Field validity of Static-99R and STABLE-2007 with 4,433 men serving sentences for sexual offences in British Columbia: New findings and meta-analysis

L. Maaike Helmus, R. Karl Hanson, Daniel C. Murrie, & Carmen L. Zabarauckas

Psychological Assessment **(in press, January, 2021)**

Author Note

L. Maaike Helmus, Department of Criminology, Simon Fraser University; R. Karl Hanson, Department of Psychology, Carleton University; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia; Carmen L. Zabarauckas, Ministry of Attorney General (British Columbia) and Department of Criminology, Simon Fraser University .

This work was completed on the traditional and unceded territories of the Algonquin Anishnaabeg People (where the city of Ottawa currently resides), the Coast Salish Peoples (where the city of Burnaby currently resides), specifically the Squamish, Tsleil-Waututh, Musqueam, and Kwikwetlem Peoples, and also the Lekwungen-speaking Peoples and the Songhees and Esquimalt First Nations (where the city of Victoria currently resides).

We thank BC Corrections for sharing data with us. We thank Natasha Usenko at SFU for verifying the analyses. The views expressed in this article are those of the authors and do not necessarily reflect those of the Ministry of Attorney General for British Columbia, or B.C. Corrections.

Correspondence concerning this article should be addressed to L. Maaike Helmus, 10322 Saywell Hall, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6. Maaike_helmus@sfu.ca

Abstract

Many forensic assessment measures are developed and validated under research conditions but applied in the field, where professionals or paraprofessionals have varied training, unknown fidelity to administration procedures, and contextual pressures related to their institutions or legal system. Yet few studies examine the generalizability of psychometric properties of these scales as actually applied in field settings. This study examined 4,433 individuals assessed by probation officers on the Static-99R or STABLE-2007 sexual recidivism risk scales in British Columbia, Canada. Sexual, violent, and any recidivism were examined. Static-99R and STABLE-2007 had moderate accuracy in discriminating recidivists from non-recidivists, and both scales added incrementally in predicting all three outcomes (with Static-99R demonstrating higher accuracy). Organizing the items into constructs, sexual criminality, general criminality, and youthful stranger aggression incrementally predicted all three outcomes. For violent and any recidivism, the incremental effect of sexual criminality was in the negative direction (i.e., high sexual criminality was associated with relatively lower rates of violent and any recidivism). Calibration analyses indicated that recidivism rates were lower than what would be predicted by the norms for the scales. The current study also presented a meta-analysis of 15 field validity studies of Static-99R and 4 field validity studies of STABLE-2007. Results of the current study and meta-analysis support the field application of Static-99R and STABLE-2007, while emphasizing the importance of training and proper implementation.

Keywords: risk assessment; field validity; recidivism; predictive accuracy; sexual offences

Public Significance Statement: Although there are no guarantees that measures developed in research studies will generalize to applied settings, this study found that two sexual recidivism

risk tools (Static-99R and STABLE-2007) predicted reoffending in a large, field validity study of men with a history of sexual offending. Variability in predictive accuracy in previous studies suggest that special efforts (e.g., appropriate training) are required to establish high quality implement recidivism risk tools with high quality.

Field validity of Static-99R and STABLE-2007 with 4,433 men serving sentences for sexual offences in British Columbia: New findings and meta-analysis

Most psychological assessment measures—including forensic assessment measures created for applied use in the criminal justice system—are developed and tested through rigorous research studies. Clinicians, courts, and policy-makers often cite these studies to justify using these scales for individual legal decisions and broader justice policies. There are many reasons, however, that findings from formal research studies of instruments may not generalize to the "real world" field settings in which the instruments are applied (Edens & Boccaccini, 2017; Guarnera & Murrie, 2017; Guarnera et al., 2017; Wood et al., 1996).

Research settings typically prioritize training, fidelity, and inter-rater reliability; raters who are not reliable are re-trained or replaced. In contrast, field practitioners may experience a number of subtle contextual pressures, due to the institutions they serve or their role in an adversarial legal system (i.e., adversarial allegiance; Murrie & Boccaccini, 2015). Where examinee participation is included, research studies typically afford confidentiality or anonymity and no consequences for refusal to participate or for the results of the assessment. In the field there is no anonymity and consequences are not just likely - they are indeed the point of the assessment (i.e., to inform decisions).

Following Edens and Boccaccini (2017), we consider "field" studies to refer to contexts where the assessment data are collected specifically to inform decisions for that individual (i.e., the assessment has real consequences for the evaluee). Of course, "both traditional validity studies and field validity studies are crucial. The former details whether an instrument *can* demonstrate validity under optimal conditions. The latter demonstrates the extent to which an

instrument *does* demonstrate validity as commonly used under routine practice conditions in the field" (Boccaccini & Murrie, 2014, p.7-2).

The emphasis on field studies is surprisingly recent in forensic assessment research (for review, see Edens & Boccaccini, 2017), and field studies tend to reveal weaker reliability and validity compared to development studies or other formal research studies (DeMatteo et al., 2019; Edens & Boccaccini, 2017), although there have been promising field validity studies for scales such as the Level of Service/Case Management Inventory (Wormith et al., 2012), Violence Risk Scale – Sexual Offence version (Olver et al., 2014), and the Sex Offender Risk Appraisal Guide (Rettenberger et al., 2017). This paper presents a large field study and meta-analysis of Static-99R and the STABLE-2007, two scales commonly used to assess static and dynamic risk (respectively) among men charged with sexual offences (Bourgon et al., 2018; Kelley et al., 2020; Neal & Grisso, 2014).

Field Validity Research on Static-99R

Hanson and Morton-Bourgon (2009)'s comprehensive meta-analysis on the accuracy of risk assessment scales for individuals charged/convicted of sex offences had 63 validation studies of Static-99 (N > 20,000; Cohen's d = .67, which roughly corresponds to an AUC of .68 – see Rice & Harris, 2005), of which 62 were research studies. The only field study was a government report (Hanson et al., 2007) finding strong predictive accuracy for the scale as scored by probation and parole officers trained by one of the scale's co-developers (AUC = .74; updated results are in Hanson et al., 2015).

However, later that year the first independent, peer-reviewed field study of Static-99 was published, yielding discouraging results for the scale among individuals screened for civil commitment in Texas (AUC = .57; Boccaccini et al., 2009; analyses updated and subsumed in

Boccaccini & Murrie, 2014; Boccaccini et al., 2017). This large study raised concerns that Static-99 may not generalize well in field practice, particularly outside Canada. Since then, more field validity studies for Static-99R have accumulated, such that a meta-analysis is warranted. There are fewer field validity studies of STABLE-2007, but minimally enough for a meta-analysis.

Static-99R and STABLE-2007: What is Being Measured and How?

Examining the field validity of risk scales can benefit from understanding what they are intended to measure and how. Whereas most psychological assessment measures are norm-referenced, risk assessment scales are fundamentally criterion-referenced (Joint Committee, 2014). Norm-referenced measures compare individuals to a reference group, with results commonly expressed as some form of percentile rank (Crawford & Garthwaite, 2009). In contrast, criterion-referenced prediction tools estimate the likelihood of an outcome that does not currently exist and may never exist. Consequently, different indicators are required to evaluate the validity of prediction tools than those used to evaluate norm-referenced scales (Helmus & Babchishin, 2017). Although prediction tools can be used to assess latent constructs (as in norm-referenced measures), the most fundamental validity evidence for a prediction tool is prospective, empirical associations with the outcome of interest. Specifically, predictive accuracy is indexed by discrimination (the extent to which individuals with the outcome are different from individuals without the outcome) and calibration (the correspondence between the expected and observed rates in validation studies; Steyerberg, 2019).

Although actuarial risk assessment can be atheoretical in origin, there are benefits to understanding the underlying constructs or propensities measured (e.g., integrating results of multiple risk scales; improving predictive accuracy; assessing changes in risk-relevant constructs; stronger linkages between risk assessment and risk reduction). Towards this goal,

some recent research has focused on understanding the major dimensions of sexual recidivism risk (e.g., Brouillette-Alarie et al., 2018; Olver et al., 2016). This task is complicated given that existing risk scales (particularly static scales) are often developed with efficiency in mind, leaving an insufficient pool of items for meaningful factor analyses.

Combining seven samples with both Static-99R and Static-2002R items, Brouillette-Alarie et al. (2016) found evidence of three latent constructs: sexual criminality, general criminality, and youthful stranger aggression. Olver et al. (2016) replicated these constructs using the Static-99R and static items from the Violence Risk Scale – Sex Offence version. Brouillette-Alarie and Hanson (2015) examined these three constructs along with the items of the STABLE-2007 and found that the STABLE items (except loneliness and capacity for relationship stability) could be mapped onto either sexual criminality or general criminality; STABLE-2007 items did not provide meaningful measurement of youthful stranger aggression.

Examining diverse correlates, Brouillette-Alarie et al. (2018) found that sexual criminality primarily assesses deviant sexual interests but is also marked by dysregulation of sexual behaviour, and a general lack of intent to harm the victim. General criminality reflects a propensity for rules violation, as well as antisocial features such as impulsivity and lack of empathy. The youthful stranger aggression construct has been trickiest to understand but this study suggested that it reflects a clear intent to harm the victim, which may be related to some combination of sexual sadism and/or pervasive anger.

Another empirical issue pertains to whether and how Static-99R and STABLE-2007 should be combined. Currently, STABLE-2007 is intended to measure putatively dynamic constructs and is recommended to be used in conjunction with a static risk scale, such as Static-99R (Brankley et al., 2017), with surveys suggesting they are often used in tandem (Bourgon et

al., 2018; Kelley et al., 2020). Given the added time and depth of information required to score STABLE-2007 (and the limited resources of most correctional systems), the incremental value of the scale beyond Static-99R is an important topic, with a recent meta-analysis supporting its added value (Brankley et al., 2019).

Purpose of Study

Recent years have seen an emphasis on field validity research for forensic measures. The current study uses a large field validity sample (N = 4,433) of Static-99R and STABLE-2007 assessments as scored in British Columbia, Canada. To appropriately examine predictive validity, we looked at both discrimination and calibration properties (Helmus & Babchishin, 2017). Furthermore, we provide the first meta-analysis of the field research on these scales. Specifically, this study had five aims:

1) Examine the relative predictive accuracy (discrimination) of items, total scores, and risk categories for Static-99R and STABLE-2007 as assessed by community correctional staff in British Columbia. We expected that Static and STABLE items would demonstrate small predictive accuracy and the scales would demonstrate moderate to large accuracy.

2) Examine how recidivism rates for Static-99R total scores and Static/STABLE combined risk levels compared to the normative data for the scales (calibration). We expected the combined Static/STABLE norms (Brankley et al., 2017) to overestimate observed recidivism in this sample because our data were based on local (i.e., provincial) criminal records and without access to offence details to determine sexual motivation of seemingly non-sexual charges/convictions. In contrast, the estimated recidivism norms for Static-99R and STABLE-2007 (also scored in the field) are based on comprehensive recidivism information from numerous sources and with access to offence descriptions for most violent charges/convictions in

order to identify sexual motivation. For the Static-99R recidivism norms, we did not have much a priori expectation. The sample is a routine correctional sample so the routine norms would be a plausible fit; however, the routine norms overestimated recidivism in two previous studies(Boccaccini et al., 2017; Lee et al., 2016), suggesting they may overestimate recidivism for this sample as well.

3) Assess whether STABLE-2007 provides incremental predictive accuracy above and beyond Static-99R. Consistent with the results of Brankley et al's (2019) meta-analysis (for which this dataset was included), we expected to find incremental validity.

4) Explore the predictive and incremental validity of the three proposed latent constructs for sexual recidivism risk. We expected all three constructs to significantly and incrementally predict sexual recidivism. We did not expect sexual criminality to significantly or incrementally predict violent or any criminal recidivism.

5) Provide a meta-analysis of field validity studies for Static-99R and STABLE-2007, incorporating the current findings. To date, there has been no meta-analysis of field studies of sexual recidivism risk scales, nor has field study status been examined as a moderator of meta-analytic effects, as has been recommended (Edens & Boccaccini, 2017).

Method

Sample

The original dataset included all individuals (N = 4,511) supervised in the community by British Columbia (B.C.) Corrections and who received either a Static-99R, STABLE-2007, or ACUTE-2007 assessment between January 1, 2005 and June 4, 2013. The current study was restricted to men with a Static-99R or STABLE-2007 assessment and who were released prior to June 4, 2013 (N = 4,433). Inclusion in the dataset did not necessarily require that the community

supervision sentence was linked to a current (index) sexual offence. For example, individuals on probation for a non-sexual theft conviction could be included in the sample if they had a prior sexual conviction or a charge for a sexually motivated offence that did not result in a conviction or provincial community supervision sentence (e.g., it may have resulted in a federal sentence), as long as they were assessed on the Static or STABLE during the probation term for the theft (i.e., during his supervision, there was concern about risk of a new sexual offence). The number of individuals charged or convicted of a sexually motivated offence in that timeframe but who were not supervised by B.C. Corrections in that same time period is unknown but likely small. This sample would be considered a routine/complete sample under the definition used by Hanson et al. (2016).

The dataset included a nontrivial number of individuals who had served federal sentences (i.e., custodial sentences of two years or more). In Canada, roughly 11% of sex offence sentences include federal custody (Canadian Centre for Justice Statistics, 2008; for more information on how this estimate was generated, see Hanson et al., 2012[1]). In the current dataset, 256 individuals (5.8%) had a federal sentence associated with one part of their index offence cluster (for explanation of index clusters, see Static-99R coding manual: Phenix, Fernandez et al., 2016), suggesting that federal offenders were reasonably well represented in this dataset.[2]

For 16.7% of the men, it appeared as though the Static and/or STABLE were first assessed pre-conviction (likely as part of a pre-sentence assessment). In 72.2% of cases, the Static/STABLE assessments had a logical link to a sentence for a clear sexual offence. For the remaining cases, either the sentence they were linked to was for a sexually motivated offence but

[1] Thank you to Kelly Morton-Bourgon for locating these data for us.
[2] Also note that of the additional 78 offenders deleted from the original dataset, 57 were offenders who did not have follow-up because they were still serving their federal sentence.

the rapsheet charge/conviction was not explicitly sexual (e.g., a sexually motivated offence pled down to a violent or non-violent charge), or they were being supervised for a non-sexual offence but were assessed on Static/STABLE presumably due to sexual offences in their criminal history. In 21 cases we could not identify any provincial sentence that mapped onto their assessment data, but we retained these assessments on the assumption that the sentence information must be missing or coded with error. In cases where offenders had multiple, non-clustered (see below) sentences for sexual offences with linked Static/STABLE assessments in the timeframe of the dataset, we randomly selected which sex offence was used as the current sentence for this study, with the others thereby categorized as either priors or recidivism incidents. Randomly selecting one of multiple dates reduces the bias that would be introduced by systematically selecting the first (increasing recidivism rates) or the last (decreasing recidivism rates).

For 1,623 offenders (36.6%), their current sentence included at least some time in custody. On average, offenders were 40.8 years old (*SD* = 13.7). Self-reported race/ethnicity data (available for 4,286 offenders) indicated that the sample was predominantly White (*n* = 2,705, 63.1%), followed by Indigenous (which includes First Nations, Métis, and Inuit, *n* = 954, 22.3%), East Indian (*n* = 157, 3.7%), East Asian (*n* = 117, 2.7%), Black (*n* = 64, 1.5%), Hispanic (*n* = 60, 1.4%), and 'other' (*n* = 229, 5.3%). Of those with self-reported data on education (*N* = 4,090), roughly half the sample (52.9%) had completed high school. In terms of highest level of education attained, 17 had no education (0.4%), 134 had some elementary education (3.3%), 548 completed either grades 7, 8, or 9 (13.4%), 1,229 completed grades 10 or 11 (30.0%), 1,398 completed high school (34.2%), 364 completed vocational post-secondary education (8.9%), and 400 completed university (9.8%).

Measures

Static-99R. Static-99R (Hanson & Thornton, 2000; Helmus, Thornton et al., 2012) is a 10-item actuarial risk assessment scale designed to predict sexual recidivism among adult males charged or convicted of a sexually motivated offence. The only scoring difference between Static-99 and Static-99R is the weights and cutoffs for the age at release item. In the current sample, roughly 2/3 of offenders were scored on the original Static-99 before B.C. switched to Static-99R. For these cases, their age was calculated and used to update the score to Static-99R.

Static-99R can be scored with high rater reliability (although there are also some exceptions; for a review, see Phenix & Epperson, 2016) and has moderate ability to discriminate recidivists from non-recidivists in predominantly research studies (AUC ~ .70; Helmus, Hanson et al., 2012). Static-99R total scores range from -3 to 12, and correspond to the following risk levels: I - very low risk (scores of -3 and -2), II - below average risk (scores of -1 and 0), III – average risk (scores of 1, 2, and 3), IVa - above average risk (scores of 4 and 5), and IVb - well above average risk (scores of 6 and higher; Hanson et al., 2017).

STABLE-2007. The STABLE-2007 (Fernandez et al., 2014; Hanson et al., 2007) is an empirically derived risk scale assessing stable dynamic risk factors relevant to treatment and supervision of individuals with a sexual offending history. The scale has 13 items, each scored as 0, 1, or 2 (reflecting no, some, and considerable concern). Total scores range from 0 to 26 for offenders with a victim under 14 years old, and 0 to 24 for others, because one item (emotional identification with children) is only scored for the former group. Total scores were combined with Static-99R following the latest Evaluator Workbook (Brankley et al., 2017). The STABLE-2007 has acceptable field interrater reliability (ICC = .86; Fernandez & Helmus, 2018). A recent meta-analysis found moderate accuracy in predicting sexual recidivism (AUC = .67) and it was incremental to Static-99R (Brankley et al., 2019).

There are some differences between the STABLE-2007 and its precursor scale, the STABLE-2000 (see Hanson et al., 2007), which was used for nearly half of the study sample. For the current study, we present analyses of the altered items separately for the 2000 and 2007 version of the scales. For simplicity, we approximated STABLE-2007 total scores for all offenders by using the 2000 version of the items as applicable.

Constructs in Risk Scales. Following Brouillette-Alarie et al. (2015) and Brouillette-Alarie and Hanson (2015), we used our data to form simple (i.e., summed) subscales to approximate the constructs of sexual criminality, general criminality, and youthful stranger aggression. These are not exact replications as we were missing Static-2002R items. Additionally, we halved each STABLE item (scores of 0/0.5/1) so they would be worth equal weight as most Static items. The internal consistency of the scales was examined using Revelle's omega total from the psych package in R statistics (Revelle, 2020). In comparison to Cronbach's alpha, omega total is appropriate when items are not expected to be equally related to the latent construct (see McNeish, 2018).

Our sexual criminality scale included seven items: non-contact sex offence, male victim, prior sex offences, emotional congruence with children, sexual preoccupation, sex as coping, and deviant sexual interests (Revelle's Omega Total = .80). General criminality included nine items: prior sentencing occasions, prior nonsexual violence, hostility towards women, lack of concern for others, impulsivity, poor problem-solving, cooperation with supervision, significant social influences, and negative emotionality/hostility (Revelle's Omega Total =.76). Youthful stranger aggression included six items: age, never lived with lover, index nonsexual violence, unrelated victim, stranger victim, and capacity for relationship stability (Revelle's Omega Total = .86). All omega values were calculated using polychoric correlations, which minimize the restriction of

range effects commonly found for dichotomous and ordinal items (see McNeish, 2018). Although the STABLE-2007 item for capacity for relationship stability did not form a factor with the youthful stranger aggression Static items in the previous study by Brouillette-Alarie and Hanson (2015), it was included here for face validity reasons, as relationship history forms a core part of rating this item.

Recidivism. The dataset contained records of all charges and convictions within B.C. up to June 4, 2013. Re-organizing the dataset into "sentencing occasion" clusters (see below) meant that pseudo-recidivism was not counted (i.e., new charges for old behaviour). We excluded charges or fines for non-criminal offences (e.g., dog leash laws, fishing out of season). Offence details to determine sexual motivation were not available to us, so all classification decisions were based on the offence name in the charge/conviction. We classified offences into the following categories: non-contact sexual, contact sexual, non-sexual violent, non-violent, and technical (e.g., breach of probation). Where applicable, offence categorization followed Static-99R coding rules (Phenix, Fernandez et al., 2016) and was done independently by the first two authors, with any discrepancies consensified. From the categorizations above, this study analyzed three recidivism outcomes: any sexual recidivism (contact or non-contact), any violent recidivism (which included contact sex offences but not non-contact), and any new crime recidivism (which included all offences but excluded technical breaches).

Follow-up started at conviction date for those without custodial sentences. Community supervision start date after custodial sentences was often available; if not, we estimated it at two thirds of the sentence to account for either time served or early release (e.g., for federally sentenced offenders in that time period, roughly 70% are released at two-thirds of their sentence; Public Safety Canada, 2018). For analyses based on survival data, recidivism (within each

category) was counted based on the date the offence was committed. If offence date was not available (<5% of recidivism incidents), the earliest known charge or conviction date was used. The average length of follow-up (from start date until June 4, 2013) was 4.5 years (*SD* = 2.5). Of the full sample (4,433), recidivism rates were 4.6% for sexual recidivism (*n* = 202), 17.7% for violent recidivism (*n* = 786), and 24.0% for any recidivism (*n* = 1,062).

Procedure

These data were shared by the government of British Columbia pursuant to a research agreement with Public Safety Canada (where the first two authors worked at the time). Subsequent IRB approval was obtained from the primary author's new university affiliation.

Community supervision officers received certified training in scoring Static-99R and STABLE-2007 and underwent an annual peer-review process to ensure quality control. We received information on all Static-99R and STABLE-2007 scores entered by community supervision officers in B.C.'s electronic offender management database in the study timeframe. The data were provided via Excel workbooks extracted from B.C.'s databases. Although we could not verify the accuracy of the scoring by the original officers, we did do some data cleaning (e.g., if a total score did not correspond to the sum of item scores, we updated the total score to reflect the item data).

The criminal history information received included one entry for each step in processing (e.g., charges, bail, sentencing, beginning community supervision). In most cases (particularly for data in 2005 onwards), date of offence was recorded, allowing us to link entries for the same offence (or spree of offences). Where possible, we organized the criminal history data into the concept of 'sentencing occasions' as per Static-99R coding rules (Phenix, Fernandez et al., 2016), where if an offender committed multiple offences on multiple dates before being detected

(i.e., charged) for one of those offences, all offences (and their associated charges, convictions, and sentences, which could have spanned numerous different dates) were grouped as a single offence cluster. Then, we linked the assessment data to the provincial community supervision sentence that most plausibly was the source of the assessments.

The data were a snapshot of all assessments recorded by B.C. Corrections, including repeated assessments over time. Information necessary to verify the reason for the re-assessment (e.g., a new sentence, a new probation officer, new information, assessing change) was not available to the authors. For both scales, we used the first assessment that was plausibly linked to the relevant index offence cluster. Future studies will examine the value of re-assessments.

Although these methods and decision rules add extra approximations, we believe this dataset is fairly generalizable to routine correctional samples of men convicted of sexual offences, including the messiness that typically accompanies real-world implementations of risk scales (e.g., individuals assessed at various stages of criminal justice system processing; individuals with non-sexual offences assessed because of recent sexual offence history; individuals without a sexual offence on their documented criminal history but who were considered to have committed sexually motivated offences, etc.), and likely including real-world staff error, such as potentially using the scale in some cases where it should not have been used.

Overview of Analyses

All analyses were independently verified to check accuracy. Predictive and incremental validity analyses were conducted using Cox regression (Singer & Willett, 2003) with the *survival* package (Therneau, 2020) in R (version 3.6.0 or 4.0.3). Harrell's concordance index (Harrell et al., 1996) was reported as the primary effect size to examine discrimination (from the *survival* package in R) and can be interpreted similar to AUCs, with *C*s of .556, .639, and .714

corresponding to small, moderate, and large Cohen's *d* values (M. E. Rice & Harris, 2005). Results are statistically significant if the 95% confidence interval does not include .50.

For Cox regression models with multiple predictors, we also reported the Bayesian Information Criterion (BIC = -2 LL + [k*ln(n)], where k is the number of parameters and n is the number of recidivists; Raftery, 1995; Volinsky & Raftery, 2000). The BIC compares non-nested models, with smaller BIC values suggesting better fitting models. BIC differences (absolute value) of 0-2, 2-6, 6-10, and 10 and higher, respectively, represent "weak," "positive," "strong," and "very strong" evidence of model fit (Gordon, 2012).

To examine calibration between predicted recidivism rates (from normative data) and observed rates from the current study, we used the E/O index (Hanson, 2017; Rockhill et al., 2003). The E/O index is the ratio of the expected number of recidivists (E) divided by the observed number (O). If the 95% confidence interval includes 1, then the expected recidivism rate is not significantly different than the observed rate. We obtained expected recidivism rates from the 5-year sexual recidivism estimates for Static-99R routine/complete samples (Phenix, Helmus, & Hanson, 2016) and for combined Static-99R/STABLE-2007 risk levels for sexual, violent, and any recidivism (Brankley et al., 2017) and compared them to observed recidivism rates in the current study, restricted to cases with fixed 5-year follow-up data.

Meta-analyses followed the formulae of Borenstein et al. (2009), with one author using the *metafor* program in R (Viechtbauer, 2010) and another using SPSS syntax for verification. Although random-effects analyses are often conceptually preferable, they are unstable when the number of studies is small (< 30, Schulze, 2007); consequently, we reported both fixed- and random-effects analyses. Variability in findings across studies was reported using Cochran's *Q* statistic and the I^2 s effect size statistic (Borenstein et al., 2009). As a rough heuristic, I^2 values of

25%, 50%, and 75% can be considered low, moderate, and high variability, respectively (Higgins et al., 2003). Where not specified otherwise, analyses were conducted in SPSS, version 25 or 27.

Results from Current Study

Online Supplement A includes means for scales and constructs, and frequency data for scale items. It also presents means and standard deviations for sexual recidivists and non-recidivists for items, scales, and constructs. Online Supplement B presents correlations between scales and constructs.

Predictive Accuracy of Static-99R and STABLE-2007

Table 1 presents the discrimination analyses for Static-99R and STABLE-2007 items, total scores, risk levels (including the risk level from combining Static-99R and STABLE-2007), and constructs. For predicting sexual recidivism, all Static-99R items were statistically significant, with the exception of index non-sexual violence, which closely approached significance (confidence interval lower limit = .50). Effect sizes ranged from .53 to .62, with prior sex offences and 4+ sentencing occasions demonstrating the largest effects (*C* values above .60). Static-99R total scores approached a large effect size (.70). There was a small drop in accuracy examining risk levels instead of total scores (*C* = .69), which is typical when clumping scores into groups (Howard, 2017). Effect sizes were similar for the five risk levels based on Static-99R scores alone versus combined Static-99R and STABLE-2007 (*C* = .687 vs .694).

Results were similar between the prediction of violent and any criminal recidivism. Some Static-99R items (particularly those specific to sex offending) were not significant. The strongest items were age, prior non-sexual violence, and 4+ prior sentencing occasions (*C*s above .60). Total scores and risk levels predicted violent and any recidivism with *C*s between .68 to .70.

STABLE-2007 total scores predicted sexual, violent, and criminal recidivism with significant, moderate, and similar accuracy (*C*s of .67, .66, and .67, respectively), albeit Static-99R had significantly stronger predictive accuracy for violent and any recidivism (from non-overlapping 84% confidence intervals). All STABLE-2007 items significantly predicted sexual recidivism, with *C* values of .55 and above. Impulsivity and poor cognitive problem-solving were the strongest predictors of sexual recidivism, with moderate effect sizes (> .64). For violent and any criminal recidivism, some of the sexual-specific risk factors were no longer significant predictors. For the general criminality items, effect sizes tended to be higher for violent and any recidivism compared to sexual recidivism (specifically: significant social influences, hostility towards women, lack of concern for others, impulsivity, poor cognitive problem-solving, negative emotionality/hostility, and cooperation with supervision).

Examining the STABLE-2000 items for sexual recidivism, emotional identification with children and child molester attitudes did not significantly predict sexual recidivism. Sexual entitlement and rape attitudes were significant with small effect sizes (.57 and above), indicating their removal from STABLE-2007 may have been premature. Relationship stability in the STABLE-2000 did significantly predict sexual recidivism (*C* = .57), but with notably lower accuracy than the revised item for STABLE-2007 (*C* = .62). Deviant sexual interests predicted significantly and similarly for both the STABLE-2000 definition (*C* = .59) and the STABLE-2007 definition (*C* = .58).

The three latent constructs predicted sexual recidivism with similar moderate accuracy (.64 to .67). For violent and any recidivism, sexual criminality was, unsurprisingly, a poor predictor (*C*s ≤ .54), youthful stranger aggression was a moderator predictor (*C* = .64), and general criminality was a strong predictor (*C* = .74).

Calibration Analyses

Table 2 presents the calibration analyses using the E/O index, which compares the expected number of recidivists at five years (from normative data; Brankley et al., 2017; Phenix, Helmus, & Hanson, 2016) to the observed number of recidivists (restricted to a fixed five-year follow-up). Data are presented overall and by risk level. For Static-99R, the routine/complete normative data predicted nearly twice as many sexual recidivists as there were observed (E/O = 1.81), and this overestimation was significant for the overall sample and for Risk Levels III, IVa, and IVb. In Risk Level I, 6 out 110 offenders sexually reoffended, which was a higher recidivism rate than Risk Levels II and III. Consequently, Static-99R significantly underpredicted recidivism for Risk Level I (E/O = 0.20).

For the risk levels combining Static-99R and STABLE-2007, as hypothesized, the expected number of sexual recidivists was significantly higher than what was observed in British Columbia. Specifically, for the total sample, the norms predicted more than two times the number of sexual recidivists than what were observed (E/O = 2.21), and the overestimation was significant in all Risk Levels except I, where the *n* was too small for meaningful analysis (3 recidivists compared to 3.6 expected recidivists).

Contrary to our hypothesis, there was no significant difference overall between expected and observed violent recidivists (E/O = 1.07). For any criminal recidivism, the expected total number of recidivists was significantly higher than observed, but the magnitude of this ratio was fairly small (E/O = 1.12), with the scale predicting 12% more recidivists than observed. For violent and any recidivism, the results are a little more muddied when examining the risk levels: the norms significantly over-predicted recidivism for Risk Levels II (below average) and IVa (above average) but significantly under-predicted recidivism for Risk Level III (average risk).

This is likely a reflection of instability in the original recidivism norms, which are based on a much smaller sample (*n*'s of 122 to 237 across Risk Levels II, III, IVa) and show less differentiation among Risk Levels II and III.

Incremental Validity

Table 3 presents the Cox regression analyses of incremental predictive accuracy of Static-99R and STABLE-2007 in one model, and of the three proposed sexual recidivism constructs in another model. For each analysis, we present the hazard ratio of each predictor, which indicates how the rate of recidivism changes for each one-point increase in the predictor, averaged across time, and controlling for the other predictor(s) in the model. Normally it is difficult to compare magnitude of prediction from hazard ratios because they are tied to the units of the predictor (e.g., Static-99R scores have a smaller range, varying from -3 to 12, whereas STABLE scores can vary from 0 to 26). Measures with larger ranges tend to have smaller hazard ratios because each one-point increase makes a small contribution to increasing recidivism. Fortunately for the analyses of sexual recidivism constructs, all three predictors had observed scores on a 10-point scale so the hazard ratios can be directly compared. We also present the Harrell's *C* index for the overall model combining all predictors, and BIC values to compare models.

Static-99R and STABLE-2007 both added significant and positive incremental validity in predicting sexual, violent, and criminal recidivism. In the combined regression model (which optimizes the weight of the total scores of each scale for this sample), effect sizes were slightly higher than the comparable effect size for the five risk levels from combining Static-99R and STABLE-2007 (presented in Table 1). Specifically, effect sizes for the regression model (compared to the five risk levels) were .716 for sexual recidivism (compared to .694), .706 for violent recidivism (compared to .693), and .716 for any criminal recidivism (compared to .699).

Examining the three dimensions of sexual recidivism risk (sexual criminality, general criminality, and youthful stranger aggression), in bivariate analyses (Table 1), all three predicted sexual recidivism with similar accuracy. In the Cox regression analysis, all three constructs added significant incremental predictive accuracy, with a large effect size for the overall model (*C* = .719). Controlling for the other constructs, general criminality had the smallest effect (HR = 1.13) and sexual criminality had the largest effect (HR = 1.24).

For violent and any criminal recidivism in bivariate analyses, sexual criminality had a positive *C* value (above .50), but was a poor predictor (*C*s < .54). After controlling for general criminality and youthful stranger aggression, however, sexual criminality demonstrated significant incremental predictive accuracy in the negative direction (i.e., a suppression effect; HR = 0.86 for violent recidivism and 0.90 for general recidivism; Table 3). Specifically, controlling for general criminality and youthful stranger aggression, each one-point increase in sexual criminality was associated with 10% to 14% less nonsexual recidivism, averaged across the follow-up period. General criminality and youthful stranger aggression both added significant positive incremental accuracy to predicting sexual and violent recidivism. For both outcomes, the effect size for general criminality (HR ~ 1.4) was roughly twice the effect size for youthful stranger aggression (HR ~ 1.2).

Looking just at the *C* values, the models combining the three constructs had higher accuracy than the models combining Static/STABLE total scores. For a more detailed comparison of the two models within each outcome, BIC values were examined to also account for the number of parameters. For sexual recidivism, the difference between models was on the cusp between positive and strong (favouring the Static/STABLE model; BIC difference = -6.18).

For violent and any recidivism, however, the difference was very strong and favoured the model with the three constructs (BIC differences > 239).

Results from Meta-Analysis

Search and Descriptive Information

We sought any field study examining predictive accuracy of Static-99R or STABLE-2007 for sexual recidivism. For each scale, the first author searched Google Scholar for the scale name, and PsycINFO with scale name anywhere in the abstract. Searches included variations (e.g., with and without hyphen, and removing the "R" to include things like Static-99/R"). The first author was familiar with most of the validation studies and examined the methods section of ones known or suspected to be field studies, or any study where the context was unknown. This search procedure was narrow in scope and focused on research published since 2007 (when the first of the two scales was developed).

Including the current study, 13 field validity studies of Static-99R were identified, representing 15 non-overlapping samples (e.g., Boccaccini & Murrie, 2014 was subsumed in Boccaccini et al., 2017; Hanson et al., 2014 was subsumed in Lee et al., 2018). One study of Static-99R and STABLE-2007 (Veith, 2018) was excluded because it did not report results for any sexual recidivism.

Basic descriptive data for these studies are provided in Online Supplemental Material C, and predictive accuracy data are in Table 4. Smeth (2013) included two separate samples. We separated samples pre-2004 and 2004-onwards from Boccaccini et al. (2017), because the 2004-onwards cohort was assessed with significantly higher inter-rater reliability (A. K. Rice et al., 2014) and significantly better predictive accuracy (Boccaccini et al., 2017), possibly indicating a

qualitative difference in field validity implementation after the 2003 coding manual was published (Harris et al., 2003).

For the 15 samples, 10 were from the U.S., 3 were from Canada, and one each was from Austria and Australia. All the studies were of fairly routine (i.e., representative) samples of offenders released from prison and/or on probation (for Smeth [2013], details of sample selection were not provided), although two studies were preselected for those with custodial sentences of 1 (Smallbone & Rallings, 2013) and 2 years (Olver et al., 2014). Combined together, the studies include 47,925 sex offenders (over twice the sample size from Hanson and Morton-Bourgon's 2009 meta-analysis of 63 studies). In all but two studies (Rettenberger et al., 2013; Smallbone & Rallings, 2013), the scale was scored primarily by parole and probation officers or other non-psychologist corrections staff. In some studies the original Static-99 was scored, with the age item later updated by researchers to obtain Static-99R scores.

From Table 4, AUCs ranged from .60 to .81, with an unweighted median of .690. The two samples from Texas and two studies from California reported calibration data (Table 4). Combining these results with the current study, the Static-99R would have predicted roughly twice as many recidivists as there were observed, E/O index = 1.88, 95% CI [1.76, 2.01].

For STABLE-2007, we found four studies (including the current study), summarized in Online Supplement D. One field validity study was excluded as there were only two sexual recidivists (out of 245 offenders), precluding analyses (Tamatea, 2014). Etzler et al.'s (2020) data are from the same Austrian evaluation center as the Rettenberger et al. (2013) study of Static-99R but in contrast, it is not a routine sample of offenders. All inmates in Austria are scored on Static-99R, but only roughly 60% are scored on STABLE-2007. In that study, STABLE-2000 scores were converted to STABLE-2007 scores as applicable, and significant

social influences was omitted due to insufficient information. Relatedly, the Hanson et al. (2015) study is the same sample that was used to revise the STABLE-2000 scale into the STABLE-2007, so the modifications in the scale were optimized based on that sample. For the Smeth (2013) study, only one of the samples (Study 2) had STABLE-2007 scores. Across these three studies, AUCs for predicting sexual recidivism varied from .62 to .67.

Meta-Analytic Findings

Table 5 presents the meta-analytic results of Static-99R and STABLE-2007 field validity studies, both prior to and including the current study. Given that AUCs and Harrell's *C* values can be interpreted the same way, they are treated interchangeably. Myer (2019) and Buttars et al. (2015) did not report a standard error or confidence interval for the AUC. For these studies, the AUC was converted to a Cohen's *d* (following M. E. Rice & Harris, 2005), and then the standard error of *d* was calculated based on the number of recidivists and non-recidivists (Borenstein et al., 2009). Confidence intervals for Cohen's *d* were converted to AUCs, from which we worked backwards to get the standard error of the AUC.

Before adding the current study, Static-99R demonstrated moderate accuracy in predicting sexual recidivism (fixed-effect weighted AUC/C = .660, random-effects = .686). Adding in the current study led to a very small increase in these effects (fixed-effect = .665, random-effects = .688). In both analyses, the variability in predictive accuracy across studies was significant and large; 75% of the variability was above and beyond what you would expect purely from sampling error (I^2 = 75%). Using fixed-effect cumulative meta-analysis (Hanson & Broom, 2005), the change in *Q* as a result of adding in the current study was statistically significant (Q_{change} = 5.1, *df* = 1, *p* = .024), indicating that the current study was significantly higher than the meta-analytic average without it.

None of the studies met Hanson and Bussière's (1998) definition of an outlier in the overall analysis; however, it is possible for a single study (or studies) to have an undue influence on the findings even if it is not the most extreme finding (e.g., Helmus et al., 2013). In particular, unusually large studies can be given too much weight in fixed-effect meta-analysis, obscuring the contribution of the remaining studies. Given that the two samples from Texas comprise 72% of the cases in this overall meta-analysis, this study was weighted heavily. Specifically, the weight assigned to these studies combined was 16,718, which was roughly 6 times more weight than the next largest study (2,890) and roughly 167 times more weight than the smallest study (100).

Consequently, we conducted an additional meta-analysis excluding Texas. The remaining 13 field validity samples demonstrated higher AUC values (fixed-effect and random-effects AUC/C = .697), although there was still significant variability in effects across studies (I^2 = 65%). This difference was not substantial and indicated that although Texas had strong weight in the meta-analysis, results are not dramatically different depending on whether it is included or excluded (so it is included for remaining analyses). Excluding Texas, however, the current study was no longer significantly different than the meta-analytic average (Q_{change} = 0.28, df = 1, p = .597), indicating that the previous finding could be explained by significantly different results between Texas and B.C.

To further explore the significant variability in findings across studies, we examined two moderator variables commonly discussed in field validity research: author involvement and quality of training/implementation. When classifying studies based on whether one of the co-authors of the research study was also part of the scale's development team, results were identical for fixed-effect analyses (AUC/C = .66) and nearly identical for random-effects

analyses (AUC/C = .70 when a scale author was involved compared to .67 when an author was not involved). This moderator was not statistically significant (fixed-effect $Q_{Between} = 0$).

For quality of training/implementation, we classified the studies into two categories. There were six studies where staff were trained by a certified Static-99R trainer or followed a comparable accredited jurisdiction-wide system for training (Hanson et al., 2015; Lee et al., 2016, 2018; Olver et al., 2014; Rettenberger et al., 2013; and the current study). In the other nine samples, training is not discussed or it is unknown whether it would be considered certified training as per the scale's coding manual recommendations (Phenix, Fernandez et al., 2016). Studies with appropriate training demonstrated a large effect size (AUC/C = .72 in fixed and random-effects models), whereas studies where the appropriateness of the training was unknown demonstrated moderate accuracy (AUC/C = .64 to .65). With this classification, results were consistent across studies with appropriate training, but there was still significant and moderate variability in the group with unknown training. The variability between these two groups was significant ($Q_{between} = 35.5$, $df = 1$, $p < .001$), indicating that training (as categorized here) is a significant moderator of predictive accuracy.

For STABLE-2007, prior to adding the current study, the three previous field validity studies demonstrated moderate predictive accuracy (fixed and random-effects AUC = .65) with no significant variability in the findings across studies ($I^2 = 0.0\%$). Adding in the current study increased the average effect from .65 to .66, still with no significant variability ($I^2 = 0.0\%$).

Discussion

Many forensic assessment measures are developed through carefully controlled research procedures, including coders trained to high reliability and proper administration procedures. However, the same instruments are usually applied in the field by professionals or

paraprofessionals (e.g., correctional or probation staff) who may have varied training, unknown fidelity to administration procedures, and contextual pressures related to the institutions or the adversarial legal system in which they work. Consequently, there have been calls for "field studies" to shed light on how the instruments perform when applied under routine conditions in the field (Edens & Boccaccini, 2017). We responded to this call by presenting a large field study of two instruments commonly used to assess sexual recidivism risk, and meta-analyzing previous field studies of these instruments. Next to the study from Texas which examined Static-99R only (Boccaccini et al., 2017), this is the largest field validity study of Static-99R and STABLE-2007 to date.

The findings from this routine/complete field sample support the continued use of both scales in British Columbia, and generally found that the scales predicted similar to the average of other field validity studies (and Static-99R predicted higher than average when the Texas samples are retained). All items of both scales significantly predicted sexual recidivism as intended, with the exception of the Static-99R item for index non-sexual violence This is consistent with the meta-analysis of Helmus and Thornton (2015) who found this item to be the weakest Static-99R item, although it did predict in North American samples (not so much elsewhere).

These encouraging findings coincide with promising field results from other recidivism prediction tools including the VRS-SO (Olver et al., 2014), LS/CMI (Wormith et al., 2012), and SORAG (Rettenberger et al., 2017). Additionally, our meta-analysis (to our knowledge the first meta-analysis of field validity) found comparable predictive validity for Static-99R and STABLE-2007 to predominantly research studies (Brankley et al., 2019; Helmus, Hanson et al.,

2012), although the below-average results from the large Texas studies do somewhat pull down the findings for Static-99R.

These results are in contrast to previous reviews finding that field reliability/validity tend to be weaker compared to instrument development or research studies (DeMatteo et al., 2019; Edens & Boccaccini, 2017). We suspect two primary reasons for this discrepancy. One is that much of this previous research has focused on the Psychopathy Checklist-Revised (PCL-R), which involves more subjectivity and inferences in scoring, especially compared to Static-99R. Consequently, the PCL-R may be vulnerable to greater decreases in reliability or administration fidelity, compared to Static-99R and the STABLE-2007, when applied in routine practice. Conversely, however, all the studies in our meta-analysis examined uses of the scales in correctional settings, whereas much PCL-R field research (and Static-99R field interrater research) has considered more adversarial court settings, particularly civil commitment in the United States. It is possible that adversarial contexts add distinct external pressures, which lower the reliability/validity of forensic measures. Most likely, both explanations are credible. One randomized study found that Static-99R is susceptible to adversarial allegiance biases in civil commitment proceedings, albeit markedly less so than the PCL-R (Murrie et al., 2013).

This meta-analysis also revealed that the quality of training and implementation are critical components for the accuracy of risk tools in field assessments. Although several of the studies included did not provide sufficient detail for a comprehensive assessment of implementation issues, a preliminary moderator analysis based on the information we did have found that studies with appropriate training systems in place for the officers administering the tools such as B.C. Corrections, found meaningfully and significantly higher accuracy than studies where the appropriateness of training was unknown. In contrast, there were no differences in predictive

accuracy based on whether one of the co-authors of the scales was also a co-author on the study, suggesting no author allegiance effects, as reported in other reviews (Blair et al., 2008).

The routine recidivism norms of Static-99R significantly overestimated sexual recidivism in the current sample. This was also consistent with the studies from Texas and California (included in the meta-analysis). This could be due to several reasons. Firstly, the studies from Texas, California, and the current study only had access to provincial or state recidivism data, whereas many samples from the Static-99R normative data had nationwide recidivism sources (Phenix, Helmus, & Hanson, 2016). Additionally, Texas, California, and B.C. did not have access to detailed data on circumstances of recidivism incidents to identify sexual motivation, which was done in some of the normative samples as well. Consequently, the lower recidivism rates in the current sample could reflect less comprehensive recidivism data. Another possibility is that the routine recidivism norms for Static-99R (which were based predominantly on offenders released pre-2000) need updating. Modern samples in Canada and the U.S. may have lower sexual recidivism rates than the earlier studies.

Comparing the calibration results for the combined Static/STABLE risk levels for sexual, violent, and any recidivism provides some insight into the reasons for the low sexual recidivism rates in the current sample. The current recidivism norms (Brankley et al., 2017) were derived from a single sample (Hanson et al., 2015) that accessed numerous sources of recidivism information including national and provincial criminal history records, police agencies, newspapers, and community supervision officers. Recidivism incidents were usually associated with a charge, but not always. Additionally, for many offences, police agencies were contacted to determine the circumstances so that sexual recidivism included sexually motivated offences, regardless of what charge was laid. In contrast, the current study used one source of information,

restricted to the province, and with no details to determine motivation. The timeframes of the data sources would be roughly similar, as Hanson et al. (2015) included offenders on community supervision between 2001-2005, whereas the 5-year data for the current study would have included offenders supervised between 2005-2008, so cohort effects are not expected.

Given these differences in methodology, it is somewhat surprising that the Static/STABLE recidivism norms did not significantly overestimate violent recidivism, and the overestimation of any recidivism was quite small (E/O index = 1.12). This suggests that rates of new charges and convictions were actually quite similar in B.C. and in the normative samples. Consequently, the normative data's overestimation of sexual recidivism may be due to lack of access to offence circumstances in the current study as opposed to different rates of recidivism, as it would be less likely for sexual recidivism rates to be half of the normative rates, but violent recidivism rates to be comparable.

The construct validity analyses reinforced the view that risk for sexual recidivism is multidimensional (Brouillette-Alarie et al., 2016; Doren, 2004; Olver et al., 2016; Seto, 2019). Sexual crimes are crimes; consequently, the factors associated with general rule violation (e.g., prior criminal convictions, lifestyle impulsivity) also increase the risk of sex crime recidivism. There are, however, sex-crime-specific risk factors, which are only weakly related to other types of rule violation. In the current study, the sex-crime-specific construct displayed weak positive associations with violent and general recidivism in the univariate analyses; however, in the multivariate analyses, they were negatively associated with these outcomes after controlling for general criminality and youthful stranger aggression. A similar negative association has been previously observed (Babchishin et al., 2016), which motivated the Static Development Team to recommend against using Static-99R to predict nonsexual violent and general recidivism;

instead, they recommend that violent and general recidivism be assessed by young age and the general criminality factor from the Static-2002R (Babchishin et al., 2013, 2016).

One interesting finding was that for violent and any recidivism, the overall predictive accuracy (discrimination) was higher when the items were organized by constructs than when they were organized by type of variable (i.e., demographic and criminal history variables form the total score of Static-99R; STABLE-2007 total scores are based on evaluators' ratings of psychological and community adjustment, largely from interview). Neither Static-99R nor STABLE-2007 were intended to be internally consistent; instead, the items were selected to cover a range of risk factors and diverse latent constructs. The total scores were presented as a convenient way of expressing the relative density of risk factors present in a specific case.

The current results suggest that information may be lost by only considering total scores. It may be possible to improve predictive accuracy by considering the constructs assessed by risk tools. Rather than a list of risk factors, future risk tools could include subscales addressing latent constructs, and the overall assessment of risk could be based on combining subscale scores. Such an approach also has the potential of identifying psychologically meaningful propensities relevant to intervention and risk management strategies (e.g., Mann et al., 2010). Future advances could include risk models whereby constructs are weighted differentially depending on what outcomes is being predicted.

Limitations

The current study had many limitations common to real-world data. Administrative data is built for administrative purposes and day-to-day supervision of clients. Linkages between charge(s), conviction(s), assessments, and treatment plan for each client may be visible to corrections staff but not maintained when the data are extracted from the 'back end' of the

system. For researchers, this means administrative data appear scattered and 'messy' and require significant preparation and effort to re-link. Particularly for linking Static/STABLE assessments with provincial corrections sentences, it meant we were making decisions without access to the same information that officers were using when they applied the assessment scales. Additionally, no inter-rater reliability information was available for officer scorings, or additional data to verify the quality/accuracy of the assessment scores, or whether they were scoring the scales on appropriate cases. Consequently, errors and approximations in the dataset (and in the original scores) are unavoidable.

We also did not have access to federal criminal history records, which limited our recidivism data to new charges or convictions within British Columbia. The inability to access offence details for recidivism incidents and the lack of national criminal history records would certainly explain some of the gap between the observed number of recidivists and the number predicted by the normative data for the risk scales; unfortunately, how much of the gap is explained by these limitations is unknown. Given the limitations of the recidivism information, we do not support using this dataset to develop new normative recidivism data for Static-99R or Static/STABLE-2007 at this point in time.

Further Research

This meta-analysis was restricted to the field use of Static-99R and STABLE-2007 in correctional settings. More research is needed to determine how these results would generalize to court settings (e.g., civil commitment or Dangerous Offender proceedings), where the stakes are higher and pressures different. Additionally, the meta-analysis appears to support the critical role of appropriate training and implementation, but there is currently insufficient empirical evidence

to provide more detailed guidance for providing optimal training and ongoing support (e.g., ideal length of training, number of participants, modality of training, testing standards, follow-up).

Importantly, this paper only examined the first assessment during community supervision. The main purpose of risk assessment is to guide intervention and supervision decisions (e.g., intensity, targets) and re-assess risk as appropriate. This paper supports Static-99R and STABLE-2007 as baseline assessments of static and putatively dynamic risk near the start of community supervision. Future research is needed to examine how these assessments inform case management, and how re-assessments of dynamic risk factors (particularly for the STABLE-2007 and the ACUTE-2007, the latter of which measures more rapidly changing risk factors) can improve prediction and case management. Such research with this dataset is currently underway.

Conclusions and Implications for Practice

This study supports the continued use of both Static-99R and STABLE-2007 in British Columbia; further research would be beneficial to explore subgroups based on offence types (e.g., internet offences, non-contact offenders), cross-cultural validity (e.g., predictive accuracy with ethnic subgroups), and re-assessments over time. This study (and the cumulative meta-analysis) adds to the burgeoning literature suggesting that actuarial risk scales such as Static-99R and STABLE-2007 can be scored with reasonable fidelity and accuracy in real-world settings.

References

Babshishin, K. M., Hanson, R. K., & Blais, J. (2016). Less is more: Using Static-2002R subscales to predict violent and general recidivism among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 28*(3), 187-217. http://dx.doi.org/10.1177/1079063215569544

Babchishin, K. M., Hanson, R. K., & Blais, J. (2013). *User Guide for the Brief Assessment for Recidivism Risk – 2002R (BARR-2002R)*. Available from www.static99.org.

Blair, P.R., Marcus, D.K., & Boccaccni, M.T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessments as an exemplar. *Clinical Psychology Science and Practice, 15,* 346-360.

Boccaccini, M.T., & Murrie, D.C., (2014). Keeping up with the field in "field reliability" risk assessment research. In A. Schlank (Ed.), *The Sexual Predator* (Vol 5). Kingston, NJ: Civic Research Institute.

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*(4), 278-314. http://dx.doi.org/10.1037/a0017232

*Boccaccini, M. T., Rice, A. K., Helmus, L. M., Murrie, D. C., & Harris, P. B. (2017). Field validity of Static-99/R scores in a statewide sample of 34,687 convicted sexual offenders. *Psychological Assessment, 29*(6), 611-623. http://dx.doi.org/10.1037/pas0000377

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, U.K.: Wiley.

Bourgon, G, Mugford, R, Hanson, R. K., & Coligado, M. (2018). Offender risk assessment practices vary across Canada. *Canadian Journal of Criminology and Criminal Justice, 60*(2), 167-205. http://dx.doi.org/10.3138/cjccj.2016-0024

Brankley, A. E., Babchishin, K. M., & Hanson, R. K. (2019). STABLE-2007 demonstrates predictive and incremental validity in assessing risk-relevant propensities for sexual offending: A meta-analysis. *Sexual Abuse.* Advance online publication. http://dx.doi.org/10.1177/1079063219871572

Brankley, A. E., Helmus, L. M., & Hanson, R. K. (2017). *STABLE-2007 Evaluator Workbook – Revised 2017*. Ottawa, ON.

Brouillette-Alarie, S., Babchishin, K. M., Hanson, R. K., & Helmus, L. M. (2016). Latent constructs of static risk scales for the prediction of sexual recidivism: A 3-factor solution. *Assessment, 23*(1), 96-111. http://dx.doi.org/10.1177/1073191114568114

Brouillette-Alarie, S., & Hanson, R. K. (2015). Comparaison de deux mesures d'évaluation du risque de récidive des délinquents sexuels. *Canadian Journal of Behavioural Science, 47*(4), 292-304. http://dx.doi.org/10.1037/cbs0000019

Brouillette-Alarie, S., Proulx, J., & Hanson, R. K. (2018). Three central dimensions of sexual recidivism risk: Understanding the latent constructs of Static-99R and Static-2002R. *Sexual Abuse, 30*(6), 676-704. http://dx.doi.org/10.1177/1079063217691965

*Buttars, A., Huss, M. T., & Brack, C. (2015). Sex offender risk assessment: A reexamination of the coffee can study. *International Journal of Law and Psychiatry, 42-43*, 31-36. http://dx.doi.org/10.1016/j.ijlp.2015.08.004

Canadian Centre for Justice Statistics. (2008). *Adult Criminal Court Survey*.

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist*, *23*(2), 193–204. https://doi.org/10.1080/13854040801968450

DeMatteo, D., Murrie, D. C., Edens, J. F., & Lankford, C. (2019). Psychopathy in the courts. In M. DeLisi (Ed.), *Routledge international handbook of psychopathy and crime* (pp. 645–664). New York, NY: Routledge/Taylor & Francis Group.

Doren, D. M. (2004). Toward a multidimensional model for sexual recidivism risk. *Journal of Interpersonal Violence*, *19*(8), 835-856. https://doi.org/10.1177/0886260504266882

*Duwe, G., & Rocque, M. (2018). The home-field advantage and the perils of professional judgment: Evaluating the performance of the Static-99R and the MnSOST-3 in predicting sexual recidivism. *Law and Human Behavior, 42*(3), 269–279. http://dx.doi.org/10.1037/lhb0000277

Edens, J. F. & Boccaccini, M. T. (2017). Taking forensic mental health assessment 'out of the lab' and into 'the real world': Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29*(6), 710-719. http://dx.doi.org/10.1037/pas0000475

*Etzler, S., Eher, R., & Rettenberger, M. (2020). Dynamic risk assessment of sexual offenders: Validity and dimensional structure of the STABLE-2007. *Assessment, 27*(4), 822-839. http://dx.doi.org/10.1177/1073191118754705

Fernandez, Y. M., Harris, A. J. R., Hanson, R. K., & Sparks, J. (2014). *STABLE-2007 coding manual – revised 2014*. Unpublished report. Ottawa, ON: Public Safety Canada.

Fernandez, Y. M.., & Helmus, L. M. (2017). A field examination of the inter-rater reliability of the Static-99 and STABLE-2007 scored by Correctional Program Officers. *Sexual Offender Treatment, 12*(2), 1-9.

Gordon, R. (2012). *Applied statistics for the social and health sciences*. Routledge.

Guarnera, L., Murrie, D. C., & Boccaccini, M. T. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science, 3*(2), 143-152. https://doi.org/10.1037/tps0000114

Guarnera, L. & Murrie, D. C. (2017). Field reliability of competence and sanity opinions: A systematic review and meta-analysis. *Psychological Assessment, 29*(6), 795-818. http://dx.doi.org/10.1037/pas0000388

Hanson, R. K. (2017). Assessing the calibration of actuarial risk scales: A primer on the E/O Index. *Criminal Justice and Behavior, 44*(1), 26-39. http://dx.doi.org/10.1177/0093854816683956

Hanson, R. K., Babchishin, K. M., Helmus, L. M., Thornton, D., & Phenix, A. (2017). Communicating the results of criterion-referenced prediction measures: Risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. *Psychological Assessment, 29*(5), 582-597. http://dx.doi.org/10.1037/pas0000371

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*(4), 357-373. http://dx.doi.org/10.1007/s11194-005-8049-1

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sex offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*(2), 348-362. http://dx.doi.org/10.1037/0022-006X.66.2.348

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 Research User Report No. 2007-05). Ottawa, ON, Canada: Public Safety Canada.

*Hanson, R. K., Helmus, L. M., & Harris, A. J. R. (2015). Assessing the risk and needs of supervised sexual offenders: A prospective study using STABLE-2007, Static-99R, and Static-2002R. *Criminal Justice and Behavior, 42*(12), 1205-1224. http://dx.doi.org/10.1177/0093854815602094

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*(1), 11-23. http://dx.doi.org/10.1080/14999013.2012.667511

Hanson, R. K., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management, 1*(2), 102-117. http://dx.doi.org/10.1037/tam0000014

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), 1-21. http://dx.doi.org/10.1037/a0014421

*Hanson, R. K., Newstrom, N., Brouillette-Alarie, S., Thornton, D., Robinson, B. E., & Miner, M. H. (2020). Does reassessment improve prediction? A prospective study of the Sexual Offender Treatment and Intervention Progress Scale (SOTIPS). *International Journal of Offender Therapy and Comparative Criminology*. Advance online publication. doi. 10.1177/0306624X20978204

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24*(1), 119-136. http://dx.doi.org/10.1023/A:1005482921333

Hanson, R. K., Thornton, D., Helmus, L. M., & Babchishin, K. M. (2016). What sexual recidivism rates are associated with Static-99R and Static-2002R scores? *Sexual Abuse: A Journal of Research and Treatment, 28,* 218-252. doi:10.1177/1079063215574710

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine, 15,* 361-387.

Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, ON: Solicitor General Canada. www.static99.org/pdfdocs/static-99-coding-rules_e.pdf

Helmus, L. M., & Babchishin, K. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior, 44*(1), 8-25. http://dx.doi.org/10.1177/0093854816678898

Helmus, L., Hanson, R. K., Babchishin, K. M., & Mann, R. E. (2013). Attitudes supportive of sexual offending predict recidivism: A meta-analysis. *Trauma, Violence, and Abuse, 14,* 34-53. doi:10.1177/1524838012462244

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 & Behavior, 39*(9), 1148-1171. http://dx.doi.org/10.1177/0093854812443648

Helmus, L. M., & Thornton, D. (2015). Stability, predictive, and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism: A meta-analysis. *Criminal Justice and Behavior*, *42*(9), 917-937. http://dx.doi.org/10.1177/0093854814568891

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*(1), 64-101. http://dx.doi.org/10.1177/1079063211409951

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560. http://dx.doi.org/10.1136/bmj.327.7414.557

Howard, P. D. (2017). The effect of sample heterogeneity and risk categorization on Area Under the Curve predictive validity metrics. *Criminal Justice and Behavior, 44*(1), 103-120. http://dx.doi.org/10.1177/0093854816678899

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Kelley, S. M., Ambroziak, G., Thornton, D., & Barahal, R. M. (2020). How do professionals assess sexual recidivism risk? An updated survey of practices. *Sexual Abuse*, *32*(1), 3-29. http://dx.doi.org/10.1177/1079063218800474

*Lee, S. C., Hanson, R. K., Fullmer, N., Neeley, J, & Ramos, K. (2018). *The Predictive validity of Static-99R over 10 years for sexual offenders in California: 2018 update.* State

Authorized Risk Assessment Tools for Sex Offenders (SARATSO). http://saratso.org/pdf/Lee_Hanson_Fullmer_Neeley_Ramos_2018_The_Predictive_Validity_of_S_.pdf

*Lee, S. C., Restrepo, A., Satariano, A, & Hanson, R. K. (2016). *The predictive validity of Static-99R for sex offenders in California: 2016 update.* State Authorized Risk Assessment Tools for Sex Offenders (SARATSO). http://saratso.org/pdf/ThePredictiveValidity_of_Static_99R_forSexualOffenders_inCalifornia_2016v1.pdf

Mann, R. E., Hanson, R. K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. *Sexual Abuse: A Journal of Research and Treatment*, *22*(2), 191-217. http://dx.doi.org/10.1177/1079063210366039

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

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*(4), 944–953. http://dx.doi.org/10.1037/a0028411

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among forensic experts. *Annual Review of Law and Social Science, 11,* 37-55. http://dx.doi.org/10.1146/annurev-lawsocsci-120814-121714

Murrie, D. C., Boccaccini, M. T., Caperton, J. D., & Rufino, K. A. (2012). Field validity of the Psychopathy Checklist-Revised in sex offender risk assessment. *Psychological Assessment*, *24*(2), 524-529. http://dx.doi.org/10.1037/a0026015

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24*(10), 1889-1897. doi:10.1177/0956797613481812

*Myer, A. J. (2019). Examining the predictive validity of the Static-99R on Native American sex offenders. *Justice Evaluation Journal*. https://doi.org/10.1080/24751979.2019.1636614

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*(12), 1406-1421. http://dx.doi.org/10.1177/0093854814548449

Olver, M. E., Klepfisz, G., Stockdale, K. C., Kingston, D. A., Nicholaichuk, T. P., & Wong, S. C. P. (2016). Some notes on the validation of VRS-SO static scores. *Journal of Sexual Aggression, 22*(2), 147-160. http://dx.doi.org/10.1080/13552600.2015.1116625

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

Phenix, A., & Epperson, D. (2015). Overview of the development, reliability, validity, scoring and uses of the Static-99, Static-99R, Static-2002, and Static-2002R. In A. Phenix & H. M. Hoberman, (Eds.), *Sexual offending: Predisposing conditions, assessment and management* (pp. 437-455). New York: Springer.

Phenix, A., Fernandez, Y. M., Harris, A. J. R., Helmus, L. M., Hanson, R. K., & Thornton, D. (2016). *Static-99R coding rules revised – 2016.* Available from www.static99.org

Phenix, A., Helmus, L., & Hanson, R. K. (2016). *Static-99R and Static-2002R Evaluator's Workbook.* Available from www.static99.org

Public Safety Canada. (2018). *2018 Corrections and Conditional Release Statistical Overview.* Ottawa, ON, Canada: Author.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. JSTOR. https://doi.org/10.2307/271063

*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*(12), 1413-1433. http://dx.doi.org/10.1177/0093854813492518

Rettenberger, M., Rice, M. E., Harris, G. T., & Eher, R. (2017). Actuarial risk assessment of sexual offenders: The psychometric properties of the Sex Offender Risk Appraisal Guide (SORAG). *Psychological Assessment, 29,* 624-638. https://doi.org/10.1037/pas0000390

Revelle, W. (2020). *psych: Procedures for personality and psychological research*. Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych Version = 2.0.12.

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*(4), 1085-1094. http://dx.doi.org/10.1037/pas0000009

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*, *Law and Human Behavior, 29*(5), 615-620. http://dx.doi.org/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*(9), 856-861. http://dx.doi.org/10.1016/S0895-4356(03)00124-0

Schulze, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Zeitschrift für Psychologie/Journal of Psychology, 215*, 90-103. http://dx.doi.org/10.1027/0044-3409.215.2.90

Seto, M. C. (2019). The motivation-facilitation model of sexual offending. *Sexual Abuse, 31*(1), 3-24. https://doi.org/10.1177/1079063217720919

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

*Smallbone, S. & Rallings, M. (2013). Short-term predictive validity of the Static-99 and Static-99-R for Indigenous and nonindigenous Australian sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 25*(3), 302-316. http://dx.doi.org/10.1177/1079063212472937

*Smeth, A. (2013). *Evaluating risk assessments among sex offenders: A comparative analysis of static and dynamic factors.* Unpublished MA thesis. Carleton University, Ottawa, Ontario, Canada.

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer.

Tamatea, A. J. (2014). Predictive validity of the Stable-2007: A New Zealand study. *Sexual Abuse in Australia and New Zealand, 6*(1), 57-71.

Therneau, T. (2020). A package for survival analysis in R. Version 3.2-7 (software). http://CRAN.R-project.org/package=survival.Veith, A. (2018). *The predictive validity of the Static-99r and Stable-2007 in a community sample of sex offenders.* Doctoral dissertation, University of North Dakota. *Theses and Dissertations*. 2373. https://commons.und.edu/theses/2373

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://www.jstatsoft.org/v036/i03.

Volinsky, C. T., & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics, 56*(1), 256–262. https://doi.org/10.1111/j.0006-341X.2000.00256.x

Wood, J., Nezworski, M., & Stejskal, W. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*(1), 3-10. https://doi.org/10.1111/j.1467-9280.1996.tb00658.x

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39(12),* 1511-1538. https://doi.org/10.1177/0093854812455741

Table 1

Predictive Accuracy for Static-99R and STABLE-2007 Items, Total Scores, Risk Levels, and Constructs

		Sexual recidivism			Violent recidivism			Any criminal recidivism		
	N	*n* recid	*C*	[95% CI]	*n* recid	*C*	[95% CI]	*n* recid	*C*	[95% CI]
Total scores/levels										
Static-99R total score	4,373	201	.705	[.669, .741]	776	.693	[.675, .711]	1,045	.699	[.683, .715]
Static-99R risk level	4,373	201	.687	[.652, .722]	776	.678	[.662, .694]	1,045	.683	[.669, .697]
STABLE-2007 total score	4,291	200	.670	[.633, .706]	772	.655	[.635, .675]	1046	.667	[.651, .683]
Static/STABLE combined risk level	4,232	199	.694	[.659, .729]	762	.693	[.675, .711]	1,029	.699	[.685, .713]
Static/STABLE constructs										
Sexual criminality	4,215	199	.661	[.618, .704]	757	.517	[.495, .539]	1,023	.534	[.514, .554]
General criminality	4,183	195	.663	[.628, .698]	753	.735	[.717, .753]	1,016	.744	[.730, .758]
Youthful stranger aggression	4,226	199	.645	[.608, .682]	760	.643	[.623, .663]	1,027	.639	[.623, .655]
Static-99R Items										
Age	4,373	201	.584	[.547, .621]	776	.618	[.600, .636]	1,045	.603	[.587, .619]
Never lived with lover	4,370	201	.592	[.557, .627]	775	.569	[.551, .587]	1,044	.566	[.550, .582]
Index non-sexual violence	4,373	201	.527	[.500, .554]	776	.561	[.545, .577]	1,045	.550	[.536, .564]
Prior non-sexual violence	4,373	201	.590	[.555, .625]	776	.664	[.646, .682]	1,045	.658	[.642, .674]
Prior sex offences	4,373	201	.619	[.582, .656]	776	.569	[.551, .587]	1,045	.581	[.565, .597]
4+ prior sentencing occasions	4,373	201	.612	[.577, .647]	776	.648	[.630, .666]	1,045	.663	[.647, .679]
Non-contact sex offence	4,373	201	.575	[.542, .608]	776	.513	[.499, .527]	1,045	.503	[.491, .515]
Unrelated victim	4,373	201	.544	[.513, .575]	776	.530	[.512, .548]	1,045	.534	[.520, .548]
Stranger victim	4,373	201	.584	[.549, .619]	776	.515	[.499, .531]	1,045	.535	[.521, .549]
Male victim	4,373	201	.530	[.505, .555]	776	.510	[.500, .520]	1,045	.508	[.498, .518]

Table continues on next page

Table 1 continued

		Sexual recidivism			Violent recidivism			Any criminal recidivism		
	N	*n* recid	*C*	[95% CI]	*n* recid	*C*	[95% CI]	*n* recid	*C*	[95% CI]
STABLE-2007 Items										
Significant social influences	4,272	199	.553	[.514, .592]	770	.631	[.611, .651]	1043	.637	[.619, .655]
Hostility towards women	4,281	200	.547	[.510, .584]	771	.601	[.581, .621]	1044	.592	[.576, .608]
Loneliness/social rejection	4,280	200	.585	[.550, .620]	770	.553	[.535, .571]	1044	.569	[.553, .585]
Lack of concern towards others	4,280	199	.571	[.534, .608]	769	.602	[.582, .622]	1043	.607	[.591, .623]
Sex as coping	4,276	200	.591	[.554, .628]	768	.496	[.478, .514]	1042	.507	[.491, .523]
Sexual preoccupation	4,279	200	.599	[.560, .638]	768	.507	[.489, .525]	1041	.517	[.501, .533]
Impulsivity	4,271	199	.646	[.611, .681]	769	.663	[.645, .681]	1042	.674	[.658, .690]
Poor cognitive problem-solving	4,272	199	.637	[.600, .674]	769	.646	[.628, .664]	1042	.652	[.636, .668]
Negative emotionality/hostility	4,273	199	.547	[.512, .582]	768	.584	[.564, .604]	1041	.591	[.575, .607]
Cooperation with supervision	4,278	199	.579	[.542, .616]	771	.636	[.616, .656]	1043	.643	[.627, .659]
Emotional identification with kids	2,250	90	.555	[.504, .606]	351	.535	[.515, .555]	477	.526	[.508, .544]
Relationship stability	2,609	103	.618	[.567, .669]	409	.609	[.584, .634]	556	.613	[.591, .635]
Deviant sexual interests	2,609	103	.580	[.525, .635]	409	.520	[.493, .547]	556	.530	[.506, .554]
STABLE-2000 Unique Items										
Emotional identification with kids	1,674	97	.528	[.483, .573]	361	.515	[.493, .537]	488	.510	[.492, .528]
Relationship stability	1,678	97	.568	[.521, .615]	362	.551	[.526, .576]	489	.552	[.530, .574]
Deviant sexual interests	1,670	97	.586	[.529, .643]	358	.515	[.488, .542]	485	.503	[.479, .527]
Sexual entitlement attitudes	1,672	97	.616	[.561, .671]	361	.568	[.541, .595]	488	.572	[.548, .596]
Rape attitudes	1,669	96	.565	[.516, .614]	361	.576	[.551, .601]	488	.574	[.552, .596]
Child molester attitudes	1,665	96	.535	[.486, .584]	361	.515	[.493, .537]	488	.506	[.486, .526]

Table 2

Calibration Analyses (E/O Index)

Risk Level	Expected recidivism		Observed recidivism			E/O	95% CI	
	Proportion	*N* recid (E)	*N*	*N* recid (O)	Prop. recid			
Static-99R – Sex recid								
I Very low	-	1.22	110	6	.05	**0.20**	**[0.09,**	**0.45]**
II Below average	-	8.83	369	3	.01	2.94	[0.95,	9.13]
III Average	-	49.19	824	29	.04	**1.70**	**[1.18,**	**2.44]**
IVa Above average	-	50.17	390	23	.06	**2.18**	**[1.45,**	**3.28]**
IVb Well above average	-	62.93	228	34	.15	**1.85**	**[1.32,**	**2.59]**
All	-	172.34	1921	95	.05	**1.81**	**[1.48,**	**2.22]**
Static/STABLE - Sex								
I Very low	.028	3.56	127	3	.024	1.19	[0.38,	3.68]
II Below average	.053	21.73	410	8	.020	**2.72**	**[1.36,**	**5.43]**
III Average	.075	52.57	701	23	.033	**2.29**	**[1.52,**	**3.44]**
IVa Above average	.136	47.06	346	19	.055	**2.48**	**[1.58,**	**3.88]**
IVb Well above average	.268	82.81	309	41	.133	**2.02**	**[1.49,**	**2.74]**
All		207.73	1,893	94	.050	**2.21**	**[1.81,**	**2.71]**
Static/STABLE - Violence								
I Very low	.028	3.56	127	4	.031	0.89	[0.33,	2.37]
II Below average	.133	54.53	410	22	.054	**2.48**	**[1.63,**	**3.76]**
III Average	.135	94.64	701	122	.174	**0.78**	**[0.65,**	**0.93]**
IVa Above average	.318	110.03	346	87	.251	**1.26**	**[1.03,**	**1.56]**
IVb Well above average	.404	124.84	309	126	.408	0.99	[0.83,	1.18]
All		387.60	1,893	361	.191	1.07	[0.97,	1.19]
Static/STABLE – Any crime								
I Very low	.028	3.56	127	6	.047	0.59	[0.27,	1.32]
II Below average	.192	78.72	410	33	.080	**2.39**	**[1.70,**	**3.36]**
III Average	.208	145.81	701	170	.243	**0.86**	**[0.74,**	**0.997]**
IVa Above average	.467	161.58	346	116	.335	**1.39**	**[1.16,**	**1.67]**
IVb Well above average	.541	167.17	309	171	.553	0.98	[0.84,	1.14]
All		556.84	1,893	496	.262	**1.12**	**[1.03,**	**1.23]**

Note: Prop. = proportion. E/O values in bold are statistically significant (p <.05)

Table 3

Incremental Validity Analyses from Cox Regression (N = 4,175)

Predictor	*N* recid	Hazard ratio	[95% CI]	*C*	[95% CI]	Null BIC	Model BIC	Model 1 BIC – Model 2 BIC
Sexual Recidivism								
Model 1: Static/STABLE	195			.716	[.679, .753]	3,139.97	3,035.75	
Static-99R		1.255	[1.178, 1.338]					
STABLE-2007		1.050	[1.019, 1.082]					
Model 2: Constructs	195			.719	[.684, .754]	3,139.97	3,041.93	-6.18
Sexual criminality		1.245	[1.161, 1.335]					
General criminality		1.128	[1.056, 1.205]					
Youthful stranger aggression		1.191	[1.103, 1.287]					
Violent Recidivism								
Model 1: Static/STABLE	749			.706	[.688, .724]	11,977.90	11,646.34	
Static-99R		1.221	[1.182, 1.260]					
STABLE-2007		1.045	[1.030, 1.062]					
Model 2: Constructs	749			.753	[.737, .769]	11,977.90	11,407.00	239.34
Sexual criminality		0.858	[0.820, 0.898]					
General criminality		1.405	[1.360, 1.451]					
Youthful stranger aggression		1.202	[1.155, 1.251]					
Any Crime Recidivism								
Model 1: Static/STABLE	1,012			.716	[.700, .732]	16,204.89	15,714.03	
Static-99R		1.228	[1.192, 1.259]					
STABLE-2007		1.054	[1.040, 1.068]					
Model 2: Constructs	1,012			.757	[.743, .771]	16,204.89	15,440.62	273.41
Sexual criminality		0.895	[0.861, 0.930]					
General criminality		1.414	[1.375, 1.454]					
Youthful stranger aggression		1.186	[1.146, 1.227]					

Note. All models demonstrated very strong evidence of model fit compared to the null model (difference between Model BIC and Null BIC ranged between 98 and 764).

Table 4

Field Validity Studies of Static-99R – Predictive Accuracy for Sexual Recidivism

			Discrimination Accuracy			Comparing Expected (Norms) to Observed Number of Recidivists		
	n recid / Total	Recid rate (%)	AUC/ C	SE	[95% CI]	E/O *N*'s	E/O Index	[95% CI]
Boccaccini et al. (2017)[a]	847/15,680	5.4%	.633	.0096	[.61, .65]	1,080.6/568[c]	1.90	[1.75, 2.07]
Boccaccini et al. (2017)[b]	399/19,007	2.1%	.667	.0137	[.64, .69]	251.6/123[c]	2.05	[1.71, 2.44]
Buttars et al. (2015)	65/1,437	4.5%	.608	*(.0342)*	- -			
Duwe & Rocque (2018)	26/650	4.0%	.682	*(.0574)*	[.569, .794]	-	-	- -
Hanson et al. (2015)	83/764	10.9%	.734	*(.0309)*	[.674, .795]	-	-	- -
Hanson et al. (2020)	13/565	2.3%	.596	.089	[.423, .770]	-	-	- -
Helmus et al. (2021) – Current study	201/4,373	4.6%	.705	.0186	[.669, .741]	172.34/95	1.81	[1.48, 2.22]
Lee et al. (2018)	23/371	6.2%	.806	*(.0536)*	[.701, .911]	30.79/23	1.34	[0.89, 2.01]
Lee et al. (2016)	78/1,626	4.8%	.756	*(.0281)*	[.701, .811]	133.4/78	1.71	[1.37, 2.13]
Myer (2019)	118/986	12.0%	.599	*(.0270)*	- -	-	-	- -
Olver et al. (2014)	42/673	6.2%	.71	*(.0332)*	[.65, .78]			
Rettenberger et al. (2013)	71/1,077	6.6%	.71	*(.0306)*	[.65, .77]	-	-	- -
Smallbone & Rallings (2013)	19/399	4.8%	.77	*(.0485)*	[.67, .86]			
Smeth (2013) Study 1	9/210	4.4%	.69	.08	[.53, .85]			
Smeth (2013) Study 2	15/207	7.2%	.63	.10	[.43, .83]			
Combined	2,009/48,025	4.2%	-	-	- -	1,669/887	1.88	[1.76, 2.01]

Note. AUC standard errors in parentheses were estimated from confidence interval data reported in the study. Where possible we reported numbers to three significant figures, but sometimes articles reported less than that.

[a]refers to cases scored pre-2004 whereas [b] refers to cases scored 2004 onwards.

[c]For Boccaccini et al. (2017), the AUCs are based on all follow-up data (*M* = 5.2 years) whereas the E/O indices are based on cases with fixed 5-year follow-up data (*N*'s = 14,077 cases pre-2004 and 3,378 cases from 2004 onwards).

Table 5

Meta-analysis of Field Validity Studies of Static-99R and STABLE-2007 Predicting Sexual Recidivism

				Fixed-effect		Random-effects		Variability measures		
	k	*N* recid	*N*	AUC/C	95% CI	AUC/C	95% CI	*Q*	*p*	I^2
Static-99R										
Before current study	14	1,808	43,652	.660	[.648, .673]	.686	[.651, .721]	51.4	<.001	74.7
Adding current study	15	2,009	48,925	.665	[.653, .677]	.688	[.656, .719]	56.5	<.001	75.2
Removing Texas (current study still included)	13	763	13,338	.697	[.678, .715]	.697	[.661, .734]	34.5	.001	65.2
No author involvement	8	365	5,639	.665	[.638, .692]	.673	[.626, .720]	17.7	.013	60.6
Author involvement	7	1,644	42,386	.665	[.652, .679]	.700	[.656, .745]	38.7	<.001	84.5
Appropriately trained	6	498	8,884	.724	[.701, .746]	.724	[.701, .746]	5.2	..395	3.3
Training unknown	9	1,511	39,141	.643	[.629, .657]	.647	[.618, .675]	15.8	.045	49.4
STABLE-2007										
Before current study	3	129	1,342	.649	[.601, .697]	.649	[.601, .697]	0.6	.735	0.0
Adding current study	4	329	5,633	.663	[.634, .691]	.663	[.634, .691]	1.1	.780	0.0

Online Supplement A

Descriptive Information for Items, Constructs, and Scales

	N						No sexual recidivism			Sexual recidivism	
Total scores/levels -				*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Static-99R total score	4,373			2.42	2.48	4172	2.33	2.44	201	4.21	2.68
Static-99R risk level (from 1 to 5)	4,373			3.14	1.00	4172	3.11	0.98	201	3.80	1.02
STABLE-2007 total score	4,291			7.47	4.81	4091	7.33	4.74	200	10.36	5.25
Static/STABLE combined risk level (from 1 to 5)	4,232			3.15	1.10	4033	3.11	1.09	199	3.90	1.03
Static/STABLE constructs											
Sexual criminality	4,215			1.59	1.61	4016	1.53	1.56	199	2.67	2.07
General criminality	4,183			2.57	2.11	3988	2.51	2.09	195	3.75	2.24
Youthful stranger aggression	4,226			1.68	2.01	4027	1.63	2.01	199	2.55	1.92
Static-99R Items											
Age		-3	-1	0	1						
	4,373	438	1,754	568	1,613	4172	-0.347	1.250	201	-0.045	1.205
				0	1						
Never lived with lover	4,370			2,824	1,546	4169	0.345	0.476	201	0.527	0.500
Index non-sexual violence	4,373			3,756	617	4172	0.138	0.346	201	0.194	0.396
Prior non-sexual violence	4,373			2,963	1,410	4172	0.314	0.464	201	0.502	0.501
Prior sex offences		0	1	2	3						
	4,373	3,210	743	306	114	4172	0.366	0.706	201	0.836	1.038
				0	1						
4+ prior sentencing occasions	4,373			2,908	1,465	4172	0.325	0.468	201	0.547	0.499
Non-contact sex offence	4,373			3,630	743	4172	0.164	0.370	201	0.298	0.459
Unrelated victim	4,373			1,507	2,866	4172	0.652	0.476	201	0.731	0.444
Stranger victim	4,373			3,189	1,184	4172	0.263	0.440	201	0.428	0.496
Male victim	4,373			3,881	492	4172	0.109	0.312	201	0.184	0.389
STABLE-2007 Items			0	1	2						
Significant social influences	4,272		2,228	1,445	599	4073	0.611	0.714	199	0.769	0.802
Hostility towards women	4,281		2,735	1,253	293	4081	0.423	0.615	200	0.555	0.670

Loneliness/social rejection	4,280	1,849	1,932	499	4080	0.674	0.669	200	0.890	0.671
Lack of concern towards others	4,280	2,686	1,178	416	4081	0.459	0.659	199	0.678	0.763
Sex as coping	4,276	2,924	1,020	332	4076	0.381	0.619	200	0.650	0.742
Sexual preoccupation	4,279	2,658	1,289	332	4079	0.444	0.627	200	0.715	0.746
Impulsivity	4,271	1,917	1,754	600	4072	0.674	0.698	199	1.055	0.712
Poor cognitive problem-solving	4,272	1,546	2,055	671	4073	0.778	0.684	199	1.146	0.727
Negative emotionality/hostility	4,273	2,883	1,027	363	7074	0.404	0.639	199	0.528	0.688
Cooperation with supervision	4,278	2,907	1,017	354	4079	0.393	0.631	199	0.618	0.728
Emotional identification with kids	2,250	1,762	414	74	2160	0.243	0.495	90	0.422	0.653
Relationship stability	2,609	608	1,156	845	2506	1.078	0.740	103	1.398	0.691
Deviant sexual interests	2,609	964	1,156	489	2506	0.808	0.721	103	1.058	0.752
STABLE-2000 Unique Items										
Emotional identification with kids	1,674	1,291	315	68	1577	0.265	0.523	97	0.340	0.593
Relationship stability	1,678	485	221	972	1581	1.277	0.888	97	1.505	0.831
Deviant sexual interests	1,670	924	478	268	1573	0.591	0.737	97	0.876	0.869
Sexual entitlement attitudes	1,672	1,010	568	94	1575	0.433	0.585	97	0.763	0.747
Rape attitudes	1,669	1,274	346	49	1573	0.256	0.492	96	0.438	0.646
Child molester attitudes	1,665	1,255	346	64	1569	0.278	0.521	96	0.396	0.657

.

Online Supplement B

Kendall's Tau-b Correlations Between Scales and Constructs

	STABLE-2007	Sexual criminality	General criminality	Youthful stranger aggression
Static-99R Total	.332	.280	.402	.675
	(*n* = 4,232)	(*n* = 4,215)	(*n* = 4,183)	(*n* = 4,226)
STABLE-2007 Total		.445	.635	.237
		(*n* = 4,215)	(*n* = 4,183)	(*n* = 4,226)
Sexual criminality			.216	.058
			(*n* = 4,176)	(*n* = 4,213)
General criminality				.183
				(*n* = 4,181)

Note. All correlations are significant at $p < .001$

Online Supplement C

Field Validity Studies of Static-99R – Descriptive Information

	Largest *N*	Location	Sample	Release Timeframe	Follow-up (Years)	Notes
Boccaccini et al. (2017)	34,687	Texas	Routine	1999-2011	*M* = 5.2	*N* varies by analyses; results separated pre/post 2003 given significant differences in accuracy
Buttars et al. (2015)	1,437	Iowa	Routine	2002-2007	5 (fixed)	Location not confirmed but likely Iowa
Duwe & Rocque (2018)	650	Minnesota	Routine	2012	4 (fixed)	Staff trained in 99R (unclear by who)
Hanson et al. (2015)	764	Canada	Routine	2001-2005	*M* = 7.4	Officers trained by scale co-developer
Hanson et al. (2020)	565	Arizona & New York	Probationers	2011-2015	*M* = 3.3	104 of 159 recidivism incidents were of 'unknown' type
Helmus et al (2021) – Current study	4,373	Canada	Routine	2004-2013	*M* = 4.5	Community supervision officers underwent certified training
Lee et al. (2018)	371	California	Parolees	2006-2007	5 (fixed)	Trained by certified trainers
Lee et al. (2016)	1,626	California	Routine	2009-2010	5 (fixed)	Trained by certified trainers
Myer (2019)	986	Midwest state	Routine	2005-2014	5 (fixed)	Staff trained (unclear by who). Unclear if Static-99 and 99R were combined
Olver et al. (2014)	673	Canada	Prison treatment	2000-2008[a]	*M* = 6.3	Trained staff for National Sex Offender Program (multi-intensity treatment)
Rettenberger et al. (2013)	1,077	Austria	Routine inmates	2001-2007	*M* = 6.4	Psychologists/psychiatrists trained by certified trainers
Smallbone & Rallings (2013)	399	Australia	1+ years custodial sentence	-	*M* = 2.4	Staff trained (unclear by who).
Smeth (2013) Sample 1	210	Iowa	Unknown	-	*M* = 1.8	7 recidivists excluded because recidivism happened prior to dynamic assessment
Smeth (2013) Sample 2	207	Iowa	Unknown	-	*M* = 3.5	10 recidivists excluded because recidivism happened prior to STABLE-2007 assessment

Note. All studies reported results for sexual rearrest, except Rettenberger et al., (2013), who defined recidivism as sexual reconviction. Some studies reported both outcomes; in these cases, the broader outcome of rearrest was used. [a] Some cases may have been released in 2009 – 2011.

Online Supplement D

Field Validity Studies of STABLE-2007 Predicting Sexual Recidivism

	Location	Release Timeframe	Recidivism criteria	Follow-up (Years)	*n* recid / Total	Recid rate (%)	AUC/C	SE	[95% CI]
Etzler et al. (2020)	Austria	2001-2011	Reconviction	5 (fixed)	48/520	9.2%	.638	*(.0398)*	[.560, .716]
Hanson et al. (2015)	Canada	2001-2005	Rearrest	*M* = 7.4	66/615[a]	10.8%	.669	*(.0362)*	[.598, .740]
Helmus et al. (2021) - Current study	Canada	2004-2013	Charge	*M* = 4.5	200/4,291	4.7%	.670	.0185	[.633, .706]
Smeth (2013)	Iowa	-	Reconviction	*M* = 3.5	15/207	7.2%	.62	.06	[.49, .74]

Note. AUC standard errors in parentheses were estimated from confidence interval data reported in the study. Where possible we reported numbers to three significant figures, but sometimes articles reported less than that.
[a]Number of recidivists was estimated based on overall base rate for larger sample