be moderate, and .71 would be large.

Calibration. The fit between the expected and observed recidivism rates was examined using intercepts from logistic regression, and the *E/O* index (expected number of recidivists divided by observed number of recidivists). Calibration analyses were only computed for Static-99R (not Static-99) because recidivism rate norms for Static-99 are not available (the original Static-99 norms from Hanson and Thornton, 2000, are no longer supported by the test developers).

The *E/O* index is the expected number of recidivists divided by observed number of recidivists. Perfect calibration is indicated by an *E/O* index of 1.0. Following Rockhill, Byrne, Rosner, Louie, and Colditz (2003), the 95% CIs for the *E/O* indices were computed as follows:

$$95\%\ \mathrm{CI}(E/O) = (E/O)\exp\left(\pm 1.96\sqrt{1/O}\right)$$

The expected number of recidivists was based on the 5-year sexual recidivism rates for routine samples reported by Phenix et al. (2012).

A second method of testing calibration was to examine the extent to which logistic regression intercept values (centered on Static-99R scores of 2) differed from those previous reported. Because the logistic regression parameters for the norms were not publically available, the parameters used for comparison were the logistic regression parameters for other routine samples reported by Hanson, Babchishin, Helmus, and Thornton (2013; Appendix Table B1: $B0_2 = -2.941$, $SE = 0.238$; $B1 = 0.331$, $SE = 0.044$). Specifically, the $B0_2$ represents the expected recidivism rate for a Static-99R score of 2 (p_2) in logit units ($\ln[p_2/\{1 - p_2\}]$). Differences between the parameters in the current sample and those of previous studies was tested using fixed-effect meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hanson & Broom, 2005).

Results

Overall, 48.4% (230/475) of offenders were charged with any offense, of which 4.8% (23/475) were charged with a sexual offense during the fixed 5-year follow-up period. Of the 23 sexual recidivists, 15 were charged with any contact sexual offenses and 8 were charged with only noncontact sexual offenses.

The average Static-99 score was 2.6 (median = 2, $SD = 2.1$, range = 0–10). Using the fixed 5-year follow-up, the AUC with any sexual recidivism was .824 (95% CI = .724–.923) and .809 (95% CI = .681–.937) for contact sexual recidivism. The average Static-99R score was 2.2 (median = 2, $SD = 2.2$, range = −3 to 11). The AUC with any sexual recidivism was .817 (95% CI = .716–.919) and .798 (95% CI = .664–.932) for contact sexual recidivism.

The relationship between Static-99R scores (centered on a score of 2) and sexual recidivism acceptably fit a logistic distribution (Hosmer-Lemeshow test was nonsignificant: $\chi^2 = 7.22$, $df = 5$, $p = .205$; $B0_2 = -3.778$, $SE = 0.341$; $B1 = 0.548$, $SE = 0.0970$). The resulting logistic equation indicated a relative risk increase of 1.73 for each increase in Static-99R score ($e^{0.548} = 1.73$), and an adjusted 5-year sexual recidivism rate of 2.2% for a Static-99R score of 2 ($[1/\{1+e^{-(-3.778)}\}] = 0.022$).

When compared with other routine samples (Hanson et al., 2013), the adjusted (score of 2) base rate was significantly lower ($B0_2$ of −3.778 vs. −2.941; $Q_{between} = 4.05$, $df = 1$, $p = .044$), and the discrimination (change in relative

risk) was significantly larger (*B1* of 0.548 vs. 0.331; $Q_{between} = 4.15$, $df = 1$, $p = .042$).

In comparison to norms for routine samples, the observed 5-year overall recidivism rate in the current sample was slightly lower (4.84% vs. 6.31%), although the difference was not significant (*E/0* = 1.30, 95% CI = 0.87–1.96; Table 3). Figure 1 provides a plot of the observed recidivism rates per Static-99R risk score, the rates based on the smoothed logistic curve fitted to this data, and the recidivism rate norms for routine samples (Phenix et al., 2012). The values are plotted up to a score of 9, because there were no offenders with a score of 10, and only one with a score of 11 (a sexual recidivist). As can be seen in Figure 1, the general pattern is that the recidivism rates in the current California sample were lower than expected for scores below 5 and higher than expected for scores of 6 or more. When comparing each of the four Static-99R risk categories (Table 3), only the Low-Moderate category (scores of 2 or 3) was significantly different from the expected values (1.2% vs. 5.6%; *E/O* = 4.58, 95% CI = 1.15–18.31).

Across racial groups, there were significant differences in the average Static-99, $F(3, 471) = 9.17$, $p < .001$, and Static-99R scores, $F(3, 471) = 4.40$, $p = .005$. As can be seen from Table 4, Blacks scored highest and Hispanics scored lowest on both measures. Race, however, did not have a significant effect on predictive accuracy. The AUC values were large (>.74) for all groups, although not all were statistically significant because of the small number of recidivists in some groups. In these analyses, offenders of other races or for whom race was unknown were grouped into a single category ($n = 36$).

The logistic regression equations for Static-99R predicting 5-year sexual recidivism rates are presented in Table 5 for Hispanic, White, and Black offenders. The "other/unknown" racial group was not included in the logistic regression analysis because the model was misspecified (only 1 sexual recidivist). Across these three racial groups, there were no significant differences in either the adjusted base rate ($B0_2$ is the predicted value for a Static-99R score of 2) or in the rate of change in relative risk for a one unit increase in Static-99R score (*B1* is the log odds ratio for adjacent scores). For both parameters, the variability across racial groups was less than would be expected by chance ($Q < df$, for both $B0_2$ and *B1*).

Discussion

This prospective study found high predictive accuracy of Static-99 and Static-99R among sexual offenders scored by the California Department of Corrections and Rehabilitation. Specifically, both instruments discriminated between recidivists and nonrecidivists with AUC and relative risk indicators larger than typically reported for these tools (for comparisons, see meta-analyses by Babchishin, Hanson, & Helmus, 2012 and Hanson et al., 2013). The overall sexual recidivism base rate was low (4.8% after 5 years) and significantly lower than the norms for the Low-Moderate risk group (Static-99R scores of 2 and 3). Although there were some differences in total risk scores for Hispanic, Black, and White offenders, there were no significant differences in the predictive ability of Static-99R for these three racial groups.

Table 3
Comparison of Expected and Observed Sexual Recidivism Rates for Static-99R Risk Categories

Risk category	Sample size	Recidivism rates (%)		E/O index	95% confidence interval	
		Observed	Expected			
Low	188	1.6	3.0	1.86	0.60	5.76
Low-moderate	163	1.2	5.6	4.58	1.15	18.31
Moderate-high	90	8.9	9.7	1.09	0.54	2.17
High	34	29.4	19.2	0.66	0.35	1.22
Total	475	4.84	6.31	1.30	0.87	1.96

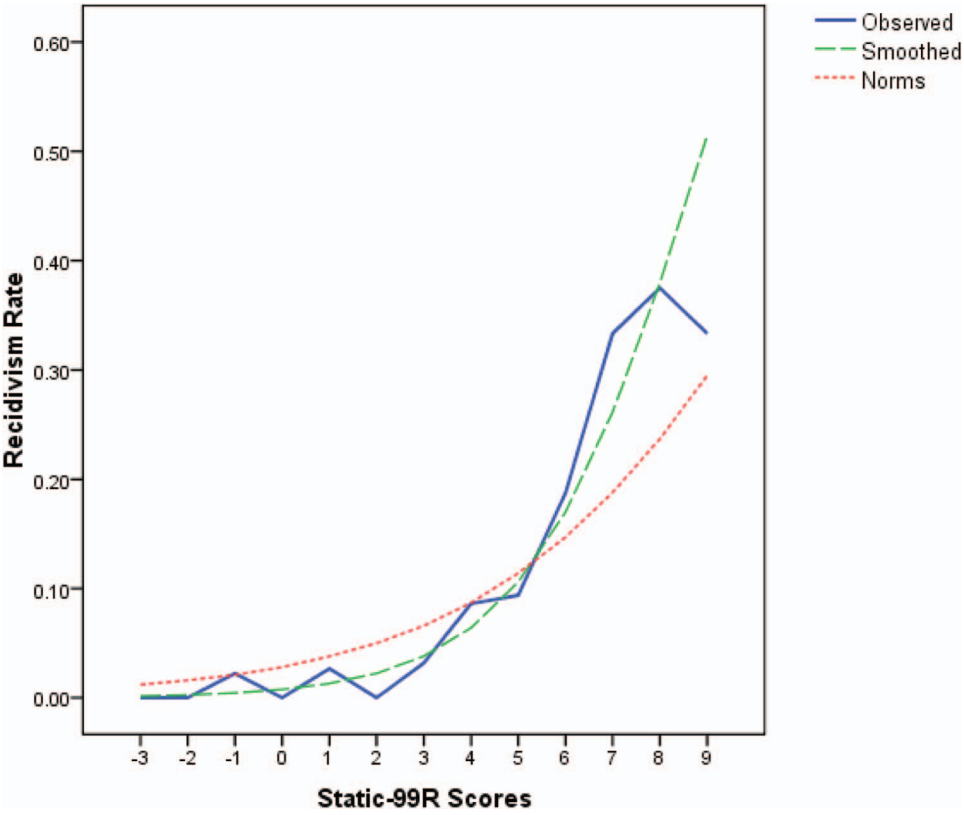

Figure 1. Observed and expected recidivism rates based on Static-99R 5-year sexual recidivism rates for routine samples. The color version of this figure appears in the online article only.

The current findings suggest that it is possible to implement empirically derived risk tools in applied settings with no appreciable degradation in predictive ability. Although the reasons for the relatively high AUCs are not fully known, it is likely that California's careful approach to implementing and monitoring Static-99 scoring was a contributing factor.

The relatively low recidivism rates in this sample were consistent with the low sexual recidivism rates in recent U.S. cohorts (Minnesota Department of Corrections, 2007; Zgoba et

Table 4
Five-Year Sexual Recidivism Rates and Static-99 and Static-99R Scores for Hispanic, White, and Black Offenders

Racial group	Recidivism rate (%)	Number of recidivists/total	*M* (*SD*)		AUC (95% confidence interval)					
			Static-99	Static-99R	Static-99	Lower	Upper	Static-99R	Lower	Upper
Hispanic	2.5	5/200	2.0 (1.8)	1.8 (2.2)	**.748**	.402	1.000	**.734**	.406	1.000
White	7.1	10/140	2.9 (2.3)	2.3 (2.4)	**.856**	.718	.994	**.850**	.722	.979
Black	7.1	7/99	3.2 (2.1)	2.7 (2.1)	**.747**	.548	.945	**.765**	.561	.968
Other/Unknown	2.8	1/36	2.2 (2.0)	1.8 (2.0)	**.957**	.000	1.000	**.986**	.000	1.000
Total	4.8	23/475	2.6 (2.1)	2.2 (2.2)	**.824**	.724	.923	**.817**	.716	.919

Note. AUC = area under the curve.

Table 5
Logistic Regression Parameters for Static-99R Predicting 5-Year Sexual Recidivism for Hispanic, White, and Black Offenders

	Base rate (Static-99R score of 2 in logit units)		Relative risk (log odds)			
Racial group	$B0_2$	*SD*	*B1*	*SD*	χ^2	p (df = 1)
Hispanic	−3.953	0.566	0.383	0.195	3.86	.049
White	−3.638	0.592	0.611	0.167	19.25	<.001
Black	−3.410	0.619	0.484	0.172	8.37	.004
Average (fixed-effect)	−3.683	0.341	0.505	0.104	23.51	<.001
Q (df = 2)	0.428, p = .81		0.808, p = .67			
I^2	0.00		0.00			

al., 2012). These low rates mean that a substantial proportion of released sexual offenders would be expected to have sexual recidivism rates that are not appreciably different from the rate of "out of the blue" sexual offenses by offenders with no prior convictions for sexual offenses. Langan, Schmitt, and Durose (2003) found that 1.3% of nonsexual offenders were rearrested for a sexual offense after 3 years of follow-up. In comparison, the expected sexual recidivism rate after 5 years for a Static-99R score of 1 was 1.3%, meaning that 39.6% of this sample (188/475) would have expected recidivism rates indistinguishable from those of nonsexual offenders. The extent to which California's policies concerning sexual offender treatment and management contribute to this low recidivism rate is unknown.

Although we found no meaningful differences in accuracy across racial groups, the current study was underpowered to detect such differences (10 or fewer recidivists in each group). Previous research in Sweden found that Static-99 worked poorly for offenders of non-European background (Långström, 2004), and there is some evidence that actuarial risk tools work less well for Canadian-Aboriginals than other Canadian sex offenders (Babchishin et al., 2012). Further research is needed before strong conclusions can be made concerning the predictors of sexual recidivism for sex offenders of substantially different cultural and racial backgrounds.

Summary and Concluding Discussion

Policies that differentially apply to sexual offenders at different risk levels require defensible procedures for classifying offenders into risk categories. The current study demonstrated that a risk assessment procedure that has demonstrated moderate predictive accuracy in research studies (Static-99/R) can be implemented in a large jurisdiction (California) without meaningful degradation in predictive accuracy. The individuals responsible for scoring Static-99R in California had sufficient training to adequately apply the scoring rules with acceptable rater reliability. The scores used in routine practice were strongly related to subsequent sexual recidivism, with discrimination indices higher than average. Overall, there was acceptable fit between the expected and observed recidivism rates, with the only exception being that the sexual recidivism rates were lower than expected for the Low-Moderate risk category.

The field validity of correctional practices cannot be assumed. The literature is replete with examples of assessment and intervention practices that worked less well in routine practice than in the carefully controlled research projects in which they were developed (Andrews, 2006; Flores, Lowenkamp, Holsinger & Latessa, 2006; Rhine, Mawhorr & Parks, 2006). This problem was well known to those responsible for implementing Static-99/R in California, and the current results provide some evidence that their efforts to promote high-fidelity implementation were successful. These efforts included obtaining explicit legal, political, and management support for the use of this risk tool, implementing a standardized system of training by qualified trainers, and requiring users be retrained every 2 years.

Although the Static-99/R scales worked as well (or better) than expected, decision-makers still need to consider whether it works well enough for any particular decision. The current results support previous findings (Hanson & Morton-Bourgon, 2009) that Static-99 is an efficient method of sorting offenders into risk categories. In the current sample, the two lowest risk categories, comprising 74% of the sample, had observed recidivism rates of less than 2% after 5 years, whereas the two higher risk categories had rates that were substantially higher (9% and 29%). Such differences are large enough to be meaningful, but demonstrate far from perfect ability to discriminate recidivists from nonrecidivists.

One challenge for applied users of Static-99R, however, is that it does not measure all risk relevant characteristics. Research studies have demonstrated that it is possible to improve the statistical prediction of sexual recidivism by considering additional information drawn from interviews, behavioral history, and psychological tests (e.g., Olver, Beggs Christofferson, Grace, & Wong, 2013; Thornton, 2002; Thornton & Knight, 2013). Evaluators are then left with the familiar problem of knowing that certain information is relevant to the risk assessment, but lacking an empirically validated method of including it in the overall evaluation of risk (Hanson, 1998).

A related problem is that there is considerable variation in the absolute recidivism rates associated with specific Static-99R scores across settings and samples (Helmus, Hanson et al., 2012). The developers of Static-99/R have responded to this variation by clustering samples into groups with different recidivism base rates (Phenix et al., 2012). These groups include norms for routine samples; that is, those roughly representative of the complete population of sexual offenders, as well as norms for sexual offenders who have been preselected to be high risk based on characteristics external to Static-99R. Consequently, evaluators interested in absolute recidivism rates must decide which normative group is most applicable to the case at hand.

For system-wide implementation of Static-99R risk assessment, administrator can set the normative group through central policy. For example, California's policy is to exclusively use the Static-99R routine norms for the purposes of probation and parole. For high stakes evaluations, such as civil commitment, a case-specific justification of the appropriate normative group is required. One approach to determining the normative group is by considering the density of external risk factors as measured by a structured risk assessment tool (Thornton, Hanson, & Helmus, 2010). The SARATSO committee, for example, has mandated that an external measure of dynamic risk factors be used along with Static-99R for sexual offenders in the community. Originally, the SARATSO committee mandated the use of the Structured Risk Assessment–Forensic Version (Thornton & Knight, 2013) and, more recently, STABLE-2007 (Hanson, Harris, Scott, & Helmus, 2007). The validity of using either of these measures to select Static-99R normative groups has yet to be evaluated in any applied setting.

The current results raise questions about which recidivism rates should be applied to Static-99R scores for routine sex offenders in California: the existing Static-99R routine norms, or the recidivism rates in the current study. In general, confidence in norms should increase with (a) the similarity between the offenders in the normative samples and the individual offender being assessed, (b) the consistency of previous research results, and (c) the number of offenders upon which the norms are based. Although there are no absolute rules for determining the sample size required for stable recidivism rate norms, a sample of 100 recidivists is a useful heuristic (Vergouwe, Steyerberg, Eijkemans, & Habbema, 2005). The 2012 Static-99R 5-year recidivism rate norms for routine samples are based on 8 samples, with an aggregated sample of 2,406, including 145 recidivists (Helmus, Hanson, & Babchishin, 2009). Given that there were only 23 sexual recidivists in the current study, creating stable local norms would require aggregating the current findings with other data, either from subsequent California cohorts, or from other comparable jurisdictions.

Although there were some differences, the current recidivism rates were largely consistent with the existing Static-99R norms for routine samples. The only statistically significant difference was for the Low-Moderate group, and the magnitude of the difference would have little practical significance (5.6% in norms vs. 1.2% in the current study). The recidivism rate for the

high risk group in the current sample was higher than expected (29.4% vs. 19.2%), but there were only 34 individuals in the high risk group (10 recidivists) and the E/0 index was not significant (Table 3). Consequently, it is reasonable for California to continue using the existing Static-99R routine norms where they are currently being used, or to create updated, aggregated norms that include the current results.

It is important to remember that this study did not address the question of selecting nonroutine normative groups for Static-99R recidivism estimates. The current study examined a routine sample and found recidivism rates comparable to routine samples in other jurisdictions. It is quite possible that identifiable subgroups of offenders in this cohort may be riskier (or less risky) than the overall average based on factors not measured in this study. The extent to which evaluators can reliably detect such variation is an important topic for further research.

Limitations

Although the initial sample size was close to 500 individuals, the low sexual recidivism rate (<5%, $n = 23$ recidivists) limited statistical power. Additional research using larger samples is needed to further explore the need for local norms, and to explore potential variability across ethnic groups.

The available data did not identify periods when the offender was not at risk during the follow-up time because of incarceration or hospitalization. Furthermore, there is a time lag between the date at which a sexual crime is detected by police and the date when records of the crime became available for research purposes. Both of these factors would contribute to underestimating the observed recidivism rates in the current study.

Another potential limitation is that several of the authors of the current report have been deeply involved in the development of Static-99 and its implementation in California in particular. Although it is impossible to evaluate potential bias in the hundreds of small decisions required in a research study, we aspired to objectivity throughout. Some confidence in our findings is provided, however, by the consistency of our results with those of other, independent research teams (Rettenberger et al., 2013; Storey et al., 2012).

Another potential limitation is that the practices examined in the current study may not be representative of current practice. It is likely that the state employees who volunteered for the reliability study were not a representative sample, with an overrepresentation by those interested in risk assessment and confident in their scoring abilities. As well, the risk assessments used in Study 2 were drawn from the first cohort routinely assessed by Static-99. Consequently, there is a risk that subsequent (i.e., current) scoring has been degraded by rater drift and decreased vigilance.

Risk Assessment and Sexual Offender Registries

Despite its adoption of risk assessment in 2006, California still does not have a tiered registration system that explicitly takes risk of reoffense into account. Forty-six states require offenders to register for a defined time period (typically 10 years for low-risk offenders, 20–25 years for moderate-risk offenders, and life for high-risk offenders) based, in part, on somebody's assessment of their risk of reoffense.

Unfortunately, despite embracing evidence-based practices in utilizing risk assessment, California still uses an offense-based system in classifying sex offenders for registration. Only the type of sex offense conviction governs whether a sex offender is posted on the public Megan's Law Web site (www.meganslaw.doj.ca.gov), and the extent of information disclosed. Neither the Static-99R score nor other types of risk assessment scores are a factor in whether an offender is disclosed on the Web site, or registration requirements. Currently, the tools are in place for California to move forward with a registration law that does not treat all sex offenders the same. If risk is going to be a major contributor to registration policies, then the time offense-free in the community also needs to be considered along with their risk at time of release. As with other types of crimes (Blumstein & Nakamura, 2009; Bushway, Nieubeerta, & Blokland, 2011), the longer sexual offenders have been offense-free, the lower their expected recidivism rate in the future (Harris & Hanson, 2004).

Conclusions

Our desire to protect ourselves from interpersonal threats has led to the development of a number of policies uniquely applicable to sexual offenders. There is considerable variation, however, in the likelihood among individuals convicted of a sexual offense that they will go on to commit another sexual offense. Previous research has identified a number of factors reliably related to recidivism risk, and these factors have been combined into structured risk tools. The current study demonstrates that one of these empirically derived risk tools (Static-99/R) can be implemented with adequate fidelity in a large jurisdiction (California). The existence of such validated tools makes it now possible to refine global public protection policies by targeting finite resources toward the highest-risk offenders.

References

Andrews, D. A. (2006). Enhancing adherence to risk-need-responsivity: Making quality a matter of policy. *Criminology and Public Policy, 5,* 595–602. doi:10.1111/j.1745-9133.2006.00394.x

Babchishin, K. M., Blais, J., & Helmus, L. (2012). Do static risk factors predict differently for aboriginal sex offenders? A multi-site comparison using the original and revised Static-99 and Static-2002 scales. *Canadian Journal of Criminology and Criminal Justice/La Revue canadienne de criminologie et de justice pénale, 54,* 1–43. doi:10.3138/cjccj.2010.E.40

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012). Even highly correlated measures can add incrementally to predicting recidivism among sex offenders. *Assessment, 19,* 442–461. doi:10.1177/1073191112458312

Blumstein, A., & Nakamura, K. (2009). Redemption in the presence of widespread criminal background checks. *Criminology: An Interdisciplinary Journal, 47,* 327–359. doi:10.1111/j.1745-9125.2009.00155.x

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 civil commitment as sexually violent predators. *Psychology, Public Policy, and Law, 15,* 278–314. doi:10.1037/a0017232

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A. K., & Jeglic, E. L. (2012). Implications of Static-99 field reliability findings for score use and reporting. *Criminal Justice and Behavior, 39,* 42–58. doi:10.1177/0093854811427131

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, United Kingdom: Wiley. doi:10.1002/9780470743386

Bushway, S. D., Nieubeerta, P., & Blokland, A. (2011). The predictive value of criminal background checks: Do age and criminal history affect time to redemption? *Criminology: An Interdisciplinary Journal, 49,* 27–60. doi:10.1111/j.1745-9125.2010.00217.x

California Penal Code. *Retrieved from* http://www.leginfo.ca.gov/.html/pen_table_of_contents.html

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

Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory–Revised: The importance of implementation integrity. *Journal of Criminal Justice, 34,* 523–529. doi:10.1016/j.jcrimjus.2006.09.007

Freeman, N. J., & Sandler, J. C. (2010). The Adam Walsh Act: A false sense of security or an effective public policy initiative? *Criminal Justice Policy Review, 21,* 31–49. doi:10.1177/0887403409338565

Hanson, R. K. (1998). What do we know about sexual offender risk assessment? *Psychology, Public Policy, and Law, 4,* 50–72. doi:10.1037/1076-8971.4.1-2.50

Hanson, R. K., Babchishin, K. M., Helmus, L., & Thornton, D. (2013). Quantifying the relative risk of sex offenders: Risk ratios for Static-99R. *Sexual Abuse: A Journal of Research and Treatment, 25,* 482–515. doi:10.1177/1079063212469060

Hanson, R. K., & Broom, I. (2005). The utility of cumulative meta-analysis: Application to programs for reducing sexual violence. *Sexual Abuse: A Journal of Research and Treatment, 17,* 357–373. doi:10.1177/107906320501700402

Hanson, R. K., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* [Corrections User Report No 2007–05]. Ottawa, Canada: Public Safety Canada.

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 scales. *The International Journal of Forensic Mental Health, 11,* 9–23. doi:10.1080/14999013.2012.667511

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

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

Harris, A. J. R., & Hanson, R. K. (2004). *Sex offender recidivism: A simple question* [User Report 2004–03]. Ottawa, Canada: Public Safety and Emergency Preparedness Canada.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, Canada: Department of the Solicitor General of Canada.

Helmus, L., Hanson, R. K., & Babchishin, K. M. (2009). *Detailed recidivism rate tables: November, 2009*. Retrieved from at www.static99.org

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

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: Journal of Research and Treatment, 24,* 64–101. doi:10.1177/1079063211409951

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley. doi:10.1002/0471722146

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. *Sexual Abuse: A Journal of Research and Treatment, 19,* 425–448. doi:10.1007/s11194-007-9062-3

Joint Legislative Audit and Review Commission. (2011). *Review of the civil commitment of sexually violent predators*. Report to the Governor and the General Assembly of Virginia. Richmond: Commonwealth of Virginia. Retrieved from http://jlarc.virginia.gov

Langan, P. A., Schmitt, E. L., & Durose, M. R. (2003). *Recidivism of sex offenders released from prison in 1994* [NCJ Report no. 198281]. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: Journal of Research and Treatment, 16,* 107–120. doi:10.1177/107906320401600202

Ley, P. (1972). *Quantitative aspects of psychological measurement: An introduction*. London, United Kingdom: Duckworth.

Looman, J., & Abracen, J. (2013). The Static-99R: Are there really differences between normative groups? *International Journal of Offender Therapy and Comparative Criminology, 57,* 888–907. doi:10.1177/0306624X12443657

Mancini, C., Barnes, J. C., & Mears, D. P. (2013). It varies from state to state: An examination of sex crime laws nationally. *Criminal Justice Policy Review, 24,* 166–198. doi:10.1177/0887403411424079

McGrath, R. J., Cumming, G. F., Burchard, B. L., Zeoli, S., & Ellerby, E. (2010). *Current practices and emerging trends in sexual abuser management: The Safer Society 2009 North American Survey* (ISBN: 978–1-884444–85-2). Brandon, VT: Safer Society Press.

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental validity with Static-99R. *Sexual Abuse: Journal of Research and Treatment, 24,* 431–458. doi:10.1177/1079063211432475

Miller, C. S., Kimonis, 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

Minnesota Department of Corrections. (2007). *Sex offender recidivism in Minnesota*. St. Paul, MN: Author.

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *British Medical Journal, 338,* 1317–1320. doi:10.1136/bmj.b375

Olver, M. E., Beggs Christofferson, S. M., Grace, R. C., & Wong, S. C. P. (2013). Incorporating change information into sexual offender risk assessments using the Violence Risk Scale-Sexual Offender Version. *Sexual Abuse: Journal of Research and Treatment*. Advance online publication. doi:10.1177/1079063213502679

Phenix, A., Helmus, L., & Hanson, R. K. (2012). *Static-99R & Static-2002R evaluators' workbook*. Retrieved from www.static99.org

Rettenberger, M., Haubner-MacLean, T., & Eher, R. (2013). The contribution of age to the Static-99 risk assessment in a population-based prison sample of sexual offenders. *Criminal Justice and Behavior, 40,* 1413–1433. doi:10.1177/0093854813492518

Rhine, E. E., Mawhorr, T. L., & Parks, E. C. (2006). Implementation: The bane of effective correctional programs. *Criminology and Public Policy, 5,* 347–358. doi:10.1111/j.1745-9133.2006.00382.x

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63,* 737–748. doi:10.1037/0022-006X.63.5.737

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior, 29,* 615–620. doi:10.1007/s10979-005-6832-7

Rockhill, B., Byrne, C., Rosner, B., Louie, M. M., & Colditz, G. (2003). Breast cancer risk prediction with a log-incidence model: Evaluation of accuracy. *Journal of Clinical Epidemiology, 56,* 856–861. doi:10.1016/S0895-4356(03)00124-0

Royston, P., Moons, K. G. M., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *British Medical Journal, 338,* 1373–1377. doi:10.1136/bmj.b604

Storey, J. E., Watt, K. A., Jackson, K. J., & Hart, S. D. (2012). Utilization and implications of the Static-99 in practice. *Sexual Abuse: Journal of Research and Treatment, 24,* 289–302. doi:10.1177/1079063211423943

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: Journal of Research and Treatment, 14,* 139–153. doi:10.1177/107906320201400205

Thornton, D., Hanson, R. K., & Helmus, L. (2010). Moving beyond the Standard Model for Actuarial Assessment for Sexual Offenders. *California Coalition on Sexual Offending, CCOSO Quarterly Newsletter, Perspectives*. Retrieved from http://ccoso.org/newsletter.php

Thornton, D., & Knight, R. (2013). Construction and validation of the SRA-FV need assessment. *Sexual Abuse: Journal of Research and Treatment*. Advance online publication. doi:10.1177/1079063213511120

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbeman, J. D. F. (2005). Substantial effective samples sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology, 58,* 475–483. doi:10.1016/j.jclinepi.2004.06.017

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

Zgoba, K. M., Miner, M., Knight, R., Letourneau, E., Levenson, J., & Thornton, D. (2012). A multi-state recidivism study using Static-99R and Static-2002 risk scores and tier guidelines from the Adam Walsh Act. Report submitted to the US National Institute of Justice. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/240099.pdf

Received August 28, 2013
Revision received February 4, 2014
Accepted April 9, 2014 ■