Psychological Assessment
2017, Vol. 29, No. 6, 611–623

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000377

Field Validity of Static-99/R Scores in a Statewide Sample of 34,687 Convicted Sexual Offenders

Marcus T. Boccaccini and Amanda K. Rice
Sam Houston State University

L. Maaike Helmus
Global Institute of Forensic Research, LLC,
Great Falls, Virginia

Daniel C. Murrie
University of Virginia

Paige B. Harris
Sam Houston State University

The Static-99 (and revision, the Static-99R) reflect the most researched and widely used approach to sex offender risk assessment. Because the measure is so widely applied in jurisdictions beyond those on which it was developed, it becomes crucial to examine its field validity and the degree to which published norms and recidivism rates apply to other jurisdictions. We present a new and greatly expanded field study of the predictive validity (M = 5.23 years follow-up) of the Static-99 as applied system-wide in Texas (N = 34,687). Results revealed stronger predictive validity than a prior Texas field study, especially among offenders scored after the release of an updated scoring manual in 2003 (AUC = .66 to .67, d = .65 to .69), when field reliability was also stronger. But calibration analyses revealed that the Static-99R routine sample norms led to a significant overestimation of risk in Texas, especially for offenders with scores ranging from 1 to 5. We used logistic regression to develop local Texas recidivism norms (with confidence intervals) for Static-99R scores. Overall, findings highlight the importance of revisiting and updating field study findings, and the potential benefits of using statewide data to develop local norms.

Keywords: Static-99, Static-99R, field validity, sex offender, risk assessment, local norms, calibration

The 10-item Static-99 (Hanson & Thornton, 2000) is the most widely used measure for assessing risk for sexual recidivism among sexual offenders (Chevalier, Boccaccini, Murrie, & Varela, 2015; Neal & Grisso, 2014). Static-99 items rely on offender and offense information available in offender files, such as age at release, prior sentencing dates, prior sexual offenses, prior non-sexual violence convictions, and victim characteristics (stranger victim, male victim, unrelated victim). Although there is now a revised version of the Static-99 (Static-99R; Helmus, Thornton, Hanson, & Babchishin, 2012), the two measures include the same 10 items and differ only in the scores assigned for the age-at-release item.

In addition to being widely implemented, the Static-99 is also the most widely researched sex offender risk assessment scale. Hanson and Morton-Bourgon's (2009) meta-analysis included 63 validation studies of the scale, which was approximately twice the number of validation studies for the next most researched scale, and roughly six times the number of validation studies for most other sex offender risk scales. However, all but one of these studies examined Static-99 scores assigned for research purposes (typically by research assistants or psychologists conducting retrospective evaluations). There was a single field study (i.e., in which the scale was scored and used as part of routine practice) in an unpublished government report from Canada, which found high predictive accuracy (AUC = .74) for Static-99 scores assigned by community supervision officers trained by one of the scale developers (Hanson, Harris, Scott, & Helmus, 2007). The Static-99 scores included in this lone field study were from assessments voluntarily submitted by community supervision officers; it was not a comprehensive examination of a jurisdiction-wide implementation.

In 2009, we published the first field validity study of Static-99 scores in the United States (Boccaccini, Murrie, Caperton, &

Editor's Note. Yossef S. Ben-Porath, PhD, served as the sole action editor for this submission.

Marcus T. Boccaccini and Amanda K. Rice, Department of Psychology and Philosophy, Sam Houston State University; L. Maaike Helmus, Global Institute of Forensic Research, LLC, Great Falls, Virginia; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia; Paige B. Harris, Department of Psychology and Philosophy, Sam Houston State University.

The research contained in this document was coordinated in part by the Texas Department of Criminal Justice (587-AR09). The contents of this document reflect the views of the authors and do not necessarily reflect the views or policies of the Texas Department of Criminal Justice. L. Maaike Helmus is a coauthor of Static-99R and a certified trainer for the tool. The copyright for this instrument is held by the Government of Canada. In her current position, L. Maaike Helmus does not receive any income from trainings or use of the scale.

Correspondence concerning this article should be addressed to Marcus T. Boccaccini, Department of Psychology and Philosophy, Sam Houston State University, Huntsville, TX 77341. E-mail: boccaccini@shsu.edu

Hawes, 2009). It was also the first study of a jurisdiction-wide implementation of the Static-99 for real-world decision making. The 1,928 offenders in the study had been released from custody after they were scored on the Static-99 as part of Texas's screening process for identifying offenders who might meet criteria for civil commitment as sexually violent predators (SVPs). The Static-99 scores had been assigned by correctional staff, including master's level mental health clinicians and parole officers. Overall, Static-99 scores were much weaker predictors of future sexual offending than expected. The 2009 meta-analysis of Static-99 scores reported a median predictive effect of $d = .74$ ($M = .67$, 95% CI [.62, .72]) across 63 studies (Hanson & Morton-Bourgon, 2009), suggesting that we should have found a predictive effect in this moderate to large effect-size range. But the predictive effect for Static-99 scores in the Texas field study was only $d = .36$ (AUC = .60), well outside the 95% confidence interval from the meta-analysis. The predictive effect was somewhat stronger among offenders who had been discharged than among those released under mandatory supervision, but even this effect ($d = .49$, AUC = .63) was outside the meta-analytic confidence interval.

Because the Static-99 did not perform as expected in the Texas field validity sample, we provided "local" recidivism rate norms for both Static-99 and Static-99R scores in Texas, using only the 847 offenders who had been released for at least five years (Boccaccini et al., 2009; Boccaccini & Murrie, 2014). These Texas norms suggested that recidivism rates for each possible score were lower in Texas than in the Static-99 and Static-99R normative samples, although recidivism rates did tend to increase as scores increased. But the Texas norms were based on observed sexual recidivism rates, included only 30 recidivists, and did not include confidence intervals, raising questions about their stability. The Static-99 developers have recommended that recidivism norms be based on logistic regression results (Hanson, Helmus, & Thornton, 2010) from samples with at least 100 recidivists to allow for the calculation of stable recidivism estimates (see Hanson, Lunetta, Phenix, Neeley, & Epperson, 2014; Vergouwe, Steyerberg, Eijkemans, & Habbema, 2005).

Possible Explanations for Texas Field Study Effects

There are several possible explanations for the smaller than expected Texas field study effects. It may be that smaller predictive effects are the norm in field settings, where records can be messy and incomplete, heavy workloads and deadlines can lead to hasty scoring practices, and there is no researcher oversight to ensure rater reliability and adherence to scoring guidelines. It is also possible that scores assigned by parole officers and other types of correctional staff may be less accurate than those assigned by researchers and research assistants, who have completed or are in the process of completing more advanced assessment training. Researchers may have specialized training in the risk scales and complete regular reliability checks, whereas training and scoring practices may drift in real-world settings. For example, staff who have received thorough training may be in charge of training newer staff, but staff turnover can lead to situations where none of the remaining staff members completed the original training. These realities of field work might lead to more measurement error in field scores, which would lead to weaker predictive effects.

It could also be that the weaker than expected field study effects were attributable to the Texas sample differing in important ways from other Static-99 samples, including those used to develop and subsequently study the measure. The Static-99 developers selected items based on research with offenders from Europe and Canada (see Hanson & Thornton, 2000). Although a more recent meta-analysis providing support for the Static-99 items included data from US samples, these were still a small minority (Helmus & Thornton, 2015). Static-99 scores may be less predictive in US samples because the developmental samples did not include offenders from many of the racial and ethnic backgrounds in Texas and other US states (see Leguizamo, Lee, Jeglic, & Calkins, 2015; Varela, Boccaccini, Murrie, Caperton, & Gonzalez, 2013).

The Texas sample was also unique among Static-99 studies in that the sample included only offenders being screened for SVP civil commitment. In Texas, only offenders with two or more qualifying (i.e., contact) sexual offenses are eligible for commitment. It could be that the select nature of the Texas sample led to the weaker than expected predictive effects. The Texas sample included only ostensibly high-risk offenders (i.e., only those with two or more convictions), but excluded those who were presumably at the highest risk for reoffending (i.e., SVP committed), which may have led to range restriction in risk. Indeed, researchers recently reported weaker predictive accuracy for Static-99R scores in samples preselected to exclude lower-risk offenders (Hanson, Thornton, Helmus, & Babchishin, 2016). A more generalizable method for studying the field validity of Static-99 scores would be to focus on a broad sample of sexual offenders that is less likely to be affected by a preselection process.

More recent Static-99 and Static-99R field studies rule out some of these possible explanations for the Texas field study effects. Researchers have reported moderate to strong predictive effects (AUC = .71 to .82) for Static-99 and Static-99R scores assigned as part of routine practice in US, European, and Canadian samples, showing that larger predictive effects are possible in field settings (see Hanson, Helmus, & Harris, 2015; Hanson et al., 2014; Rettenberger, Haubner-Maclean, & Eher, 2013). Some of these studies used scores assigned by parole, probation, or supervision officers (Hanson et al., 2014, 2015), and some used scores assigned by mental health clinicians (Rettenberger et al., 2013), showing that scores from many different types of professionals can be predictive of future offending.

But more recent studies also suggest that poor field reliability and racial/ethnic diversity may be plausible explanations for the weaker Texas field validity effects. Texas administers the Static-99 (currently the Static-99R) to all offenders who may need to register as sexual offenders after release, and about half of these offenders are scored on two or more occasions while incarcerated (Rice, Boccaccini, Harris, & Hawes, 2014). In a large field reliability study ($N = 21{,}983$) of Texas Static-99 scores, rater-agreement was significantly stronger among offenders scored after the release of a new Static-99 scoring manual (Harris, Phenix, Hanson, & Thornton, 2003) than before the release of the new manual ($ICC_{A,1} = .88$ vs. .73; Rice et al.,

2014).[1] The new coding manual contained over 50 pages of detailed text and examples to provide guidance on how to score each item and interpret the overall scale, whereas prior to this version, scoring guidance consisted of less than 2 pages of material in an Appendix to the development study (Hanson & Thornton, 1999). The new coding manual was intended to increase clarity, particularly focusing on areas of common confusion among raters.

These Texas field reliability findings suggest that there was more measurement error in earlier scores, which may help to explain the weaker than expected 2009 field validity findings. Indeed, 74.3% (n = 1,422) of the 1,928 offenders in the 2009 field validity study had been released before 2004, suggesting that they had been scored before the release of the new scoring manual (offenders are screened for SVP evaluations approximately 16 months before their anticipated release date). It may be that predictive validity for Static-99 scores has also improved over time in Texas, with stronger predictive effects for scores assigned in 2004 or later.

When it was published, the Texas field validity study included data from more Latino offenders than all other Static-99 studies combined (Varela et al., 2013). Static-99 scores were not statistically significant predictors of recidivism among Latino offenders (n = 588) in the Texas field study (AUC = .53 to .59), raising concerns about its use with Latino offenders (Varela et al., 2013). Although findings from a recent field study from California were more optimistic, with high predictive accuracy for Static-99 and Static-99R scores among Latino offenders (AUCs of .73 to 75; Hanson et al., 2014), the sample size (n = 200 Latino offenders) was insufficient to alleviate the concerns raised by the Texas data.

Because Static-99 items are based on historical information, one possible explanation for the weaker predictive effect among Latino offenders is that evaluators did not have access to necessary historical information for offenders who had immigrated to the US. Although there was no way to identify offenders born outside the US in the original Texas field validity study, researchers studying Static-99 scores from New Jersey calculated predictive effects separately for Latino offenders born outside the US (n = 268, 6 recidivists) and those born in the US or Puerto Rico (n = 215, 3 recidivists; Leguizamo et al., 2015). Although their small sample size and low recidivism rates mean that their analyses should be interpreted cautiously, they found that Static-99 and Static-99R scores were predictive of postrelease sexual offending among Latino offenders born in the US or Puerto Rico (AUC = .77 and .82), but not among those born outside the US (AUC = .47 and .52). Consequently, the relevant issue here may not be ethnicity per se, but rather the quality and accessibility of relevant information for scoring the instrument.

Current Study

The purpose of the current study was to revisit the field validity of Static-99 and Static-99R scores in Texas, using a much larger and representative sample of 34,687 sexual offenders released from custody after being scored on the Static-99. This sample includes all offenders released after being scored on the Static-99. This study is, by far, the largest Static-99 validity study ever conducted. Indeed, this sample is much larger than the combined sample used to develop the current Static-99R norms for routine correctional samples (N = 4,325 offenders from 10 samples; Hanson et al., 2016).

Our primary goal was to examine whether predictive effects for Static-99 and Static-99R scores in this large and representative sample were larger and more comparable to those from other field studies (e.g., Hanson et al., 2014) than the prior Texas field study, which only examined scores for offenders who met minimum statutory requirements for SVP commitment (Boccaccini et al., 2009). Because there appears to have been a significant improvement in the field reliability of Texas Static-99 scores beginning in 2004 (Rice et al., 2014)—after the release of the updated Static-99 scoring manual (Harris et al., 2003)—we were also interested in examining whether there was an accompanying improvement in predictive validity beginning in 2004. Another goal was to examine whether the previously documented weaker predictive effect for Static-99 scores among Latino offenders could be attributable to especially poor performance among Latino offenders born outside the US (Leguizamo et al., 2015). Our sample includes far more Latino offenders (n = 8,939) and recidivists (n = 268) than any prior study.

Finally, another goal of this study was to examine not just how well Static-99 and Static-99R scores discriminated among recidivists and nonrecidivists (as assessed by predictive accuracy statistics such as AUCs and Cohen's d), but also to examine the often-neglected area of calibration accuracy (Helmus & Babchishin, in press). Calibration (described further in the methods section) assesses the extent to which the expected recidivism rates (e.g., the norms for Static-99R) generalize in validation studies. This type of predictive accuracy is distinct from discrimination; it is possible to have excellent discrimination but poor calibration. Given survey results that 83% of experts who use the Static-99R in SVP evaluations report recidivism rate estimates (Chevalier et al., 2015), it is important to know how well those estimates generalize to various jurisdictions. For example, if the estimated sexual recidivism rate for offenders with a Static-99R score of 4 is 11% (Phenix, Helmus, & Hanson, 2015), then validation studies should find roughly 11% recidivism rates for offenders with that score. Calibration analyses provide a useful framework for comparing observed and expected (i.e., normative) recidivism rates. If the calibration analyses suggest that the Texas data are inconsistent with the current Static-99R norms, we will develop an updated set of local Texas norms with confidence intervals, using the same type of logistic regression analyses used to develop the current Static-99R norms.

Method

Study Sample

The Texas Department of Criminal Justice (TDCJ) provided information from an existing database of Static-99 scores assigned between September, 1999 and June, 2011. The database included the offender's TDCJ identification number, Department of Public Safety (DPS) identification number, the Static-99 total score or

[1] Rice et al. (2014) reported the $ICC_{A,1}$ = .88 (95% CI [.81, .82]) value for post-2003 scores (n = 12,332). We calculated the $ICC_{A,1}$ = .73 (95% CI [.72, .75]) value for pre-2004 scores (n = 9,651) for this article.

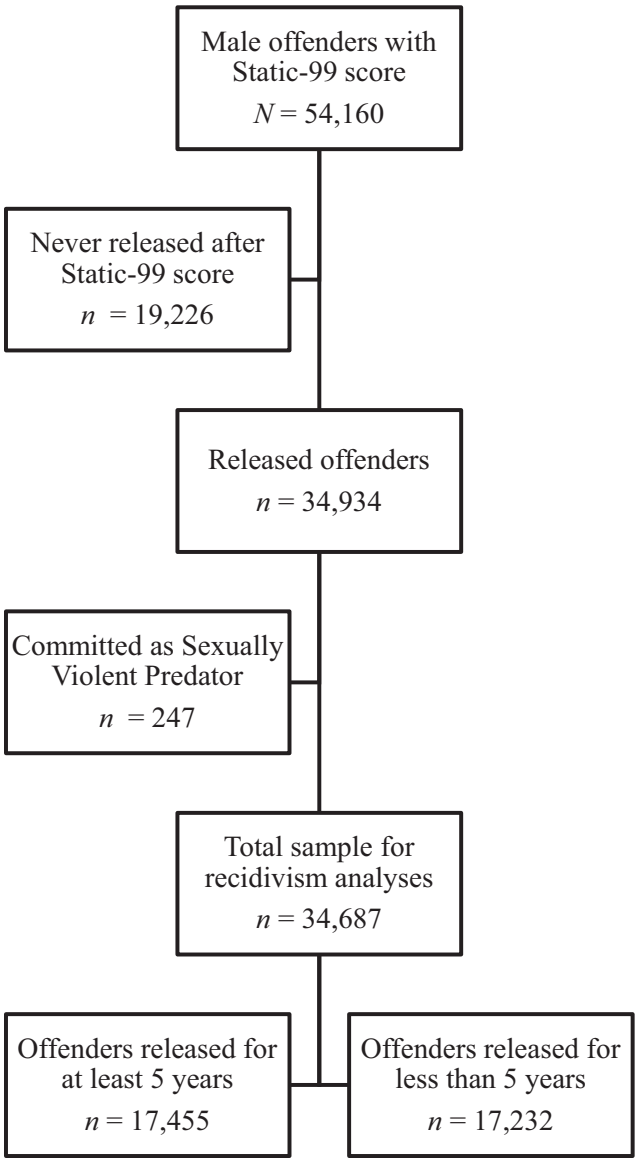

Figure 1. Flowchart of study sample.

scores assigned to the offender, the date of each Static-99 score, and the Static-99 rater's name. There were 54,160 male offenders with at least one Static-99 score in this database (see Figure 1), 34,934 of whom had been released before we collected recidivism data (June, 2011). The remaining 19,226 offenders either died before release (n = 646) or were never released after being scored on the Static-99 (n = 18,580). Of the remaining 34,943 released offenders, 247 (0.7%) were civilly committed and released to the state's SVP program. Our recidivism analyses focus on the 34,687 released offenders who had not been civilly committed.[2]

Although some offenders were released on two or more occasions after being scored on the Static-99 (n = 4,208, 12.1%), we used scores and release dates from the offenders' earliest incarceration period to allow for the longest follow-up period. Offenders had been released for an average of 5.23 (SD = 3.18) years at the time of recidivism data collection, and there were 17,455 offenders who had at least 5-years of follow-up time (i.e., time between release and collection of recidivism data). The mean age at time of release was 39.94 years (SD = 11.65). Most of the offenders had been discharged (n = 15,438, 44.5%), or released on mandatory supervision or parole (n = 14,292, 41.2%). Other offenders had been discharged with a detainer (n = 2,817, 8.1%), released on mandatory supervision or parole with a detainer (n = 2,030, 5.9%), or released in some other less common way (e.g., shock probation, pardon, probation from boot camp; n = 110, 0.3%). An offender released with a detainer may have charges pending in another jurisdiction when released, and the jurisdiction must be notified when the offender is released. Because we have no way of knowing what happened to these offenders because of their detainers (e.g., dropped charges, probation, fine, conviction) and we know that some of these offenders were arrested for new sexual offenses, we included them in the recidivism analyses.

We obtained information about race, ethnicity, and country of birth from DPS records. Although offenders identified as White are supposed to be further identified as Hispanic or non-Hispanic, this information was missing for 7,903 (31.9%) of the White offenders. A review of the offenders' names suggested that some of these offenders were Latino, but that many others were not. Although we examine predictive effects separately for this subgroup of offenders, we did not include any of them in our comparison of effects for Latino and non-Latino offenders. The remaining offenders were identified as Black (African American; n = 9,725, 28.0%), White Hispanic (Latino; n = 8,939, 25.8%), White non-Hispanic (Caucasian; n = 7,938, 22.9%), or as being from a different racial/ethnic background (n = 182, 0.5%). Of the 8,939 Latino offenders, 6,337 (70.9%) were identified in their DPS records as having been born inside the US, 2,459 (27.5%) were identified as having been born outside the US, and 143 (1.6%) had no information about their country of birth.

Measures

Static-99 and Static-99R. The Static-99 (Hanson & Thornton, 2000) is an empirical-actuarial risk assessment tool designed to predict sexual recidivism among adult male sex offenders who have been charged with or convicted of an offense that is judged to have a sexual motivation (see www.static99.org). The scale includes 10 items that broadly assess age, criminal history, victim characteristics, and relationship history. Most items are dichotomous, although prior sex offenses and age-at-release are worth more points given their stronger relationship to recidivism. Static-99 total scores are computed by summing all items, and they range from 0 to 12. Based on the total score, offenders can be assigned to one of the following risk categories: low (0 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6+). The estimated median score for the Static-99 is 2 (Hanson, Lloyd, Helmus, & Thornton, 2012).

The Static-99R (Helmus, Thornton et al., 2012) is identical to the Static-99 with the exception of an updated age-at-release item designed to better reflect reductions in risk among offenders age 40 and above. Thus, an offender may have different scores, cor-

[2] Mean Static-99 scores were marginally lower (d = −.11) among those included in the study (M = 2.47, SD = 1.62, n = 34,687) than among those excluded (M = 2.67, SD = 1.78, n = 19,473).

responding with different levels of risk, on the two measures. Static-99R total scores can range from −3 to 12 and are used to place offenders in one of five risk categories (Hanson, Babchishin, Helmus, Thornton, & Phenix, in press): Category I (−3 to −2), Category II (−1 to 0), Category III (1 to 3), Category IVa (4 to 5) and Category IVb (6+). The estimated median Static-99R score is 2 (Hanson et al., 2012). Meta-analyses have found that Static-99 (d = .67; Hanson & Morton-Bourgon, 2009) and Static-99R (AUC = .69; Helmus, Hanson, Thornton, Babchishin, & Harris, 2012) scores have moderate and similar predictive accuracy for sexual recidivism.

The TDCJ database contained Static-99 scores assigned for various purposes, including risk-level determination, prerelease evaluations, parole evaluations, program entry, and civil commitment screening. Many of the offenders had been scored on multiple occasions before being released (n = 12,800, 36.9%). When an offender had multiple scores assigned during their first period of incarceration, we used the Static-99 score that was assigned closest in time to his release from incarceration. The offender's latest Static-99 score would likely reflect any changes in anticipated age-at-release and any new charges or convictions an offender received during incarceration; therefore, the latest risk assessment score should be based upon the most accurate information. The average Static-99 score among offenders was 2.47 (SD = 1.62).

We transformed Static-99 total scores into Static-99R total scores using offender birth and release dates. We used these dates to calculate the offender's score on the Static-99 and Static-99R age-at-release items, subtracted the Static-99 age-at-release item score from the Static-99 total score, and then added the Static-99R age-at-release item score to that value to obtain a Static-99R total score. The mean Static-99R score was 2.18 (SD = 2.02), which is slightly lower than the average Static-99R score among routine correctional samples used in the normative data for the scale (M = 2.5, 95% CI of 2.1 to 2.9; Hanson et al., 2016).

The Static-99 scores had been assigned by various types of TDCJ employees, including psychologists, parole officers, and administrative staff. More than 600 different evaluators scored at least one offender on the Static-99. The education of these evaluators range from bachelor's degrees to doctoral degrees, and their training and areas of profession vary. Most evaluators have been trained internally by senior TDCJ staff members (see Boccaccini et al., 2012); however, we have no information about the experience or training of any individual evaluator. Despite the number of different evaluators, only about 3% of the variance in Static-99 scores in this sample is attributable to evaluator differences (Rice, Boccaccini, & Luna-Collier, 2012). Rice et al. (2014) examined Static-99 score rater-agreement among 21,983 Texas offenders who had been scored on multiple occasions during the same period of incarceration. Rater-agreement was good for Static-99 total scores ($ICC_{A,1}$ = .81), and stronger for scores assigned after the release of the 2003 scoring manual ($ICC_{A,1}$ = .88) than before the release of the new manual ($ICC_{A,1}$ = .73).

Recidivism. We obtained postrelease arrest data for each offender from the Texas DPS and release dates from TDCJ. DPS provided us with an electronic data file that included an arrest date and a National Crime Information Center offense-type code for each arrest. We defined recidivism as any postrelease arrest for a sexual offense. In the normative data for Static-99R, roughly half the samples used arrests or charges as the recidivism criterion, and half used convictions (Hanson et al., 2016). The rationale for combining results across these studies was that differences in the recidivism criterion did not meaningfully explain differences in recidivism rates across samples, after controlling for Static-99R scores (Helmus, 2009). Most previous field studies of Static-99R scores used arrests or charges as the recidivism criteria (Boccaccini et al., 2009; Hanson et al., 2014, 2015).

Because the primary goal of the Static-99 is to estimate risk of sexual recidivism, we do not report results for other types of recidivism (e.g., general criminal recidivism, or nonsexual violent recidivism). We considered both contact sexual offenses (e.g., sexual assault, sexual abuse, indecency with a child) and noncontact sexual offenses (e.g., exposure, online solicitation, possession of child pornography) as indicators of sexual recidivism. The base-rate of sexual recidivism was 3.6% (n = 1,241).

In some analyses, we used a fixed 5-year definition of sexual recidivism. These analyses included the 17,455 offenders (50.3%) who had been released for at least 5 years when we collected recidivism data (see Figure 1), and only counted new arrests that happened within 5 years of release as recidivism. The base rate of sexual recidivism was 4.0% (n = 691) in the fixed 5-year sample.

Overview of Analyses

We examined relative predictive accuracy (also referred to as discrimination) of the Static-99 and Static-99R using Cohen's d and the Area Under the Curve (AUC) from Receiving Operator Characteristic curve analyses. Following Cohen (1992), we considered d values of .20, .50, and .80 roughly as small, moderate, and large effects, respectively. Following Rice and Harris (2005), we considered AUCs of .56, .64, and .71 as small, moderate, and large effects as they roughly correspond to d values of .20, .50, and .80. We used the z-score formula from Hanley and McNeil (1982) to compare AUC values between different subgroups of offenders. We compared findings from our study to those from other studies using cumulative meta-analysis (Hanson & Broom, 2005), which tests if the current findings are significantly different from previous findings (either a single study or a previous meta-analytic average) and provides a new mean effect size based on the combination of effects from the two sets of findings being compared. When the effect size being compared was an AUC, we also calculated and tested the difference between the two AUCs using the formula from Hanley and McNeil (1982). Given that this method produced no differences in results from the cumulative meta-analysis, we report only the cumulative meta-analysis findings.

Relative predictive accuracy (i.e., discrimination) has been commonly analyzed in risk scale validation studies, but absolute predictive accuracy (also referred to as calibration) has been largely neglected in the offender risk assessment field (Helmus & Babchishin, in press). Whereas discrimination analyses indicate whether the scale effectively rank orders offenders in terms of their risk to reoffend, calibration analyses examine whether the recidivism estimates associated with scores on an actuarial risk scale generalize to new validation samples or subgroups.

We used calibration analyses to determine whether the routine sample recidivism norms for the Static-99R were generalizable to Texas. Calibration analyses were restricted to the Static-99R be-

cause the scale developers have recommended that the revised scale replace the original (Helmus, Thornton et al., 2012) and they have not published updated norms for the original Static-99. For the calibration analyses, we used the E/O index (Gail & Pfeiffer, 2005; Rockhill, Byrne, Rosner, Louie, & Colditz, 2003) to examine the correspondence between expected and observed recidivism rates based on Static-99R scores. For a more detailed explanation and worked-out examples of these calculations, see Hanson (in press). The E/O index is the ratio of the expected number of recidivists (E) divided by the observed number of recidivists (O; Method M_0 from Viallon, Ragusa, Clavel-Chapelon, & Bénichou, 2009). Following Rockhill et al. (2003), we calculated the 95% confidence intervals for the E/O index using the Poisson variance for the logarithm of the observed number of cases (O):

$$\textit{Lower Limit of } 95\%\ CI \textit{ of } \frac{E}{O}\ \textit{Index} = \left(\frac{E}{O}\right)e^{\left(1.96\sqrt{\frac{1}{O}}\right)}$$

$$\textit{Upper Limit of } 95\%\ CI \textit{ of } \frac{E}{O}\ \textit{Index} = \left(\frac{E}{O}\right)e^{\left(-1.96\sqrt{\frac{1}{O}}\right)}$$

An E/O index of 1 indicates that the expected number of recidivists perfectly matches the observed. An index below 1 means the scale underpredicted recidivism and an index above 1 means the scale overpredicted recidivism. If the 95% confidence interval includes 1, then the expected recidivism rate is not significantly different than the observed rate. For these analyses, we obtained expected recidivism rates from the Static-99R norms for routine correctional samples after 5 years of follow-up (Hanson et al., 2016; Phenix et al., 2015) and compared these expected rates to those we observed in the Texas fixed 5-year sample.

We developed local Texas norms for Static-99R scores using the same procedures as those used to develop the current Static-99R norms. Specifically, we used logistic regression with the fixed 5-year sample to compute estimated recidivism rates (in logits) for each Static-99R score. We calculated 95% confidence intervals for the predicted recidivism rates in logits, using the following formula to compute the standard error (from Fleiss, Levin, & Paik, 2003):

$$SE_{logit} = \sqrt{(SE_{B0})^2 + 2 \cdot x \cdot r \cdot SE_{B0} \cdot SE_{B1} + x^2(SE_{B1})^2}$$

As a final step, we transformed recidivism rates and the confidence intervals from logits into percentages.

Results

Because of the large number of offenders in this study, even small predictive validity effects are large enough to reach statistical significance ($p < .001$) in many of our analyses. Although we report p values, our emphasis is on comparing predictive effects among subgroups within our sample and comparing predictive effects from our sample to those from the Static-99 and Static-99R normative samples and prior field studies.

Static-99 and Static-99R Predictive Effects

Table 1 provides AUC and Cohen's d values for Static-99 and Static-99R scores predicting sexual recidivism among the overall sample of released offenders ($N = 34{,}687$) and various subgroups of offenders. In the overall and fixed 5-year samples, Static-99 and Static-99R scores were medium-sized predictors of sexual recidivism (AUC = .623 to .650), larger in terms of absolute value than in the 2009 Texas field study (AUC = .60). Because predictive effects were similar in the overall sample and in the fixed 5-year sample, we used data from the entire sample for subgroup comparisons.

In terms of absolute value, predictive effects for both Static-99 and Static-99R scores tended to be strongest among offenders scored after 2003 (AUC = .660 and .667) and among discharged offenders (AUC = .654 and .657). The z-value for the difference in predictive effects between those scored before and after Jan 1, 2004 was large enough to reach statistical significance ($z > 1.96$)

Table 1
Predictive Validity of Static-99 and Static-99R Total Scores for Sexual Recidivism

			Static-99				Static-99R			
Sample/Subsample	*N*	Base rate	AUC	*SE*	95% CI	*d*	AUC	*SE*	95% CI	*d*
Follow-up period										
Any follow-up	34,687	3.6%	.638	.0081	[.62, .66]	.55	.650	.0078	[.63, .66]	.56
Fixed 5-year	17,455	4.0%	.625	.0108	[.60, .65]	.51	.640	.0107	[.62, .66]	.53
Scoring timeframe										
Scored pre-2004	15,680	5.4%	.614	.0010	[.59, .63]	.44	.633	.0096	[.61, .65]	.49
Scored post-2003	19,007	2.1%	.660	.0144	[.63, .69]	.69	.667	.0137	[.64, .69]	.65
Release type										
Discharged	18,255	3.9%	.654	.0103	[.63, .67]	.61	.657	.0101	[.64, .68]	.58
Supv./Parole	16,322	3.1%	.618	.0130	[.59, .64]	.48	.634	.0124	[.61, .66]	.52
Race/Ethnicity										
African American	9,725	4.5%	.626	.0140	[.60, .65]	.47	.640	.0135	[.61, .67]	.51
Caucasian	7,938	4.9%	.646	.0145	[.62, .68]	.61	.651	.0141	[.62, .68]	.58
Latino	8,939	3.0%	.643	.0171	[.61, .68]	.61	.634	.0170	[.60, .67]	.52
Ethnicity missing	7,903	1.8%	.583	.0231	[.54, .63]	.30	.621	.0231	[.58, .67]	.43
Latino birth country										
Born in US	6,337	3.9%	.616	.0180	[.58, .65]	.50	.613	.0180	[.58, .65]	.44
Born outside US	2,459	.7%	.667	.0644	[.54, .79]	.72	.646	.0658	[.52, .78]	.58

Note. $p \leq .01$ for all effects, with the exception of those for Latino offenders born outside the U.S. ($ps < .05$).

for Static-99 scores ($z = 2.55$, $p = .005$) and approached significance for Static-99R scores ($z = 1.89$, $p = .06$). The difference in effects between offenders who were discharged and those released under mandatory supervision or parole was statistically significant for Static-99 scores ($z = 2.05$, $p = .04$), but not for Static-99R scores ($z = 1.31$, $p = .19$).

There was, however, no evidence that predictive effects were significantly weaker among Latino offenders than Caucasian or African American offenders ($zs < 1.05$, $ps > .30$). Although predictive effects were larger, in terms of absolute value, for Latino offenders born outside the U.S. than inside the U.S., they were not significantly larger ($zs < 0.70$, $ps > .30$). There were only 17 recidivists born outside the U.S. (0.7% base rate), which led to large 95% confidence intervals around the predictive effects for Latino offenders born outside the U.S.

With respect to race/ethnicity, effects were smallest among White offenders who were missing information about their ethnicity, although the predictive effect was smaller for the Static-99 (AUC = .583) than the Static-99R (AUC = .621). These offenders were significantly ($p < .001$) older at release ($M = 44.10$ years, $SD = 11.96$) than Caucasian ($M = 38.83$ years, $SD = 11.75$) and Latino ($M = 37.63$ years, $SD = 11.17$) offenders ($d = .44$ & $.56$, respectively), which may help explain why Static-99R scores—which better account for the association between age and recidivism—were better predictors among this subgroup than Static-99 scores.

Comparisons to Previous Meta-Analyses and Field Studies

Table 2 presents cumulative meta-analyses comparing the effect sizes for the total sample (see Table 1) with previous research. The predictive accuracy of Static-99 scores (AUC = .638, $d = .55$) among Texas offenders was significantly lower than the predictive accuracy reported in the Hanson and Morton-Bourgon (2009) meta-analysis of 63 studies ($d = .67$), and also significantly lower than the previous field studies from California (AUC = .824; Hanson et al., 2014), Canada (AUC = .740; Hanson et al., 2007), and Austria (AUC = .730; Rettenberger et al., 2013). The current findings were not significantly different than the previous field study from Texas ($d = .360$; Boccaccini et al., 2009), although the offenders in the previous study were subsumed in the current dataset, so this was not a test of fully independent samples. The accuracy for Static-99R (AUC = .650) scores was significantly lower than a 2012 meta-analysis (AUC = .693; Helmus, Hanson, et al., 2012), as well as in field studies from California (AUC = .817) and Canada (AUC = .734; Hanson et al., 2015). Although lower than the results from Austria (AUC = .71; Rettenberger et al., 2013), this difference only approached significance ($p = .056$). The current Static-99R effects were the same as the previous field study from Texas (AUC = .650; Boccaccini & Murrie, 2014).

Given the possibility that improved reliability of Static-99 scoring beginning in 2004 meaningfully increased accuracy, we made the same comparisons using only offenders scored on or after January 1, 2004 (see Table 3). For Static-99 scores, the predictive accuracy in these post-2003 cases ($d = .69$, AUC = .660) was significantly higher than the previous field study from Texas and no longer significantly different from the Hanson and Morton-Bourgon (2009) meta-analysis. The findings were, however, still significantly lower than the results from California, Canada, and Austria. Predictive accuracy for Static-99R scores (AUC = .667) was still significantly lower than the results for California and Canada, but was not significantly different than the results from Austria, the previous Texas field study, or from the previous meta-analysis.

Observed Recidivism Rates and Calibration Analyses

Table 4 provides fixed 5-year recidivism rates for Static-99 and Static-99R scores. For most Static-99 and Static-99R scores, these

Table 2
Comparing Current Findings (Total Sample) to Previous Research Using Cumulative Meta-Analysis

Comparison	Effect size metric	Current field study results			Previous results				Q_{change}	p	New meta average	
		N	Effect size	SE	k	N	Effect size	SE			Effect size	SE
Static-99												
Previous meta-analysis	Cohen's d	34,687	.550	.0290	63	24,089	.670	.0255	9.66	.002	.618	.0191
California field study	AUC	34,687	.638	.0081	1	475	.824	.0508	13.09	<.001	.643	.0080
Texas field study (2009)	Cohen's d	34,687	.550	.0290	1	1,928	.360	.1292	2.06	.151	.541	.0283
Canada field study	AUC	34,687	.638	.0081	1	972	.740	.0332	8.93	.003	.644	.0079
Austria field study	AUC	34,687	.638	.0081	1	1,077	.730	.0306	8.44	.004	.644	.0078
Static-99R												
Previous meta-analysis	AUC	34,687	.650	.0078	22	8,055	.693	.0092	12.97	<.001	.668	.0059
Canada field study	AUC	34,687	.650	.0078	1	764	.734	.0309	7.03	.008	.655	.0076
California field study	AUC	34,687	.650	.0078	1	475	.817	.0518	10.22	.001	.653	.0077
Texas field study (2009)	AUC	34,687	.650	.0078	1	847	.651	.0545	<.01	>.999	.650	.0077
Austria field study	AUC	34,687	.650	.0078	1	1,077	.710	.0306	3.66	.056	.653	.0076

Note. For Static-99, the previous meta-analysis was Hanson & Morton-Bourgon (2009). For Static-99R, the previous meta-analysis was Helmus, Hanson, Thornton, Babchishin, & Harris (2012). The California field study was from Hanson, Lunetta, Phenix, Neeley, & Epperson (2014). TheTexas field study was from Boccaccini, Murrie, Caperton, & Hawes (2009) for the Static-99, using data from all released offenders, and Boccaccini & Murrie (2014) for the Static-99R, using only offenders released for at least 5 years. The Canada field study for Static-99 comparisons was from an early report (Hanson, Harris, Scott, & Helmus, 2007) and the Static-99R data were from the final report (Hanson, Helmus, & Harris, 2015). The Austrian field study was Rettenberger, Haubner-Maclean, & Eher (2013).

Table 3
Comparing Current Findings (Post 2003 Data) to Previous Research Using Cumulative Meta-Analysis

Comparison	Effect size metric	Current field study results			Previous results						New meta average	
		N	Effect size	*SE*	*k*	*N*	Effect size	*SE*	Q_{change}	*p*	Effect size	*SE*
Static-99												
Previous meta-analysis	Cohen's *d*	19,007	.690	.0505	63	24,089	.670	.0255	.12	.729	.674	.0228
California field study	AUC	19,007	.660	.0144	1	475	.824	.0508	9.66	.002	.672	.0139
Texas field study (2009)	Cohen's *d*	19,007	.690	.0505	1	1,928	.360	.1292	5.66	.017	.646	.0471
Canada field study	AUC	19,007	.660	.0144	1	972	.740	.0332	4.90	.027	.673	.0132
Austria field study	AUC	19,007	.660	.0144	1	1,077	.730	.0306	4.28	.039	.673	.0130
Static-99R												
Previous meta-analysis	AUC	19,007	.667	.0137	22	8,055	.693	.0092	2.48	.115	.685	.0076
Canada field study	AUC	19,007	.667	.0137	1	764	.734	.0309	3.94	.047	.678	.0125
California field study	AUC	19,007	.667	.0137	1	475	.817	.0518	7.84	.005	.677	.0132
Texas field study (2009)	AUC	19,007	.667	.0137	1	847	.651	.0545	.08	.777	.666	.0133
Austria field study	AUC	19,007	.667	.0137	1	1,077	.710	.0306	1.64	.200	.655	.0125

Note. For Static-99, the previous meta-analysis was Hanson & Morton-Bourgon (2009). For Static-99R, the previous meta-analysis was Helmus, Hanson, Thornton, Babchishin, & Harris (2012). The California field study was from Hanson, Lunetta, Phenix, Neeley, & Epperson (2014). The Texas field study was from Boccaccini, Murrie, Caperton, & Hawes (2009) for the Static-99 and Boccaccini & Murrie (2014) for the Static-99R. The Canada field study for Static-99 comparisons was from an early report (Hanson, Harris, Scott, & Helmus, 2007) and the Static-99R data were from the final report (Hanson, Helmus, & Harris, 2015). The Austrian field study was Rettenberger, Haubner-Maclean, & Eher (2013).

rates are lower than those from the instrument norms. For example, we found recidivism rates of 5.9%, 7.5%, and 15.2% for Static-99R scores of 5, 6, and 7, but the corresponding predicted recidivism rates from the Static-99R Routine Samples norms are 15.2%, 20.5%, and 27.2% (Hanson et al., 2016; Phenix et al., 2015).

We used calibration analyses to compare the fixed 5-year recidivism rates from Texas to those from the Static-99R norms. We focused on Static-99R scores because this is the measure now used in contemporary practice, including Texas. Table 5 summarizes how the observed recidivism rates corresponded to the predicted recidivism rates from the Static-99R normative data, examining all cases, as well as offenders grouped within each total score and each risk category. We present analyses separately for offenders scored before January 1, 2004 and on or after January 1, 2004.

There were 14,077 offenders scored in 2003 or earlier in the fixed 5-year sample, 568 of whom sexually reoffended (4.0%). However, based on the Static-99R scores of these offenders, the recidivism norms would have predicted roughly 1,081 recidivists from this sample. In other words, the scale predicted roughly twice as many recidivists as there were observed (E/O Index = 1.90, 95% CI [1.75, 2.07]), which was a significant overestimation. Examining the results per score, the Static-99R underestimated recidivism for scores of −3 and −2 (E/O Indexes less than 1), but not significantly so, although the small number of recidivists reduced statistical power. For all other scores, the Static-99R overestimated recidivism, and this was statistically significant for scores of 1 through 7. When collapsing the offenders into risk categories on the basis of their Static-99R scores (see Table 5), Static-99R scores significantly overestimated recidivism for risk Categories III, IVa, and IVb (i.e., the three highest risk categories), with E/O indexes ranging between 1.7 and 2.5. Overestimation of recidivism was greatest for the highest risk category, where the scale predicted about 2.5 times the number of recidivists than were actually observed.

For offenders scored on or after January 1, 2004, calibration results were fairly similar, although with lower statistical power because of fewer recidivists (n = 123). Collapsing across all offenders, the recidivism norms predicted approximately twice as many recidivists as the observed rate (E/O Index = 2.05, 95% CI [1.71, 2.44]), which was a significant overestimation. The Static-99R norms also significantly overestimated recidivism for risk Categories III and IVa, with E/O Indexes of approximately 2.3. For the other categories, the observed recidivism rates were not significantly different than the predicted rates.

Local Recidivism Rate Norms for Texas Static-99R Scores

Given that the Static-99R norms tended to overpredict recidivism twofold in the current sample, we produced local (Texas) recidivism norms using the same logistic regression procedures used to develop the published recidivism norms for the Static-99R (Hanson et al., 2016; Phenix et al., 2015). We used only the 3,378 offenders scored on or after January 1, 2004 with at least 5-years of follow-up time for these norms, and the norms are based on a fixed 5-year definition of recidivism (n = 123 recidivists, 3.6% base rate). The mean Static-99R score among these 3,378 offenders was 2.23 (SD = 2.01). We used offenders scored in 2004 or later because the scores assigned to these offenders are most likely to represent contemporary practice. Although restricting the estimates to the more recent data substantially decreased the sample size, the remaining sample still exceeded Hanson et al.'s (2016) recommendation of at least 100 recidivists for credible recidivism norms.

The new Texas local norms are provided in the Appendix. These norms include predicted recidivism rates and 95% confidence intervals. Figure 2 compares the official Static-99R recidivism norms for routine correctional samples (Hanson et al., 2016; Phenix et al., 2015) to the new Texas norms. For most scores, the Texas estimates are roughly half of the Static-99R norms, although the magnitude of this difference remains quite small until Static-99R scores of 4 and higher, where it becomes more pronounced.

Table 4
Fixed 5-Year Sample Sexual Recidivism Rates for Each Static-99 and Static-99R Total Score

Score	Offenders	Recidivists	Recidivism rate
		Static-99	
0	1,560	22	1.4%
1	3,155	79	2.5%
2	4,274	131	3.1%
3	3,886	180	4.6%
4	2,507	116	4.6%
5	1,242	79	6.4%
6	498	28	5.6%
7	232	34	14.7%
8	72	17	23.6%
9	23	5	17.9%
10	5	0	.0%
11	1	0	.0%
12	—	—	—
Any	17,455	691	4.0%
		Static-99R	
−3	169	4	2.4%
−2	192	4	2.1%
−1	1,116	12	1.1%
0	1,858	42	2.3%
1	2,528	56	2.2%
2	3,142	107	3.4%
3	3,703	153	4.1%
4	2,575	142	5.5%
5	1,347	80	5.9%
6	535	40	7.5%
7	184	28	15.2%
8	77	17	22.1%
9	23	5	21.7%
10	5	1	20.0%
11	1	0	.0%
12	—	—	—
Any	17,455	691	4.0%

Discussion

In this largest ever field validity study of Static-99 and Static-99R scores, we found medium-sized predictive effects for sexual recidivism. When we focused on scores assigned after the release of a new Static-99 scoring manual (Harris et al., 2003), the predictive effects were stronger and fell within the range expected based on prior Static-99 and Static-99R meta-analyses (Hanson & Morton-Bourgon, 2009; Helmus, Hanson et al., 2012). Although researchers have reported larger predictive effects for Static-99 and Static-99R scores in other field studies (see Tables 2 and 3), the observation of effects within the expected range in Texas is an important finding, given the prior finding of only small predictive effects among Texas offenders released after being screened for SVP civil commitment (Boccaccini et al., 2009). Texas scores the Static-99R for thousands of sexual offenders each year, and our findings suggest acceptable performance for post-2003 scores, at least with respect to recidivists scoring higher than nonrecidivists.

But our observed recidivism rates suggested that the normative sample recidivism rates did not generalize to Texas, and our calibration analyses revealed that the norms led to a significant overestimation in risk, especially for offenders with scores ranging from 1 to 5. These findings suggested the need for local Texas recidivism rate norms. We developed Texas norms for Static-99R scores using the same type of logistic regression procedures that the Static-99R authors used to develop the current Static-99R recidivism rate norms (see Appendix). These norms replace those based on the original Texas field validity study (Boccaccini et al., 2009; Boccaccini & Murrie, 2014).

Although our findings have specific implications for evaluators in Texas, they also highlight the potential benefits of conducting large-scale field studies in any jurisdiction. Instrument manuals provide information about how well instruments work in normative samples, but those normative sample findings may or may not generalize to any particular field setting. It is only through field studies conducted within a jurisdiction that we can know if the normative sample findings apply to that jurisdiction, and effects may vary from jurisdiction to jurisdiction. Indeed, the same type of calibration analyses that showed an overestimation of recidivism in our Texas sample showed better estimation in a California field study (Hanson et al., 2014).

Our findings also show the potential benefits of revisiting and updating field study findings. Any field study finding provides a picture of psychometric properties at one point in time, or over one specific period of time. The findings from that time period may or may not generalize to other time periods. Indeed, our original Static-99 field reliability (Boccaccini et al., 2012) and field validity (Boccaccini et al., 2009) findings suggested weaker than expected performance, but our updated sample reexaminations of both field reliability (Rice et al., 2014) and field validity (this study) have suggested improved performance over time. Static-99 scores assigned since 2004 are more reliable than earlier scores, and they are better predictors of recidivism than earlier scores.

Although we separated offenders into earlier and later scoring groups based on the release of an updated Static-99 scoring manual (Harris et al., 2003), we cannot know with any degree of certainty whether the improvements in reliability and validity are attributable to the revised scoring manual. One important limitation of field research is that researchers often have no control over administration and scoring procedures, and it can be difficult to look backward to identify factors that may explain variability in field study effects. The findings of improved reliability and validity over time could be attributable to the new scoring manual, or they could be due to changes in administrative policy, personnel factors, or other factors that are outside of institutional control, such as increasing sentence lengths for sexual offenders or changes in eligibility for release and parole.

Using the Texas Norms

By definition, the Texas recidivism rate norms we have provided apply only to Static-99R scores in Texas, and there are several factors for Texas evaluators to consider before using these norms in practice. Although these new Texas norms are based on a relatively large sample of offenders (n = 3,378), there were only 123 recidivists in this sample. This is a relatively small number of recidivists compared with the Static-99R routine sample norms (358 recidivists). The Texas norms sample is large enough ($N >$ 100; Hanson et al., 2016) to allow for a fairly stable calculation of recidivism rates, but is close enough to the recommended minimum threshold of 100 recidivists that these rates should be interpreted with caution.

Table 5
Calibration Results for Static-99R Scores

Score/ Category	Expected Recid (proportion)	Cases scored 2003 or earlier						Cases scored 2004 onwards					
		N	n recid (O)	Expected recid (E)	E/O	95% CI		*N*	n recid (O)	Expected recid (E)	E/O	95% CI	
−3	.009	127	2	1.1	.55	.14	2.20	42	2	.4	**.20**	**.05**	**.80**
−2	.013	146	3	1.9	.63	.20	1.96	46	1	.6	.60	.08	4.26
−1	.019	886	11	16.8	1.53	.85	2.76	230	1	4.4	4.40	.62	31.24
0	.028	1,469	35	41.1	1.17	.84	1.64	389	7	10.9	1.56	.74	3.27
1	.039	2,070	49	80.7	**1.65**	**1.24**	**2.18**	458	7	17.9	**2.56**	**1.22**	**5.36**
2	.056	2,541	85	142.3	**1.67**	**1.35**	**2.07**	601	22	33.7	**1.53**	**1.01**	**2.33**
3	.079	2,947	133	232.8	**1.75**	**1.48**	**2.07**	756	20	59.7	**2.99**	**1.93**	**4.63**
4	.110	2,121	117	233.3	**1.99**	**1.66**	**2.39**	454	25	49.9	**2.00**	**1.35**	**2.95**
5	.152	1,084	66	164.8	**2.50**	**1.96**	**3.18**	263	14	40.0	**2.86**	**1.69**	**4.82**
6	.205	452	31	92.7	**2.99**	**2.10**	**4.25**	83	9	17.0	1.89	.98	3.63
7	.272	142	17	38.6	**2.27**	**1.41**	**3.65**	42	11	11.4	1.04	.57	1.87
8	.351	71	16	24.9	1.56	.95	2.54	6	1	2.1	2.10	.30	14.91
9	.438	16	2	7.0	3.50	.88	13.99	7	3	3.1	1.03	.33	3.20
10	.530	4	1	2.1	2.10	.30	14.91	1	0	.5	—	—	—
11	.530	1	0	.5	—	—	—	—	—				
Total	—	14,077	568	1,080.6	**1.90**	**1.75**	**2.07**	3,378	123	251.6	**2.05**	**1.71**	**2.44**
Cat. I	—	273	5	3.0	.60	.25	1.44	88	3	1.0	.33	.11	1.03
Cat. II	—	2,355	46	57.9	1.26	.94	1.68	619	8	15.3	1.91	.96	3.82
Cat. III	—	7,558	267	455.8	**1.71**	**1.51**	**1.92**	1,815	49	111.3	**2.27**	**1.72**	**3.01**
Cat. IVa	—	3,205	183	398.1	**2.18**	**1.88**	**2.51**	717	39	89.9	**2.31**	**1.68**	**3.15**
Cat. IVb	—	686	67	165.8	**2.47**	**1.95**	**3.14**	139	24	34.1	1.42	.95	2.12

Note. Bold values indicate that the E/O Index is statistically significant. Expected recidivism rates were obtained from 2015 Evaluator Workbook (Phenix, Helmus, & Hanson, 2015) from www.static99.org. Risk Categories were obtained from Hanson et al., 2016.

In addition, the Texas norms are based on Static-99R scores calculated from Static-99 scores, using birth and release dates. Because we do not know whether the anticipated release dates used to score the Static-99 matched up with the actual release dates we used to calculate Static-99 age-at-release item scores, there may have been some errors in our conversion of Static-99 scores to Static-99R scores (although such errors would be small in number and likely have minimal impact on the overall results). Ideally, Static-99R norms should be based on Static-99R scores, not those converted from the Static-99. But the Static-99R was first released in 2009 (Helmus, 2009; Hanson, Phenix, & Helmus, 2009), so it would not be possible to calculate 5-year recidivism rates for Static-99R scores in this study unless we based them on Static-99 scores. Texas evaluators now use the Static-99R, not the Static-99, and they need norms that apply to scores from the measure they are using (see Boccaccini & Murrie, 2014).

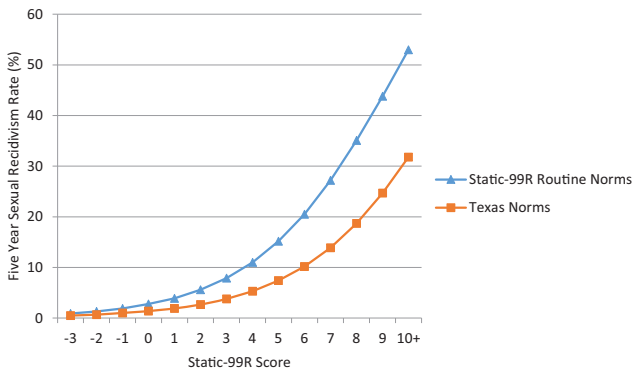

Figure 2. Comparing estimated recidivism rates from post-2003 cases in Texas to Static-99R routine correctional norms (5-year recidivism rates). See the online article for the color version of this figure.

Ultimately, it is up to each Texas evaluator to best decide how to use the Texas norms. Evaluators may decide to present only the Texas rates, or both the Texas rates and the routine sample rates from the Evaluator Workbook (Phenix et al., 2015). Both approaches offer advantages and disadvantages. With respect to representativeness, Texas offenders will better match the Texas norms than the routine samples norms, and the Texas norms are arguably more applicable to Texas offenders than those from other countries and U.S. states. But the Texas norms are based on a smaller number of recidivists, in a jurisdiction with evidence of changing performance over time. The routine samples norms include a larger number of recidivists and, thus, more stable recidivism estimates, but those norms are based on samples with a higher recidivism rate than we found in Texas. The routine samples norms are also based on Static-99R data with higher predictive accuracy, possibly suggesting better quality data.

Reporting both sets of norms (routine sample norms and Texas norms) acknowledges the uncertainty inherent in estimating recidivism rates and offers a plausible range of empirically defensible estimates. However, evaluators who report two sets of recidivism rates will inevitably be faced with the question of which set of estimates is the most applicable. We encourage evaluators to always keep in mind that these recidivism estimates are only estimates, and all of them should be interpreted with appropriate caution. No matter what estimates are used, evaluators should acknowledge the strengths and limitations of the information they are using, and the source of that information. Moreover, a recent survey of 109 SVP evaluators who used the Static-99R found that

only one (0.9%) based his or her opinion about recidivism risk entirely on the offender's Static-99R score (Chevalier et al., 2015). In other words, it is very rare for evaluators to base their risk opinions on instrument results alone, let alone one specific recidivism rate estimate. For evaluators in Texas, the Texas norms provide one additional piece of information that evaluators can consider when coming to conclusions about risk.

Limitations

Although the large number of Texas sexual offenders with Static-99 scores allowed for the largest ever Static-99 validity study, the data available for this study were limited in several important ways. We had no information about Static-99 item scores, no specific information about the training and experience of individual Static-99 raters, and no information about the ethnicity of more than 7,000 White offenders. We found that predictive effects were smaller among those with missing ethnicity information, but can only speculate as to why (e.g., poor record quality for these generally older offenders). The validity analyses were also limited by the exclusion of offenders who had been civilly committed. Although SVP commitment affected fewer offenders in this study (0.7%) than in the prior Texas field study (2.2%), exclusion of these presumably high risk offenders may have led to somewhat attenuated effects. There was, however, no clear evidence of range restriction in Static-99 or Static-99R scores because of omitting the committed offenders. For example, the standard deviation of Static-99R scores was 2.02 for both the entire sample of released offenders (which includes civilly committed offenders)[3] and the subset of 34,687 released but not committed offenders in our analyses.

The sheer number of offenders made it impossible for us to look beyond official DPS arrest records for information about postrelease sexual offenses. The low fixed 5-year recidivism rate (4.0%) that we observed using these records is almost certainly an underestimate. Any definition of recidivism based on official criminal records will underestimate the true rate of recidivism because of underreporting (e.g., Dobash & Dobash, 1995). Nevertheless, the recidivism rate in this study falls within the range reported in recent U.S. sample Static-99 and Static-99R studies. For example, researchers have reported rates of 1.9% in New Jersey (Leguizamo et al., 2015), 4.8% in California (Hanson et al., 2014), and 8.2% in New York (Sandler, 2010). As with these other studies, we coded arrests as sexual offenses only if the arrest charge clearly indicated a sexual offense. If we were able to review detailed correctional and law enforcement files, we may have found that some offenders were charged with seemingly nonsexual offenses (e.g., burglary, assault, kidnapping, failure to comply with conditional release) for sexually motivated crimes.

The fact that many offenders had been incarcerated on multiple occasions during the study period or scored multiple times during the same incarceration period forced us to make decisions about which data to use in our analyses. We used data from the earliest incarceration period (after a Static-99 score) for each offender because doing so allowed us to maximize follow-up time, but it could be that this selection process affected our findings in some unintended way. For example, this process would have favored earlier scores, and we found that earlier scores were less predictive than later scores. Fortunately, we had enough offenders to conduct separate analyses for offenders scored before and after January 1, 2004, but our overall sample predictive effects may be attenuated because of us favoring earlier scores in those analyses. Although we used January 1, 2004 to identify offenders scored after the release of the new Static-99 scoring manual, this date is only an estimate. The manual was released in 2003 and we presume—but cannot know with certainty—that it had been incorporated by the beginning of 2004. Although it seems reasonable to assume that new scoring guidelines were at least partially responsible for the improved reliability (Rice et al., 2014) and validity (this study) of Static-99 scoring after 2003, we have no way to know if this is actually the case.

Conclusion

The Static-99R is used across the world to facilitate decision-making about sexual offenders. As with any measure, there will always be variability in effects across studies, and evaluators and decision-makers who use the Static-99R want to know how well the measure works in their field setting. Our original Texas field study findings suggested that Static-99 scores assigned as part of routine practice may not work as well as those assigned for research purposes, raising important questions about using the instrument in practice (Boccaccini et al., 2009). Although subsequent field studies demonstrated that Static-99 and Static-99R scores can have strong predictive effects in field settings (e.g., Hanson et al., 2014), the Texas study has served as a reminder that research findings do not always translate to the field. By revisiting the field validity of Static-99 and Static-99R scores in Texas, we were able to show that predictive validity has improved over time, but that there may still be a need to develop local recidivism norms for the Static-99R, even when overall predictive effects are in the expected range.

The finding of improved field validity over time is promising for risk assessment research. Revising and updating instrument manuals (based on feedback from field evaluators) may be one of the best mechanisms for working to ensure that findings from controlled studies generalize to the field. It is, however, important to keep in mind that no risk measure study can provide the type of information needed to make perfectly accurate predictions about individual offenders. Sex offender risk assessment measures predict reoffending at levels that are significantly better than chance and better than unstructured clinical judgment (Hanson & Morton-Bourgon, 2009), but effects are sometimes only moderate in size. Given the complexity of human behavior and imperfections in criminal history records, no risk scale will ever achieve perfect accuracy. When there is a low base rate of recidivism, it is inevitable that some or even many relatively high scoring offenders will not reoffend, even when high scoring offenders are more likely to reoffend than lower scoring offenders.

[3] The mean Static-99R score among all 34,943 released offenders was 2.19 (SD = 2.02).

References

Boccaccini, M. T., & Murrie, D. C. (2014). Keeping up with the field in field validity research: Updated Texas norms for the Static-99 and

Static-99R. In A. Schlank (Ed.), *The Sexual Predator* (Vol. 5, pp. 7–1-7–15). 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,* 278–314. http://dx.doi.org/10.1037/a0017232

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

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior, 39,* 209–218. http://dx.doi.org/10.1037/lhb0000114

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Dobash, R. P., & Dobash, R. E. (1995). Reflections on findings from the Violence Against Women survey. *Canadian Journal of Criminology, 37,* 457–484.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/0471445428

Gail, M. H., & Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics, 6,* 227–239. http://dx.doi.org/10.1093/biostatistics/kxi005

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36. http://dx.doi.org/10.1148/radiology.143.1.7063747

Hanson, R. K. (in press). Assessing the calibration of actuarial risk scales: A primer on the E/O index. *Criminal Justice and Behavior*.

Hanson, R. K., Babchishin, K. M., Helmus, L. M., Thornton, D., & Phenix, A. (in press). Communicating the results of criterion-referenced tests: Evidence-based risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. *Psychological Assessment*.

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

Hanson, R. K., Harris, A. J. R., Scott, T.-L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (User Report No. 2007–05). Ottawa, ON: 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,* 1205–1224. http://dx.doi.org/10.1177/0093854815602094

Hanson, R. K., Helmus, L., & Thornton, D. (2010). Predicting recidivism amongst sexual offenders: A multi-site study of Static-2002. *Law and Human Behavior, 34,* 198–211. http://dx.doi.org/10.1007/s10979-009-9180-1

Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. *International Journal of Forensic Mental Health, 9,* 11–23.

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,* 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,* 1–21. http://dx.doi.org/10.1037/a0014421

Hanson, R. K., Phenix, A., & Helmus, L. (2009, September). *Reporting Static-99 and Static-2002 scores in light of new research findings*. Preconference workshop presented at the Annual Research and Treatment Conference for the Association for the Treatment of Sexual Abusers, Dallas, TX.

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sexual offenders* (User Report No. 1999–02). Ottawa, ON, Canada: Solicitor General Canada.

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24,* 119–136. 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, 28,* 218–252. http://dx.doi.org/10.1177/1079063215574710

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

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MR58443)

Helmus, L. M., & Babchishin, K. (in press). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior*.

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior, 39,* 1148–1171. 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,* 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,* 64–101.

Leguízamo, A., Lee, S. C., Jeglic, E. L., & Calkins, C. (2015). Utility of the Static-99 and Static-99R with Latino sex offenders. *Sexual Abuse*. Advance online publication. http://dx.doi.org/10.1177/1079063215618377

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior, 41,* 1406–1421. http://dx.doi.org/10.1177/0093854814548449

Phenix, A., Helmus, L., & Hanson, R. K. (2015). *Static-99R and Static-2002R evaluator's workbook*. Available from www.static99.org

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

Rice, A. K., Boccaccini, M. T., & Luna-Collier, T. L. (2012, March). *Evaluator differences in assigning Static-99 scores*. Poster presented at the meeting of the American Psychology Law Society, San Juan, Puerto Rico.

Rice, A. K., Boccaccini, M. T., Harris, P. B., & Hawes, S. W. (2014). Does field reliability for Static-99 scores decrease as scores increase? *Psychological Assessment, 26,* 1085–1094. 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,* 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,* 856–861. http://dx.doi.org/10.1016/S0895-4356(03)00124-0

Sandler, J. C. (2010). *The Static-99 and additional research-based risk factors: A statistical theory to improve sex offender risk assessment*. Available from ProQuest Dissertations and Theses database. (UMI No. 3432541)

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

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

Viallon, V., Ragusa, S., Clavel-Chapelon, F., & Bénichou, J. (2009). How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine, 28,* 901–916. http://dx.doi.org/10.1002/sim.3517

Appendix

Static-99R Sexual Recidivism Norms for Texas

	Predicted Sexual Recidivism Rates from Logistic Regression		
Static-99R Score	Predicted Recidivism Rate (%)	95% CI	
−3	.5	.3	.9
−2	.7	.4	1.2
−1	1.0	.6	1.5
0	1.4	.9	2.0
1	1.9	1.4	2.6
2	2.7	2.2	3.4
3	3.8	3.2	4.6
4	5.3	4.5	6.4
5	7.4	6.0	9.2
6	10.2	7.8	13.3
7	13.9	10.0	19.1
8	18.7	12.7	26.7
9	24.7	16.0	36.1
10+	31.8	19.8	46.7

Note. These norms were based on 3,378 sex offenders in Texas who were scored on Static-99R after 2003 and who were released to the community with at least five years of follow-up data as of November, 2011 (123 sexually reoffended). The recidivism estimates were calculated using logistic regression ($B_0 = -4.2866$, $SE_{B0} = .1394$; $B_1 = .3523$, $SE_{BI} = .0474$; correlation of estimates $= -.872$). Examples of how to incorporate this information in applied risk assessment reports can be found on www.static99.org. If these estimates are used in place of the existing norms in the Static-99R Evaluator Workbook (from www.static99.org), however, it is important to note their source so they are not confused with the official Static-99R recidivism rates.

Received March 1, 2016
Revision received June 30, 2016
Accepted June 30, 2016 ■