Static-99R: Strengths, limitations, predictive accuracy meta-analysis, and legal admissibility review

L. Maaike Helmus, Sharon M. Kelley, Annabelle Frazier, Yolanda M. Fernandez, Seung C. Lee, Martin Rettenberger, & Marcus T. Boccaccini

In press: *Psychology, Public Policy and Law*

Author Note

L. Maaike Helmus, Department of Criminology, Simon Fraser University https://orcid.org/0000-0002-5032-2548; Sharon M. Kelley, Sand Ridge Secure Treatment Center, Madison, Wisconsin https://orcid.org/0000-0002-8129-237X; Annabelle Frazier, The Committee for Public Counsel Services; Yolanda Fernandez, Society for the Advancement of Actuarial Risk and Needs Assessment (SAARNA), Kingston, ON, Canada; Seung C. Lee, Society for the Advancement of Actuarial Risk and Needs Assessment (SAARNA), Kingston, ON, Canada https://orcid.org/0000-0002-6705-4535; Martin Rettenberger, Centre for Criminology (Kriminologische Zentralstelle – KrimZ), Wiesbaden, Germany http://orcid.org/0000-0002-0979-4295; Marcus T. Boccaccini, Department of Psychology, Sam Houston State University.

The meta-analysis dataset for this paper is available from osf.io/6af3t. Findings from this paper were briefly discussed at the 2021 Annual Research and Treatment Conference for the Association for the Treatment of Sexual Abusers (specifically, a summary of the meta-analysis and legal admissibility review; Thornton & Helmus, 2021).

This work was completed on the traditional and unceded territories of the Coast Salish Peoples (where the city of Vancouver, B.C. currently resides), specifically the Squamish (Sḵwx̱wú7mesh Úxwumixw), Tsleil-Waututh (səl̓ilw̓ətaʔɬ) and Musqueam (xʷməθkʷəy̓əm) Nations; the Algonquin Anishnaabeg People (where the city of Ottawa, Ontario currently resides); the Anishinaabe, Haudenosaunee and the Huron-Wendat (where the city of Kingston, Ontario currently resides); the Kiikaapoi (Kickapoo), Peoria, Sauk and Meskwaki, Ho-Chunk (Winnebago), Myaamia, and Očhéthi Šakówiŋ Nations (where the city of Madison, Wisconsin currently resides); the Tonkawa and Bidai Nations (where the city of Huntsville, Texas currently resides); and Pawtucket, Pennacook, Wabanaki, and Pentucket Nations (where the city of Lawrence, Massachusetts currently resides).

L. Maaike Helmus, Yolanda Fernandez, and Martin Rettenberger are certified trainers of the Static-99R and STABLE-2007 risk tools; Sharon Kelley is a certified trainer of Static-99R. The copyright for both measures is held by the Government of Canada and the authors do not receive royalties for their use. The views expressed are those of the authors and not necessarily those of the Wisconsin Department of Health Services – Division of Care and Treatment Services. Data sharing is not applicable to this article as no new data were created or analyzed in this study.

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

Most risk assessment scales are developed to broadly rank individuals according to their risk to reoffend, not to inform particular legal decisions. Consequently, there is almost always a gap between what the scale measures and the referral question. This paper discusses the most commonly used and researched risk assessment tool for individuals charged or convicted of sexual offenses: Static-99R. This review summarizes relevant history, and psychometric, applied, and legal admissibility information for those who use the scale or encounter its use in court. The first section outlines the development, purpose, and evolution of Static-99R. The second section summarizes interrater reliability and predictive validity. We conducted an updated meta-analysis of 56 studies, whereby Static-99R demonstrated moderate predictive accuracy (AUCs = .68 to .69). The third section discusses field issues. Lastly, we summarize the tool's admissibility across 85 legal cases, of which 4 resulted in inadmissibility and an additional 12 in partial admissibility. We provide suggestions for cross-examination in court. We outline how Static-99R meets Daubert criteria for legal admissibility. Although Static-99R has important limitations, we conclude that it can be useful and informative in legal decisions (caveats discussed) and is demonstrably better than the alternate of not using structured risk scales.

Static-99R: Strengths, limitations, predictive accuracy meta-analysis, and legal admissibility review

Introduction and Risk Assessment

Risk assessment is a core component of correctional decision-making. Following the risk/need/responsivity model of effective correctional practices (Bonta & Andrews, 2017), resources to reduce and manage risk should be proportional to each person's risk to reoffend (risk principle), and treatment/interventions should target empirically established dynamic predictors of recidivism (need principle). This means virtually every decision in the criminal legal system should consider a person's risk to reoffend, from arrest to sentencing, custody classification, treatment allocation, parole readiness, and community supervision intensity and conditions. Where resources are limited, allocation should be primarily based on risk. For example, if a probation officer only has enough time in a week to see one quarter of the people on their caseload, they should see the quarter whose risk is highest.

For the most severe sanctions, consideration of risk is paramount and may even be central in legislative criteria. One example is Sexually Violent Predator (SVP) commitment in the United States (for a basic overview, see Association for the Treatment of Sexual Abusers and Sex Offender Civil Commitment Programs Network, 2015). SVP commitment is the (typically) indeterminate involuntary confinement of individuals who have committed sexual offenses, after they have completed their prison sentence. In Canada, the most severe sentence possible is a Dangerous Offender designation (Canada Criminal Code S. 753(1)), applicable not just for those who committed sexual offenses but also violent offenses. The precise legal criteria vary between countries and states but place a high emphasis on risk of reoffense, often defined using terms such as "likely" or "more likely than not", although the definition of such terms can be open to interpretation (e.g., Knighton et al., 2014).

Overall, risk assessment tools are used in diverse settings for multiple decisions that vary in purpose and criteria. Most tools are developed for broad purposes of ranking individuals according to their risk to reoffend. With rare exceptions, they are not developed

to inform a particular legal decision. Additionally, they are not developed to make a dichotomous decision around dangerousness. Risk is fundamentally continuous (Hanson et al., 2013) and the Practice Guidelines for assessment of adult male sexual abusers from the Association for the Treatment of Sexual Abusers (2014) eschew definitive statements about whether any particular individual will reoffend.

This raises a conundrum for evaluators and decision-makers who often need to make dichotomous decisions. For example, does this person require treatment? Should this person be paroled? Should they be civilly committed? Generally, no risk tool is designed to answer these exact questions. This means there is almost always a gap between what the scale measures and the referral question. Specifically, risk scales measure an individual's risk, on a continuous scale, and this information can then be used to inform a wide variety of decisions about punishment, rehabilitation, and risk management. But the risk scale should not be conflated with the legal decision itself.

The purpose of this paper is to review research and practice for the most commonly used risk assessment tool for individuals charged or convicted of sexual offenses: Static-99R. The intended audience of this paper is anyone who uses Static-99R or deals with it regularly in court cases (e.g., psychologists, other risk evaluators, lawyers, judges) or researchers interested in a summary of the evidence to date on Static-99R. The paper is organized around questions that we are commonly asked or that we think people *should* be asking. This paper focuses on Static-99R as a risk assessment tool and not on other scales or on particular legal decisions (e.g., civil commitment in the United States).

The first section outlines the development and purpose of Static-99R, including its evolution given that this has been a source of confusion. The second section summarizes the psychometric properties of the scale, specifically interrater reliability and predictive validity. The third section focuses on field issues related to the scale, including misuses and international applications. The last section focuses on the legal admissibility of the scale (primarily in the United States), including suggestions for examination and cross-

examination in court. We close with recommendations for future research and conclusions on the utility and admissibility of Static-99R in legal settings. Given that the revised version (Static-99R) is recommended over its predecessor (Static-99; Helmus, Thornton, et al., 2012), we focus where possible on Static-99R. However, some mention of Static-99 is included where still relevant (e.g., Rettenberger et al., 2013).

Preliminary Definitions

Before delving into Static-99R, it is important to clarify some key terms in forensic risk assessment. Risk assessment scales synthesize risk factors, which are variables that are empirically related to recidivism. One common way of classifying risk factors is static versus dynamic/criminogenic (Bonta & Andrews, 2017). Static risk factors are historical or demographic factors that do not change (e.g., prior offenses, victim characteristics) or which only change predictably with the passage of time (e.g., age). In contrast, dynamic/criminogenic factors can change, and when changed, alter the likelihood of recidivism (Bonta & Andrews, 2017; Kraemer et al., 1997). Examples include procriminal attitudes, deviant sexual interests, and intimacy deficits (e.g., Mann et al., 2010). Dynamic risk factors can further be subdivided into those that change more slowly, like personality characteristics (stable factors) and those that change more rapidly and are more sensitive to the individual's environment (acute factors; Hanson et al., 2015).

The static/dynamic distinction sounds simple but is complicated because both may measure the same underlying propensities for criminal behaviour (Mann et al., 2010). For example, history of substance abuse (a static factor) and current substance abuse (a dynamic factor) are different ways of measuring the same construct. Scales assessing static or dynamic factors each have their own advantages. Static risk factors can be scored quickly, reliably, and with minimal information (e.g., criminal history). Dynamic risk factors are harder to assess (e.g., it is easier to discern whether someone has a conviction for a violent offense than whether they have a hostile personality pattern) but are useful in guiding interventions to reduce risk and assessing changes in risk over time.

In addition to types of risk factors, we can also differentiate methods of structuring risk factors into an overall risk assessment. A useful differentiation comes from Hanson and Morton-Bourgon (2009). Empirical-actuarial scales (such as Static-99R) are those where items are specified in advance based on research, mechanically combined into a total score based on research, and total scores are empirically linked to probability estimates of recidivism. Structured Professional Judgment (SPJ) scales are those in which a non-exhaustive list of items is specified (identified based on combinations of research, theory, and/or practice) and synthesizing this information into an overall risk evaluation is left to the discretion of the evaluator. These are the two most common types of structured risk scales. Other methods of risk assessment include mechanical scales (which look like actuarial scales but are missing some of the research-based criteria for the actuarial definition), adjusted-actuarial (allowing the evaluator to override the results of actuarial scales based on professional judgment), and unstructured clinical judgment (Hanson & Morton-Bourgon's 2009 definition would include what others [e.g., Lyon & Welsh, 2017] have called the anamnestic approach, which relies on individual case analysis).

In Hanson and Morton-Bourgon's (2009) meta-analysis of sexual recidivism risk assessment, empirical-actuarial and mechanical scales performed similarly, and both had significantly higher accuracy than unstructured judgment. The accuracy of SPJ was intermediate between actuarial/mechanical and unstructured clinical judgment, not significantly different than any of them, although there were insufficient studies to draw strong conclusions because most validation studies of SPJ scales actually use them as mechanical scales. Adjusted-actuarial scales performed worse than their unadjusted counterparts.

Section 1: What is Static-99R?

Static-99R is an empirically derived actuarial risk assessment tool. As per its coding manual (Phenix, Fernandez, et al., 2016), "Static-99R is intended to position offenders in terms of their relative degree of risk for sexual recidivism based on commonly available

demographic and criminal history information that has been found to correlate with sexual recidivism in adult male sex offenders" (p. 6). Additional information can be found at saarna.org. The 10 items assess demographics (age, ever lived with a lover for 2 years), criminal history (prior non-sexual violence, current non-sexual violence, prior sexual offenses, 4+ prior sentencing dates, any non-contact sexual offense conviction), and victim characteristics (any male, unrelated, or stranger victim) based on commonly available information usually found in legal, forensic, and clinical documentation.

Who Should it Be Used On?

The appropriate population for Static-99R may not fully correspond with the population of individuals labeled as 'sex offenders' in the criminal legal system. Specifically, Static-99R is intended for individuals who have been charged or convicted of a sexually motivated offense that meets the definition of a "Category A" sexual offense (Phenix, Fernandez, et al., 2016). This includes common sexual offenses (contact and non-contact offenses), as well as non-sexual charges or convictions that are considered by the evaluator to be sexually motivated, or where someone is supervised for a non-sexual offense but has a sexual offense (charge or conviction) in their history. The scale cannot be used for individuals whose sexual offense is based on consensual sexual activity with a similar-aged peer, where the only sexual offense(s) are child pornography offenses that do not involve creation of pornography, offenses without a sexual motive (e.g., drunken public urination), individuals less than 17 years old when they committed their latest sexual offense, and women. It also cannot be used to determine guilt or innocence.

Who Can Score the Scale?

The scale is scored based on file information and can typically be scored without an interview (although where possible, interviews can provide helpful clarification). Anyone with sufficient knowledge of the criminal legal system, risk assessment, and sexual offending, should be able to accurately score it after training with a certified trainer (roughly 8-12 hours; SAARNA, 2021). This includes members of law enforcement, correctional

workers, and mental health professionals; advanced degrees are not necessary to score the scale (Phenix, Fernandez, et al., 2016).

What do the Total Scores Mean?

Total scores on Static-99R can range from -3 to 12 and interpretive data are available in the Evaluators Workbook (Helmus, Lee, et al., 2021). Although any method of clumping total scores into risk categories will necessarily result in loss of information, category labels are typically preferred by diverse mental health practitioners involved in violence risk assessment (Heilbrun et al., 2016) despite their disadvantages (Scurich, 2018). In an effort to improve the empirical defensibility and consistency of risk levels, new risk levels were assigned to Static-99R total scores (Hanson, Babchishin, et al., 2017) following the United States Council of State Governments Justice Center standardized 5-level risk classification (Hanson, Bourgon, et al., 2017).

Risk Level I, *Very Low Risk,* describes individuals who have committed a sexual offense, but whose current risk is not meaningfully different from that of individuals in the criminal legal system with no history of sexual offending. Risk Level II, *Below Average Risk,* describes individuals whose risk for sexual recidivism is higher than Level I but lower than average for individuals charged or convicted of sexually motivated offenses. Risk Level III, *Average Risk,* describes individuals in the middle of the risk distribution, who should receive typical treatment and risk management resources. This is the largest group and encompasses roughly half of individuals scored on Static-99R (Hanson, Babchishin, et al., 2017). Level IVa, *Above Average Risk*, is a group whose risk is noticeably higher than that of most individuals with a sexual offense charge or conviction. Lastly, Level IVb, *Well Above Average Risk,* describes the top 5% to 10% of the risk distribution. Individuals in Risk Level IVb are expected to have a range of criminogenic life problems, few strengths, and sexual recidivism rates three to four times higher than for individuals in the middle of the risk distribution.

This last Level is called IVb instead of V because the Justice Center risk levels designate V as referring to "virtually certain to reoffend", defined as probabilities greater than 85% over two years (Hanson, Bourgon, et al., 2017, p.13). Fortunately, detected rates of sexual recidivism, including lifetime projections, do not meet this threshold (Thornton et al., 2021). Notably, the Standardized Risk Levels were developed by an international, interdisciplinary advisory group (Hanson, Bourgon, et al., 2007) and were meant to provide general standards for classifying risk (as discussed above). They were not meant to be synonymous with any particular legal threshold or decision (e.g., civil commitment).

Three quantitative metrics were used to develop the risk levels and can also be used on their own to report and interpret Static-99R scores. These include percentiles (Hanson, Lloyd, et al., 2012), risk ratios (Hanson et al., 2013), and absolute recidivism estimates (Hanson, Thornton, et al., 2016; Lee & Hanson, 2021). Each of these metrics has strengths and weaknesses. Percentiles (e.g., top 10% in terms of risk) are a simple and commonly used metric to indicate relative risk and guide resource allocation. However, because people convicted of sexual offenses are not evenly distributed across percentiles, differences in ranking do not reflect differences in likelihood to reoffend, which may be misunderstood by some users (Hanson, Lloyd, et al., 2012).

Risk ratios (e.g., twice as likely) communicate an individual's risk relative to a reference group (also useful for resource allocation) and are in many ways most linked to what is being assessed by Static-99R (Hanson et al., 2013). However, they may be poorly understood by laypeople (Varela et al., 2014), misleading to interpret in the absence of base rate information (e.g., saying someone is two times more likely to reoffend than the typical person convicted of a sexual offense has a different meaning if the base rate is 4% vs 40%), and systematically lead to higher perceptions of risk than other metrics (Helmus et al., 2018; Hilton & Helmus, 2021).

Absolute recidivism estimates (e.g., 25% chance of recidivism in 5 years) are intuitively easy to understand and commonly reported (e.g., Chevalier et al., 2014; Kelley

et al., 2020) but may be unnecessary for many resource allocation decisions (Harris et al., 2015). Nonetheless, they tend to be of critical importance in the legislative criteria for SVP commitment decisions (ATSA & SOCCPN, 2015). Their key limitation is that they are not consistent across samples and settings (Hanson, Thornton, et al., 2016). This is likely true of all actuarial risk scales (Helmus, 2018). What it means for practice is that absolute recidivism estimates are the least robust risk metric provided by Static-99R and other actuarial scales, and should be reported with care (Helmus, 2021).

How Has it Changed Over Time?

Static-99 was created by combining risk factors from the Rapid Risk Assessment for Sex Offender Recidivism (RRASOR) and the Structured Anchored Clinical Judgment (SACJ-Min; for a detailed review of the history of Static-99, see Fernandez, 2021). Hanson and Thornton (2000) found that the two tools had incremental validity and combined them together to form Static-99. Given the lack of alternatives at the time, Static-99 was adopted rapidly by many jurisdictions. It was revised to Static-99R to incorporate accumulating evidence of a more nuanced relationship between age and sexual recidivism; the authors recommended that the revised version replace the original and they have since stopped updating normative data for the original Static-99 (Helmus, Thornton, et al., 2012).

Additionally, a related scale called Static-2002 was developed in 2002 and its age item was revised concurrently with Static-99, creating Static-2002R (Helmus, Thornton, et al., 2012). Static-2002R was developed to improve upon Static-99 in several ways, such as construct validity, consistency in scoring rules, and accuracy (Hanson & Thornton, 2003). Of Static-99R's ten items, five are identical in Static-2002R (age, non-contact sex offense, male victim, unrelated victim, stranger victim), one has the same definition but different cut-offs (prior sentencing dates), two measure similar constructs but with notably different definitions (prior sex offenses and non-sexual violence), and two items were dropped (ever lived with a lover, index non-sexual violence). Additionally, six new items were added. Although the original intent was for Static-2002 to replace Static-99, after the age revisions,

both scales had equivalent predictive accuracy (Babchishin, Hanson, & Helmus, 2012) so there was no empirical need to switch. Both scales have their advantages and disadvantages but this paper will focus solely on Static-99R, not Static-2002R.

Aside from the revision of the scale (from Static-99 to Static-99R), updates have primarily focused on ancillary products developed to assist implementation and interpretation (for review of these changes, see Fernandez, 2021). Incorporating new research, societal changes (e.g., the internet), new jurisdiction applications, and FAQs, iterations of the scoring manual have increased from 3 pages in length (Hanson & Thornton, 1999) to 15 (Phenix et al., 2000), to 33 (Harris et al., 2003), and to 94 (Phenix, Fernandez, et al., 2016). Starting in 2009, Evaluator Workbooks were provided to compile normative data separately from the coding manual. From this point forward, materials for the scale were divided into the Coding Manual (updated infrequently) to guide scoring the items, and the Evaluators Workbook (updated regularly), to offer data and guidance for results interpretation (for the latest version, see Helmus, Lee, et al., 2021). In the latest version, the absolute recidivism norms were updated to add 10-year data from routine correctional samples (Lee & Hanson, 2021) as well as 20-year projections of risk and a calculator to incorporate reductions in risk based on time offense-free in the community since release from the index sexual offense (Thornton et al., 2021).

Ironically, given the name of the tool, its development and normative data are not static. Iterations of the scale (and related dynamic tools) have evolved, sometimes at a "dizzying pace" (Harris & Hanson, 2010, p. 296). While this can be frustrating and confusing to users, it is a prerequisite of evidence-based practice, which evolves as new evidence accumulates (Dawes et al., 1989). This is why continuing education is important for professional practice. That being said, the developers of Static-99R can be criticized for unclear dissemination strategies, making it more difficult to stay up to speed. Use of the Static-99 website (www.static99.org) helped, but it was not updated with great regularity. The worldwide network of certified trainers should ideally be ongoing resources to the

jurisdictions they train, but trainers retire, leave, or fail to keep up with updates themselves. The amalgamation of Static-99R and other scales into a new non-profit organization (SAARNA: Society for the Advancement of Actuarial Risk and Needs Assessment; www.saarna.org) is intended to improve communication of best practices and resources for the scale. Although there are advantages to widespread use of a scale, quality control and dissemination of updates remain ongoing challenges. Consequently, it is not surprising to hear reports of disparate practices across jurisdictions or evaluators.

Can I Still Use Static-99?

One clear example of how disparate practices can develop pertains to the adoption of Static-99R over Static-99. Although the scale developers recommended Static-99R to replace Static-99 in all places where it is used (Helmus, Thornton, et al., 2012), not all jurisdictions have done so. For example, the Austrian prison system has continued to use the original Static-99 because, in their large dataset, age had a smaller relationship with recidivism and Static-99 performed better than Static-99R (Rettenberger et al., 2013). This could be a sampling fluke (e.g., outlier) or a real difference. Either way, it highlights complicated decisions and trade-offs to consider. Jurisdictions like Austria face a conundrum. They must choose to follow local evidence or the scale developers' recommendation, which cannot be reconciled. There are reasons to privilege local validation evidence, but important drawbacks of this are that the jurisdiction loses out on all subsequent updates (e.g., time free adjustments, updated normative data) or has the duty to independently analyze and publish methodological advances using their databases (e.g., the above-mentioned five-level risk system was recently introduced in the German-speaking language area by re-analyzing the Austrian normative data; Eher et al., 2019). Overall, jurisdictions/evaluators who deviate from the scale developers' recommendations must be aware of the drawbacks and should be prepared to defend that decision with a strong empirically based rationale.

Where and How is Static-99R Used?

STATIC-99R STRENGTHS AND LIMITATIONS

It is unknown how many jurisdictions mandate the use of Static-99R and for what purpose (e.g., treatment referrals, inmates, probationers, registration/notification). Anecdotally, many uses across a large number of jurisdictions have been reported (see discussion on international applications of the scale). Every survey we have seen examining risk assessment has found that the STATIC family of scales is by far the most frequently used for sexual offense risk assessment. Some version of the STATIC family of scales (Static-99, Static-99R, Static-2002, and/or Static-2002R) is mandated in every province and territory of Canada except for Alberta and New Brunswick (Bourgon et al., 2018). The most recent survey (Kelley et al., 2020) used professional list-serves to recruit 119 participants who conducted risk assessments of adults who sexually offended. The most frequently used scale was Static-99R (82% frequently used), followed by STABLE-2007 (42% frequently used). However, roughly 6% reported using the RRASOR and 6% reported using Static-99, despite the authors' disavowal of these previous iterations (Hanson, 2016). Examining 111 risk assessment reports for Dangerous Offender hearings in Canada, Blais and Forth (2014) found that the PCL-R was used in over 95% of cases, even though it is not designed as a risk assessment scale. The next most commonly used scale was Static-99, used in over 60% of cases, which is high given that only 64% of the cases had a current sexual offense. Static-99/R was also reportedly used in over 90% of cases for SVP commitment in the United States (Jackson & Hess, 2007) and preventive detention hearings in Australia (Doyle et al., 2011). Static-99/R was the most frequently used sexual recidivism risk assessment scale in surveys of forensic psychologists and psychiatrists (Neal & Grisso, 2014) and in sexual offense-specific treatment programs (McGrath et al., 2010).

Chevalier et al. (2015) surveyed 109 people who used Static-99R for SVP commitment evaluations in the United States. Overall, 71% used Static-99R in all SVP evaluations, and an additional 25% used them in most. Most reported giving it some (49%) or a lot (42%) of weight in their overall risk assessment. Roughly 61% used at least parts of the reporting templates in the Evaluators Workbook to guide their reporting practices.

Recidivism estimates and risk categories were the most frequently reported risk metric (83% for both), followed distantly by percentiles (35%) and risk ratios (33%), although the latter two had only been published in 2012 and 2013, respectively, not long before the study would have been conducted. Despite absolute recidivism estimates being the least robust risk communication metric available for Static-99R (Helmus, 2018), it was considered the most important metric by 54% of evaluators (with 25%, 7%, and 15% considering risk levels, percentiles, and risk ratios as most important, respectively).

Section 2: How Good Is It? Psychometric Properties

Using a scale should be based on careful consideration of the empirical support for it, including gaps in research. This section clarifies what types of validation evidence are most important, and reviews interrater reliability and predictive accuracy.

What is Important to Consider in Terms of Accuracy and Error?

No risk assessment scale is perfect. But when using any scale, it is important to be aware of the accuracy and error rates. This is important for court admissibility (e.g., *Daubert v. Merrell Dow Pharmaceuticals*, 1993) and standard ethical practice for psychologists (APA, 2013). When evaluating the accuracy of a risk tool, it is important to keep in mind the purpose for which it is developed. One informative document is the Standards for Psychological and Educational Testing, co-authored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (hereafter referred to as the Joint Committee, 2014). Specifically, test developers should clearly delineate how test scores are intended to be used, including the appropriate population(s) and the construct(s) assessed. Test developers should provide a rationale and appropriate evidence for the intended interpretation of test scoring, whereas it is evaluators' responsibility to select an appropriate tool for their purpose. Furthermore, "when a test user proposes an interpretation or use of test scores that differs from those supported by the test developer, the responsibility for

providing validity evidence in support of that interpretation for the specified use is the responsibility of the user" (Joint Committee, 2014, p. 13).

Rather than traditional delineations of validity types (e.g., construct validity, concurrent validity, face validity, predictive validity), the Joint Committee Standards (2014) group everything into the broad term of validity and emphasize that the types of evidence needed to support the validity of a scale depend on the particular use and interpretation of the scale. Not all types of validity evidence are equally applicable or informative. For example, if a scale is being used to determine who gets into graduate school, then the most appropriate validity evidence is to determine whether it does indeed identify students who perform better in graduate school.

What Type of Scale is Static-99R?

Risk assessment scales such as Static-99R are criterion-referenced, prognostic measures (Hanson, 2021; Helmus & Babchishin, 2017). Norm-referenced measures compare individuals to a reference group, often on a single latent construct, with results commonly expressed as some form of percentile rank. In contrast, criterion-referenced scales are designed to predict an outcome (here, an outcome in the future). This distinction has implications for validation efforts. Specifically, internal reliability (particularly measured with Cronbach's alpha) is uninformative because the scale is not designed to measure a single construct; in fact, maximizing predictive accuracy (while also considering parsimony) privileges assessing as many diverse risk-relevant constructs as possible, with no assumptions made about the relatedness of the specific items (Hanson, 2021; McNeish, 2018; Schmitt, 1996). Additionally, risk scales are prognostic as opposed to diagnostic (Helmus & Babchishin, 2017). This means that rather than identifying a current state of affairs, the scale is trying to predict a future event which currently does not exist and may never exist. This is important because prognoses should be communicated along a continuum (e.g., scores, probabilities) as opposed to dichotomously (e.g., this person will or will not reoffend). Given these considerations, the primary validity evidence to consider for

Static-99R is its predictive accuracy; however, in order to achieve good predictive accuracy, it is also necessary that the scale is measured with as little error as possible. One of the most direct ways to examine measurement error is through interrater reliability.

What is the Interrater Reliability?

Given that the items have been the same for Static-99 and Static-99R (except for different cut-points in the age item, which should not impact reliability), interrater reliability should be the same for both versions. Consequently, our discussion considers both. The most frequently reported interrater reliability statistic is the Intraclass Correlation Coefficient (ICC). In terms of heuristics for interpreting ICCs, Cicchetti (1994)'s commonly cited guideline is that values of .75 and above are considered excellent. More recently, Koo and Li (2016) have advocated for more stringent criteria whereby, taking into account confidence intervals, .75 to .90 would be considered good and .90+ excellent. Confidence intervals, however, are more informative about sample size than about the magnitude of reliability, and they are also rarely reported in our field.

In their review of 18 diverse studies/samples, Phenix and Epperson (2015) reported overall ICCs for Static-99 ranging from .79 to .95, with most in the mid .80 range. More recently, Leguízamo et al. (2017) found an ICC of .84 for Static-99 when comparing clinician scores to graduate students. Examining 55 Static-99 scores from prison-based program officers in real-world scoring conditions, Fernandez and Helmus (2017) reported an ICC of .95, with 7 of 10 items scored identically in 90% or more of cases. Looking at Static-99R specifically, ICC's have varied between .89 to .96 (Gonçalves, 2020; McGrath et al., 2012; Raymond, 2020; Stephens et al., 2016), generally suggesting excellent reliability.

High reliability, however, cannot be taken for granted. The lowest reliability has been found comparing opposing experts in SVP proceedings in Texas. Murrie et al. (2009) found high reliability when two experts from the same side separately evaluated the case (ICC's between .84 to .95), but ICCs dropped to .64 when comparing two opposing experts. Examining forensic psychologists from Wisconsin Corrections, Skaar (2013) found that more

experienced evaluators (those who scored 100+ previous cases) demonstrated higher Static-99 reliability (ICC = .933) than less experienced evaluators (Static-99 ICC = .855). For Static-99R, however, both groups had similar and excellent reliability (ICC = .975 for the more experienced and .930 for the less experienced). Excellent reliability was also found for the individual items (ICC = .90+ for all items, except .875 for non-contact sexual offenses). Hanson et al. (2014) similarly found differences based on level of experience among corrections and probation officers in California. In this study, 55 evaluators scored a standard set of 14 cases, with an ICC of .85 among more experienced raters (defined as scoring 26+ cases in the past year) and .71 among less experienced raters.

Rice et al. (2014) examined initial and most recent Static-99 scores in a large field sample from Texas, where multiple scorings over time are common.[1] For over 21,000 cases, the ICC was .81, but notable patterns were also found. Percent agreement decreased as the total score increased, and there was a tendency toward regression to the mean (i.e., higher scores got lower when re-scored, and lower scores got higher).

Another important statistic related to interrater reliability is the Standard Error of Measurement (SEM). Using field validity interrater data from California, Hanson et al. (2014) found an SEM of 1.08. What this means is that 68% of the time, evaluators should expect their score to be within one point of the offender's true score, and 95% of the time it should be within two points.

Important Considerations for Interrater Reliability

Static-99R scores appear generally reliable, both in research contexts when scored by trained researchers and in field settings scored by trained clinicians from professionally diverse backgrounds. That being said, experience seems to matter (Hanson et al. 2014, Skaar, 2013), suggesting supervision and mentoring are important for novice evaluators.

[1] Although the authors ruled out cases where scores changed because they reached their 25th birthday (changing the age item) or because of a new index offense in prison, it is possible that some re-scorings included new information.

Also, mistakes may be more common in higher risk cases (Rice et al., 2014), perhaps due to complexity. Higher risk cases, however, are also often higher stakes cases (e.g., SVP commitment), which may be more vulnerable to adversarial allegiances and other types of bias (Murrie et al., 2009). Overall, interrater reliability should be an ongoing goal for improvement, as opposed to something achieved once (e.g., after training) and then neglected.

Boccaccini et al. (2012) note that even with good to excellent reliability, a one-point difference can have important consequences in some decisions. Consequently, it may be useful to consider percent agreement in addition to ICCs, although this is infrequently reported. In field settings, Hanson (2001, as cited in Boccaccini et al., 2012) and Austin et al. (2003) found exact agreement below 50% of the time (43% and 41%, respectively), whereas three other field samples have found agreement in 54%-55% of cases (Boccaccini et al., 2012; Quesada et al., 2014; Rice et al., 2014). Rice et al. (2014) found that total scores were within one point 86% of the time. One potential outlier is Fernandez and Helmus (2017), who found identical total scores in 78% of cases. In this study, the scales were assessed by six correctional program officers, who were generally experienced (e.g., 40+ prior cases) but did not have advanced degrees. The high reliability may be a result of the specific attention to training and ongoing supervision that was routine in this assessment unit, again speaking to the importance of training and quality control processes.

Quesada et al. (2014) found that the most common type of scoring error was failing to follow the coding manual and Rice et al. (2014) found lower agreement for items that required counting. This suggests that supervision has the potential to produce dividends in terms of reliability. All of these observations speak to the importance of ongoing checks and balances, regular audits, and quality supervision, perhaps most particularly in cases that result in higher Static-99R total scores and among less experienced evaluators.

Finally, Murrie et al.'s (2009) finding that total scores are more often discrepant in adversarial contexts (e.g., opposing experts) is not easily addressed through training and

supervision, as these issues are more embedded in evaluators and in the adversarial process than they are in the scale. The thoroughness of the coding manual, however, can minimize these threats to reliability. Averaging the results of disparate assessments might be a method for reducing the effects of adversarial allegiance, but it is not clear how those averaged scores should be interpreted. For example, the manual does not allow for interpretations of scores that are not whole numbers (averaging would lead to decimals in many cases) and there is very limited research literature on the validity of averaged scores (Rice, 2016). If decimal values from average scores were rounded, many people would systematically round that up to the higher score (inflating risk). Overall, efforts are needed to increase awareness of these reliability issues, including further research on how to minimize them. While training and supervision are not a panacea, consultation and independent checks are likely the best available method for minimizing these biases. Although there are some ambiguous situations open to interpretation, most discrepancies in Static-99R scores can be traced to one person scoring an item (or more) incorrectly. It is preferable to try and get the right score at the front end rather than combining disparate scores later. Good record-keeping on the rationale for scoring (particularly for tricky cases or situations) can also improve practice, particularly in adversarial proceedings.

What is the Predictive Accuracy? A Meta-Analysis

There are two types of predictive accuracy: discrimination and calibration (also referred to as relative and absolute accuracy; for further discussion, see Helmus & Babchishin, 2017). Discrimination assesses how well the scale rank orders individuals according to risk. In other words, are higher risk scores associated with higher recidivism rates? It is often assessed with AUCs, among other statistics (Helmus & Babchishin, 2017). AUCs vary between 0 to 1, with 0 reflecting perfect negative predictive accuracy, 0.5 reflecting chance prediction, and 1 reflecting perfect positive predictive accuracy. As an example of interpretation, an AUC of .70 indicates that there is a 70% chance that a randomly selected recidivist from the sample would have a higher risk score than a

randomly selected non-recidivist. AUC values of .56, .64, and .71 are interpreted as small, moderate, and large predictive accuracy effects (following Rice & Harris, 2005, these values are equivalent to Cohen's *d* values of .20, .50, and .80, which are commonly used benchmarks). AUCs are good at summarizing how well the scale ranks individuals in terms of risk regardless of the base rate of recidivism, although AUCs are sensitive to the distribution of predictors; particularly, restriction of range in the predictor variable will reduce the AUC values (Howard, 2017).

Calibration assesses the applicability of the absolute recidivism estimates from the scale in new samples. Calibration is reported less often and clear conventions on statistical approaches do not yet exist, but the E/O index has been one plausible statistic used to assess calibration (Hanson, 2021). The E/O index compares the expected (i.e., predicted) number of recidivists to the observed number. For example, if a risk scale predicted 30 recidivists in a sample based on the distribution of risk scores but there were only 20 recidivists, then it can be said that the scale predicted 1.5 times as many recidivists as there actually were (E/O = 30/20 = 1.5; for further examples and worked out calculations, see Hanson, 2017).

There have been other statistics periodically recommended for predictive validity research (e.g., likelihood ratios, positive/negative predictive accuracy, sensitivity/specificity, false positives/negatives) but they are not appropriate (for a fuller discussion of these issues, see Helmus & Babchishin, 2017). These statistics are ill-suited for risk assessment for myriad reasons, including a focus on diagnosis (as opposed to prognosis) which often involves artificial dichotomization of a scale providing dimensional information and because of undue influence by the base rate of the outcome as opposed to the predictive properties of the scale.

Helmus, Hanson, et al. (2012) reported a fixed-effect meta-analytic average AUC of .69 for Static-99R (the random-effects average was .70) across 22 studies, but roughly two-thirds of the cases had also been used to develop the new age weights for Static-99R, so

there could be some overfitting. Nonetheless, there were two key findings from this early meta-analysis. Firstly, discrimination was stable across diverse samples and settings. In other words, the finding that higher risk scores were associated with higher recidivism rates was consistent regardless of the setting in which Static-99R was applied. In contrast, calibration was not stable across samples. This means the recidivism rate associated with each Static-99R total score varied significantly across samples, and this variability was moderate to large (this was the finding that led to the different reference groups for recidivism estimates). In terms of practical implications, this suggested that evaluators could be confident in Static-99R's ability to rank individuals in terms of their likelihood of reoffending but should be more cautious in using the scale to provide absolute estimates of recidivism probability. More recently, Helmus, Hanson, et al. (2021) reported a meta-analysis for 15 field validity studies of Static-99R (all of which are subsumed in the current meta-analysis), but this will be described further below. There have been no comprehensive meta-analyses of Static-99R validation studies since the scale's development in 2012.

Method

For the current meta-analysis, one co-author[2] searched the keyword "Static-99*" or "Static 99*" in the document text for the following three academic databases: PsychInfo, Criminal Justice Abstracts, and Proquest Digital Dissertations. Additionally, Google Scholar was searched with the same keywords. Results were examined from 2009 onwards as that was the earliest dissemination of the revised scale (Helmus, 2009). Studies were included if there was sufficient information to code an effect size for predictive accuracy for any sexual recidivism in the community among males charged or convicted of a sexually motivated offense. We also reached out to the ATSA list-serve to request any additional or upcoming studies we were not aware of, and ensured we included studies from previous meta-analyses (Helmus, Hanson et al., 2012, 2021). All Static-99R articles were examined in full

[2] Another co-author did some informal searching to check for obvious ommissions.

to determine if there was sufficient information to code predictive accuracy, unless it was clear from the abstract that no recidivism data were available. Online Supplement A provides the reference list for the 52 studies identified as of February 21, 2021. Online Supplement B provides three descriptive tables summarizing the studies (e.g., sample, setting, moderator variables, interrater reliability, outcome data, effect size).

All studies were coded by both the first and fifth author to ensure accuracy. Differences in scoring were consensified prior to analysis. Formal interrater reliability was not conducted because one author completed all coding before the other started. However, given that identification of sample type (particularly routine vs. preselected high risk/need) can be particularly important in court cases (discussed in the section on legal admissibility), this variable was analyzed for interrater reliability. The second coder scored all studies on this variable blind to the first coder's ratings. Studies included in previous normative samples (where sample type was already classified by the scale developers) were excluded. Of the remaining 28 studies, the two authors agreed on sample type for 22 (79%, Cohen's kappa = .661). Depending on the heuristics applied, this level of interrater reliability would be considered either substantial (Landis & Koch, 1977) or fair to good (Fleiss, 1981). Interrater reliability was likely impacted by the lack of training and calibration cases, and the amount of inference required for this task (i.e., few studies clearly reported on preselection or referral processes).

Two studies were coded as included in the Static-99R normative datasets (Olver, Sowden, et al., 2018; Rettenberger et al., 2013), although those studies had additional cases beyond what was included in the norms. Additionally, Olver, Sowden, et al.'s (2018) study contained analyses separated by Indigenous ancestry; results for the combined sample were obtained via email (Olver, personal communication, Feb. 1, 2021). For studies included in the norms for Static-99R (e.g., Hanson, Thornton, et al., 2016), we coded them based on what was used in the normative datasets, as those data were cleaned and more easily accessible than the original documents. If information was available on both charges

and convictions, we coded charges as a more inclusive measure of recidivism (also, it generally had higher statistical power). We generally coded effects from longer follow-up periods unless it resulted in a non-trivial loss of information. The full meta-analysis dataset, with all coded variables is available from osf.io/6af3t.

Meta-analyses followed the formulae of Borenstein et al. (2009), using the "metafor" package in R program (Viechtbauer, 2010). Although random-effects analyses are often conceptually preferable due to greater generalizability, they are unstable when the number of studies is small (< 30, Schulze, 2007), which was often the case in subgroup analyses. In contrast, however, the greater the variability across studies (and in our analyses, moderate to large variability was not uncommon), the more unrealistically narrow the confidence intervals from fixed-effect analyses are. Consequently, we reported both fixed- and random-effects analyses. Confidence intervals around the weighted average AUC give some indication of the robustness or confidence in the findings; when the number of studies is close to or greater than 30, then the random-effects results should be privileged.

Variability in findings across studies was reported using Cochran's *Q* statistic and the I^2 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). Moderator analyses used the $Q_{between}$ statistic (measuring variability between levels of the moderator) from both fixed- and random-effects models within subgroups (Borenstein et al., 2009). Given the amount of variability found, fixed-effect moderator analyses were overpowered (even small differences were statistically significant); we therefore based our conclusions primarily on random-effects moderator analyses. No obvious outliers were detected. All analyses were independently run by the first and fifth author to ensure accuracy.

Results

The 52 studies represented a total of 56 samples and 71,515 individuals. For Boccaccini et al. (2017), we reported results separately for cases released before and after the 2003 coding manual was released as this was associated with a significant difference in interrater reliability (Rice et al., 2014) and predictive accuracy. Following Hanson, Thornton, et al. (2016), the Knight and Thornton (2007) study was coded as two separate subsamples (those assessed but not civilly committed, and those civilly committed). Smeth (2013) reported two separate samples of men on probation, and Spiranovic (2012) reported results separately for Indigenous and non-Indigenous individuals without a composite sample.

All studies were of adult men (although a small number of people under 18 may have been included in some samples). Nearly half the studies were from the United States (k = 24; 42.9%) and a quarter were from Canada (k = 14; 25%). There were four from Australia (7.1%) and two each (3.6%) from the United Kingdom, Germany, and Sweden. There was one study each from the Netherlands, Denmark, Norway, Austria, New Zealand, Belgium, Singapore, and Switzerland. More than half of the studies did not provide information on whether the sample was mostly treated or untreated (k = 32). Of the remaining 24 studies, the majority were mostly treated (k = 16; defined as 75% or more of the sample), seven samples were mixed, and only one was identifiably mostly untreated (<25%). No studies reported meaningful variation from the coding rules (e.g., missing access to victim information), although two studies reported analyses of risk categories as opposed to total scores (McGrath et al., 2012; Reeves et al., 2018). The unweighted average follow-up across studies was 6.8 years, ranging from 1.8 to 16.8.

The overall meta-analysis (summarized in Table 1) found moderate predictive accuracy for Static-99R (fixed-effect AUC = .68, random-effects AUC = .69, k = 56, n = 71,515). These results are nearly identical to the previous meta-analysis (Helmus et al., 2012). One difference, however, is that the variability across studies in the current meta-analysis was statistically significant and moderate in magnitude (I^2 = 69.8). This means that the discrimination of Static-99R varied meaningfully across studies, which was not the case

in the normative datasets previously used for the scale (Hanson, Thornton, et al., 2016; Helmus et al., 2012). Specifically, roughly 70% of the variability in AUCs was more than what would be expected purely based on sampling error. Encouragingly, all studies found AUCs that were at least small in magnitude ($\geq$ .56). Of the 56 studies, 15 had small effects, 20 had moderate, and 21 had large. Importantly, AUC magnitudes are heavily impacted by the range of the predictor (Static-99R scores; Howard, 2017) whereas the previous meta-analyses used statistics more robust to such influences; consequently, they provide a more realistic summary of variability in discrimination. Nonetheless, moderator analyses are important to examine and explain variability. We examined several potential moderators (see Table 1). The fixed-effect results were overpowered and all moderators but one were statistically significant, often with differences small in magnitude.

Focusing on the random-effects analyses within subsamples, two moderators remained statistically significant. Sample type was significant, with AUCs lowest from samples preselected as high risk/need (random-effects AUC = .65, k = 7). This is consistent with some analyses from the normative data for Static-99R (Hanson, Thornton, et al., 2016). This is likely due to restriction of range on both measured (Static-99R) and unmeasured risk. In samples with this amount of screening to select a small minority of the perceived highest risk individuals, there would likely be few low-scoring individuals, and most would also have higher levels of criminogenic needs. This does not mean that Static-99R would not be useful for people in high risk/need settings. Many of the other samples contained high risk/need individuals, but these were integrated with a broader distribution of Static-99R scores and external risk factors. It is also unlikely that Static-99R would be predictive for samples of all individuals convicted of sexual offenses, but then stop being predictive once someone has been identified as unusually high risk/need. Consequently, this difference can be viewed primarily as a statistical artifact.

Studies that referenced the coding manual had significantly higher AUCs (random-effects = .72, k = 37) than studies that did not (.69, k = 19), although this difference was

small. This is likely a rough indicator of quality of coding. Random-effects AUCs were not significantly related to country, whether the study was used in Static-99R norms or in the original development of Static-99R, whether the scale developers were co-authors of the study, appropriate training practices noted, or recidivism criteria (charges vs convictions).

Only nine samples provided information on calibration (see Online Supplement B, Table B3), in all cases using a fixed 5-year follow-up. Studies compared their data to routine norms from Static-99R, except DeClue and Rice (2016) who reported comparisons to both routine norms and preselected high risk/need norms (with the latter reflecting a better description of their sample and used in our summary). The scale was in the direction of overestimating recidivism for seven out of nine studies. Given that the variance (used to determine weight in meta-analysis) of the E/O index is non-symmetrical and difficult to estimate, we summarized calibration data weighting by sample size (combining the number of observed and expected recidivists across the nine studies). Static-99R normative data significantly overestimated recidivism (E/O = 1.73, 95% CI of 1.63 to 1.84). Specifically, the norms estimated 73% more recidivists than there actually were. One important limitation of these comparisons, however, is that for most studies examining calibration, the quality of the observed recidivism data may be lower than the normative data. Specifically, many of the studies relied on only one source of recidivism data, often local (not national), and without details on offense circumstances to identify non-sexual offenses or technical violations which stemmed from sexual crimes. Additionally, none of these studies examined calibration using the 2021 Evaluators Workbook (Helmus, Lee, et al., 2021), which has slightly lower recidivism rates. However, the difference in the new Workbook is relatively small and would be unlikely to meaningfully change these findings.

Section 3: Field Issues and Misuses in Applying the Scale

What About Field Validity and Adversarial Allegiance?

Although the meta-analysis above presents a comprehensive summary of Static-99R research to date, it is important to note that most of those studies were conducted under

research conditions where the scale is often scored retrospectively based on file information by trained research assistants (often students selected for conscientiousness). This helps control for potential biases (allegiance effects). In contrast, field studies examine the accuracy of scales scored in routine conditions to be used for real-world decisions. Evaluators have competing time demands and likely vary in their conscientiousness and commitment to accuracy. For example, they may believe their judgment is more valuable than the scoring rules. Consequently, Edens and Boccaccini (2017) have pointed out that in some respects, research studies show the potential reliability and validity of a scale, whereas field studies show the reliability and validity in practice, as currently implemented.

Helmus, Hanson, et al. (2021) reported a meta-analysis specifically of field validity studies of Static-99R (k = 15). The average weighted AUC was .66 in the fixed-effect model and .69 in the random-effects model, representing moderate accuracy. There was significant and moderate variability in the accuracy across studies and two moderator variables were examined. An earlier meta-analysis found that effect sizes for Static-99 tended to be higher in studies co-authored by someone on the development team, suggesting author allegiance biases (Blair et al., 2008). Alternatively, studies with scale author involvement could reflect greater fidelity to training and implementation of the scale (Andrews et al., 2011). Helmus, Hanson, et al. (2021) found no difference in predictive accuracy for studies with or without a co-author from the scale's development team. In contrast, predictive accuracy was significantly higher (AUC = .72, a large effect size) in studies with appropriate training processes in place compared to studies where the training processes were not reported (AUC = .64 to .65, depending on the statistical model). This suggests that quality training is more important than scale author involvement, and that large effect sizes are achievable in field settings if staff are appropriately trained.

One field issue that has received particular attention is adversarial allegiance. Correlational and experimental research has found a tendency for evaluators' risk scores to be biased towards the party that retained them in an adversarial system (Murrie et al.,

2009, 2013). This reduces interrater reliability in adversarial proceedings, decreasing the credibility of the scale and the evaluators. Murrie and colleagues (2009, 2013) found smaller adversarial allegiance effects for Static-99 compared to the PCL-R. This may be partly because Static-99 has more structured, detailed, and objective scoring rules. Additionally, the items measured by Static-99 are generally more objective and clear-cut (e.g., any male victim) to assess than items from the PCL-R (e.g., shallow affect).

At the individual level, another important factor influencing accuracy of Static-99R (and presumably other risk scales) is the conscientiousness or level of commitment of the evaluator. In a prospective study of parole and probation officers across Canada, Hanson et al. (2015) found that Static-99R scores had significantly higher predictive accuracy (AUC = .801) for officers who completed every stage of the research process (submitted a Static-99R, STABLE-2007, and ACUTE-2007 assessment, along with an indication of whether they thought an override was warranted or not) compared to officers who missed one or more steps (AUC = .685). This was within a sample that had received appropriate training, suggesting that the benefits of appropriate training and staff commitment may be additive, with field validity results as high as .80 or above under optimal conditions.

How is the Scale Misused?

Concerns about the misuse of psychological measures, particularly within forensic contexts, have been longstanding (Rogers, 2003; Wakefield & Underwager, 1993; Zapf & Grisso, 2012). Static-99R is not immune to problems of misuse. In addition to field implementation issues and biases discussed above, this section is informed by frequent misuses observed by the current authors as well as insights gleaned from email questions submitted to the staticquestions@gmail.com email account. Common misuses fall into three main categories: 1) Scoring and application errors, likely related to inadequate or absent training; 2) Problematic user choices, including test version, norms, and reporting metrics; and 3) Interpretation problems and biases, including integration of additional information, and the use of clinical overrides.

Scoring Errors

Static-99R has a 94-page coding manual, which is unusually detailed compared to other sexual risk scales. Training from a certified trainer is recommended in addition to reading and applying the coding manual (Phenix, Fernandez, et al., 2016). Despite these recommendations, we regularly encounter scoring and application errors with the scale, which may often be traced to inadequate or absent training.

Inappropriate Applications

Static-99R is not applicable to all cases; the coding manual outlines appropriate populations. Common inappropriate applications include:

- Using Static-99R for people whose only sexual offense is Possession of Child Pornography or Statutory Rape (or other Category "B" sexual offense in Static-99R scoring rules).
- Applying Static-99R to adults when the last sexual offense occurred at age 16 or younger (or 15 or younger prior to the 2016 coding manual).
- Assuming Static-99R does not apply to people with solely non-contact offenses in their sexual offense history (e.g., exhibitionists).
- Applying Static-99R to women who have sexually offended, contrary to guidance and evidence (see Marshall et al., 2021).

Incorrect Item Scoring

To the untrained eye, the Static-99R coding sheet (list of items) seem disarmingly simple. Yet some scoring rules are deceptively counter-intuitive, particularly in complex cases (this is common feedback from trainings). Despite attempts to clarify these issues in the coding manual, common scoring errors include:

- Incorrect identification of the index sexual offense, particularly identifying clusters. This can have ripple effects in scoring other criminal history items.
- Error or failure in identification of non-sexually violent offenses, particularly those coded (non-intuitively) as both non-sexual violence and sexual violence.

- Incorrect coding of "Age at Release" based on current age, or age at release from the current offense when it is not the same as the index sexual offense.
- Inclusion of post-index information in item scoring (e.g., new relationships).
- Misunderstanding the definition of a Category "A" offense (identifiable victim) as meaning the evaluator must know the personal identity or name of the victim.
- Erroneously concluding that all Category "B" offenses are also non-contact offenses, or that all Category "A" offenses are contact.
- Insufficient familiarity with item coding exceptions for specific offense types (e.g., how to score sexual offenses that involve failure to disclose HIV status or posting/sharing of revenge pornography).

Absent or Inadequate Training

The Static-99R Training Guidelines (SAARNA, 2021) suggest that training on Static-99R should be between 8 and 12 hours depending on the experience and needs of the participants. However, it is further noted that competent use of the measure requires both training and an in-depth reading of the coding manual. Other problems related to training include:

- Users downloading the Static-99R coding form from the internet and scoring it based on a cursory review of the coding manual or based on their interpretation of the items on the scoring sheet without additional training.
- Jurisdictions creating their own abridged or replacement coding manual that does not reflect the details required to code items. Sometimes it may be intended as a supplement to the coding manual but then becomes a replacement.
- Users informally trained by colleagues rather than by certified trainers.

Problematic User Choices

Test Versions

In 2012, the scale developers recommended that users switch to Static-99R (Helmus, Thornton, et al., 2012) and future updates to the scale (e.g., norms) have been

restricted to the revised version. As discussed earlier, updates can lead to confusion and disparities in practice, related to lack of knowledge about updates or deliberate rejection of them. There are limited circumstances where such decisions have empirical merit (e.g., the discussion of Austria's findings, above), but generally if users do not follow scale developer recommendations, the onus is on them to empirically defend their practice. Misuses in this category include:

- Continued use of Static-99, or the simultaneous use of both Static-99 and 99R
- Privileging norms that provide the highest recidivism estimates (e.g., the originally published Static-99 norms for older individuals), despite the data suggesting they are outdated and will overestimate risk (e.g., Helmus, Thornton, et al., 2012).
- Selective use of the original Static-99 with specific individuals, by suggesting that a particular older individual is an "outlier" and therefore the protective effect of age is irrelevant (for contrary evidence, see Thornton & Helmus, 2021). In some cases, Mattek and Hanson's (2018) case study of a man actively reoffending into his 80s has been cited to disregard the aggregate data on age.

Recidivism Norms

A unique feature of actuarial scales is their inclusion of probability estimates for the outcome (Meehl, 1954), although these probabilities have not been stable across diverse samples and settings (Helmus, Hanson, et al., 2012). To at least partially address the sample differences in recidivism norms, the Evaluators Workbook provides two sets of norms: from routine/complete correctional samples and preselected high risk/need samples. The developers recommend that users should default to the routine norms (Hanson, Thornton, et al., 2016). Use of the high risk/need norms should be based on a judgment concerning the density of external risk factors, typically using a structured method for assessing other risk factors such as dynamic risk. Preselected high risk/need samples would

likely be characterized by unusually high levels of dynamic risk. Observed misuses related to users' choice of norms include:

- Use of preselected high risk/need norms without an appropriate justification or structured examination of external risk factors. Sometimes a vague reference to a single external/dynamic risk factor has been used to justify the high risk/need norms, rather than a balanced consideration of the density of those risk factors.
- Routinely using norms that are more favourable to the retaining party (i.e., adversarial allegiance). Specifically, evaluators hired by the defense are more likely to use the (lower) routine norms and evaluators hired by the prosecution are more likely to use the (higher) high risk/needs norms. This observation is consistent with survey data from SVP evaluators (Chevalier et al., 2015).
- Use of outdated norms due to lack of awareness of research updates.
- Use of "local" norms that lack robust methodology and sufficient follow-up periods.
- Although Babchishin, Hanson, and Helmus (2012) and Lehmann et al. (2013) have supported averaging the results from Static-99R and Static-2002R (distinct and incremental risk scales despite their intercorrelation), this finding has been misinterpreted to support averaging appropriate and inappropriate scales/norms. Specifically, people have averaged outdated norms with updated norms, or included the RRASOR in their averaging. Including more does not always improve the assessment; in fact, adding invalid/outdated data degrades the defensibility of the assessment. It is important to note that the developer of the RRASOR has long disavowed its use (Hanson, 2016) and both Hanson (2016) and Thornton (2016) have disavowed use of the original Static-99 norms. Unfortunately, we continue to see examples of their use.
- Reporting the base rate of sexual offending rather than norms associated with a specific Static-99R score. Specifically, we have seen evaluators critique and reject

the normative data and then report a base rate (not accounting for risk) instead. Here, it should not be said that Static-99R was used as a risk tool as it was rejected, and the final reported estimates are not based on the scale at all.

Reporting Metrics

In addition to the Static-99R total score (which can be considered the raw test result), the Static-99R Evaluators Workbook provides four potential risk communication metrics: standardized risk level, risk ratios, percentiles, and absolute recidivism estimates. The Workbook encourages users to report metrics that are pertinent to the purpose of the assessment, noting that for most treatment, management, and supervision decisions standardized risk levels, risk ratios, and percentiles would likely be sufficient. Observed misuses of the available risk communication metrics include:

- Selective presentation of metrics perceived to reflect higher or lower risk, possibly based on adversarial allegiance. For example, risk ratios often sound more perilous than absolute recidivism estimates (Helmus et al., 2018).
- Incorrect interpretation of reporting metrics due to insufficient familiarity with statistics (e.g., equating the 90th percentile with a 90% likelihood of recidivism).
- Use of outdated risk categories (Hanson, Babchishin, et al., 2017).
- Incorrect interpretation of the total score as a proportion of possible risk. For example, associating a score of 6+ (the highest category) as representing a 100% risk of reoffending and consequently interpreting a score of 3 as a 50% risk of reoffending.

Interpretation Problems and Biases

Integration of Additional Information

Ultimately, a risk assessment score (e.g., Static-99R) is not the same thing as a comprehensive risk assessment (Hanson, 2009). Evaluators must consider the results of the scale in terms of the bigger picture of the assessment, such as how well the evidence from the scale applies to the individual being assessed, how well the scale addresses the referral

question, and what information is not assessed by the risk scale (Hanson, 2009; Helmus, 2021). What additional information needs to be considered may depend on the purpose of the assessment. For example, it would be irresponsible to conduct a civil commitment evaluation based solely on Static-99R, without also using a structured risk scale to assess dynamic risk factors. However, static-only assessments may be fine for routine triage of large numbers of cases. Other issues external to Static-99R may be necessary to consider, including severe individual specific limitations, such as serious physical disabilities or severely failing health. Obviously, stated intentions to reoffend or clear evidence of offense planning (compiling items for a rape kit) would also warrant special consideration. Static-99R can be misused either through myopic tunnel vision that fails to contextualize Static-99R as part of a broader assessment, or by minimizing Static-99R results in the presence of potentially redundant or irrelevant external information.

Clinical Overrides

Paradoxically, although actuarial risk scales like Static-99R cannot account for all risk-relevant information, every study examining the use of professional discretion to adjust actuarial results have found that overrides degrade predictive accuracy (Cohen et al., 2020; McCallum et al., 2017; for review and discussion, see Helmus, 2021). Observed misuses relative to clinical overrides include:

- Use of unstructured or instinctual overrides to the Static-99R score. This can introduce all kinds of unchecked biases. It also tends to lead to overestimation of risk.
- Opining that the "true risk" in a case is higher or lower based on the presence of a factor that has not been empirically related to risk (e.g., offense admission or denial; e.g., McCallum et al., 2017).
- Overriding or overemphasizing the importance of age. Although age is a robust risk factor for recidivism, it is already incorporated in Static-99R and the development research found that after scoring Static-99R, age did not

incrementally improve prediction (Helmus, Thornton, et al., 2012). Nonetheless, some evaluators still adjust their results based on age (effectively double-counting it) or make empirically unfounded arguments that the reduction of risk based on age does not apply to their case (for further discussion, see Thornton & Helmus, 2021). Although it is possible that future, more nuanced research may find more optimal adjustments that better distinguish between different types of older individuals (e.g., those with more recent versus historical offenses), the current Static-99R data are based on a good representation of these different situations and should be considered broadly applicable to all age groups.

- Including constructs correlated with existing Static-99R items (e.g., diagnosis of pedophilia, psychopathy, high victim count, large number of instances of sexual offending) as a justification for an override, typically to higher risk. It is likely that overrides tend to degrade accuracy in part because evaluators overweigh a single piece of information (e.g., psychopathy) relative to a risk scale (or scales) that considers numerous risk factors already, likely correlated with the external factor. Or, the external factor may not add incrementally to the evaluation of risk (for further discussion and examples, see Helmus, 2021).
- One concerning example of the above is using the meta-analysis by Hawes et al. (2013) as an argument for overriding Static-99R results based on the PCL-R. Hawes et al. (2013) found that offenders high on both psychopathy and a measure of sexual deviance were more likely to sexually reoffend than other groups. However, sexual deviance was measured as a single construct, whereas Static-99R measures sexual criminality, general criminality, and youthful stranger aggression, the latter two of which would likely be correlated with psychopathy, particularly Factor 2 scores (Brouillette-Alarie et al., 2016; Brouillette-Alarie et al., 2018). Consequently, this meta-analysis does not speak to the added value of considering the PCL-R after having scored Static-99R. To our knowledge, only

one study has tested the incremental validity of the PCL-R over Static-99R (Looman et al., 2013); however, this was an unusually preselected high risk/need sample, with an average Static-99R score of 5.4 (compared to a median score of 2 in routine correctional samples). In this study, the PCL-R did not add unique information in predicting sexual recidivism after controlling for Static-99R scores. It did add uniquely in predicting violent recidivism, which is not an outcome that Static-99R is designed to predict. Generalizability of these results is unknown and for the prediction of violent recidivism, this might suggest consideration of both Static-99R and psychopathy, not overriding one based on the other. Looman et al. (2013) did find a significant interaction between psychopathy and sexual deviance in predicting violent recidivism among rapists after controlling for Static-99R, but with such low statistical power and such a high risk sample, the results are difficult to interpret. Specifically, Looman et al.'s (2013) findings were in the direction such that high psychopathy was protective against sexual recidivism for offenders with high sexual deviance.

How Can We Guard Against Misuses of Static-99R?

As noted, there are numerous potential misuses of Static-99R. Some reflect errors or ignorance, and others potentially reflect conscious or unconscious biases of the evaluator. Particularly given that the scale is non-proprietary, it is not feasible to police or enforce its use. However, incorrect uses can leave individuals or jurisdictions open to court challenges and ethical or license-related complaints. While it is impossible to stamp out all misuses, there are mechanisms to increase the quality of Static-99R assessments, both at the systems-level and the individual-level. As discussed previously, individual commitment to the process matters (Hanson et al., 2015), as does the training process at the systems-level (Helmus, Hanson, et al., 2021). Furthermore, it is unlikely that all training is equal. There are now Training Standards in place for the scale, which include important considerations such as ideal length, number of participants, and qualification tests (SAARNA, 2021).

Training should also be ongoing, with refreshers and quality control processes in place (e.g., for examples, see Fernandez et al., 2014).

Structured scoring guidance can also improve quality. Notably, findings from Texas suggest significant improvements in both interrater reliability (Rice et al., 2014) and predictive accuracy (Boccaccini et al., 2017) after the 2003 scoring manual was introduced. The 2016 manual is even more detailed. This is likely partly why Static-99/R is more resilient to adversarial allegiance effects compared to other assessment scales (Murrie et al., 2009; Murrie et al., 2013); notably, however, resilience does not mean immunity. It is important to try and adhere to best practices in risk assessment, including encouraging appropriate training, evaluator conscientiousness, and bias-minimization techniques. This could include methods such as blind referrals (where possible), refresher trainings, and ongoing resources (e.g., webinars, FAQs, sample cases) to enhance scoring fidelity.

Although these techniques may help, bias and adversarial allegiance are difficult to combat. These issues, discussed at length in recent research on forensic assessment (e.g., Guarnera et al., 2017; Zapf & Dror, 2017), affect patterns of misuse across many forensic assessment instruments. Given their prevalence, and their deleterious consequences for people accused of sexual offending, efforts to increase awareness of misuse and training on appropriate use of Static-99R should be extended to parties who can intervene in cases where biased evaluations are most likely to occur. Namely, developing material specifically for attorneys and judges handling sexual offending cases may be necessary to reduce the likelihood of misuse that goes unnoticed and unchallenged.

Can It Be Applied Internationally? Considerations and Observations

It may take a few years for translations and implementations to appear, but Static-99R has now been adopted across a number of different languages, countries, and jurisdictions. No comprehensive list is maintained, but we are aware of at least four official translations (i.e., led by a certified trainer fluent in the translated language; Dutch, French, German, and Latvian), and six unofficial translations (Chinese, Finnish, Korean, Polish,

Spanish, and Swedish). Although the scale was developed in Canada and England, we are aware of its international application in at least the following 29 countries: Austria, Australia, Belgium, Bermuda, Estonia, Finland, France, Germany, Hong Kong, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Macau, Namibia, the Netherlands, New Zealand, Norway, Poland, Scotland, Singapore, Slovenia, South Korea, Sweden, Switzerland, Taiwan, and the United States.

There are advantages to translating and adapting risk scales to new countries/languages as opposed to developing new scales: it is efficient (no need to re-invent the wheel) and facilitates collaboration and timely advances, and provides useful cross-cultural data (de Vogel and de Vries Robbé, 2016). As the sixth author of this paper has been involved in translating and adapting these scales into new countries/languages, we can also confirm that in retrospect these applications can be fraught with naïveté and inexperienced guesses. It is difficult to foresee all the challenges and pitfalls of translating and applying a scale to a new country, culture, and language, especially as responses to sexual offenses may evolve at different speeds and stages. Additionally, research and practice culture may also differ, such as in terms of preferences for actuarial approaches (versus other assessment methods).

But even if there is a broad willingness to implement and use actuarial instruments, there are methodological and psychometric challenges (for further discussion and recommendations, see de Vogel & de Vries Robbé, 2016): First and foremost, it is necessary to professionally translate the coding manual. There are two different translation strategies available: the literal translation (i.e., the word-by-word translation ideally including back translation, which is usually done for psychological self-report measures) or the adaptation of the instruments, which considers more strongly the juridical and cultural differences. The first process could be done by a professional translator (or ideally, multiple translators) with no experience in the criminal legal system or risk assessment. In the second process, a skilled evaluator (ideally a certified trainer on the scale) is necessary in order to understand

the principles behind the rules and apply them to unique situations in the new language/country. Generally, there is no silver bullet for translation processes, but either approach has advantages and disadvantages. It can be beneficial to consult with colleagues who have completed this process successfully before, as well as the scale developers.

The second step is cross-validating the translated instrument using a local sample (e.g., Rettenberger et al., 2013). Even with a high-quality translation (which should not in itself be taken for granted), there is no presumption that the scale will work in a new jurisdiction. The generalizability of risk scales (and this may even include application across different states/provinces/territories within a country) rests on several assumptions (Helmus et al., 2011): consistency in the outcomes being measured, the underlying risk factors, and the ways of measuring those risk factors.

The third and final step is probably the most difficult and at the same time the most important one: determining if local norms are necessary. This would be the case if local sexual recidivism rates (after accounting for Static-99R) differ significantly from the international normative data, and if yes, it requires collecting psychometrically appropriate normative data, based on appropriate sample sizes (Hanson et al., 2016).

Section 4: Is it Legally Admissible?

Legal Admissibility Considerations

Shortly after it was introduced, Static-99 (and later, Static-99R) were routinely adopted for sexual recidivism risk assessment, including in civil commitment cases under Sexually Violent Predator (SVP) laws, in both U.S. state and federal levels. Given its frequency of use, it is unsurprising that the admissibility of Static-99 instruments has been challenged in at least 20 state and federal jurisdictions, as well as in Canada. In the United States, challenges are generally governed by one of at least two distinct standards: *Frye v. United States* (1923) or *Daubert v. Merrell Dow Pharmaceuticals* (1993). The "general acceptance test" set in *Frye* emphasizes scientific consensus to determine admissibility. The more detailed *Daubert* standard includes general acceptance as well as ensuring the

scientific method be tested, subjected to peer review, have a known error rate, and standards for use (Hilbert, 2018). The *Daubert* standard assigns the trial judge the task of determining whether to admit expert testimony (or data from administration of the instrument) into evidence based on whether the underlying reasoning or methodology is scientifically valid and can be properly applied to the case at hand (*Daubert v. Merrell Dow Pharmaceuticals,* 1993). It is important to note that none of the Daubert standards are specifically required. Rather, the Daubert inquiry is flexible and will vary depending on the facts of a specific case. Consequently, *Daubert* admissibility challenges around Static-99/R have been received differently by courts.

It is also important to contextualize the *Daubert* criteria. The consideration of a known error rate is not necessarily the same as having a low or near-zero error rate. The precise error rate goes to weight of evidence, which is often evaluated by juries or judges after the evidence has been deemed admissible. In other words, a risk scale need not have high accuracy to be admissible, yet juries and judges may use information about accuracy in deciding how much to trust the evidence in their ultimate decision on the case. It is also important to consider the error rate of alternatives; a sub-optimal error rate may be preferable to unstructured clinical judgment, which has an error rate that is known with less precision and is almost certainly higher. This meta-analysis found moderate accuracy for Static-99R, and high accuracy has been found in field studies adhering to appropriate training processes (Helmus, Hanson, et al., 2021).

It can be helpful to gauge risk scale effects and error rates relative to assessments or interventions in other fields (e.g., medicine), which are commonly accepted as 'routine' practice. Marshall and McGuire (2003) contextualize effect sizes for sex offender treatment with a variety of other interventions in diverse fields. Large effect sizes are rare, but can be found in treatments for depression or public speaking anxiety. Moderate effect sizes (such as those for Static-99R) are more common for many types of therapy, which are routinely accepted for addressing diverse issues (e.g., marital problems, mental health issues). Small

effects are found for bypass surgery, which is highly dangerous and invasive, but commonly used to treat heart disease to reduce risk of heart attack and stroke. Other commonly-accepted medical interventions have even less-than-small effect sizes, in some cases quite trivial, such as aspirin to reduce heart attack, or chemotherapy to treat breast cancer (the latter was as of the early 1990s; it is quite possible that effects have increased since then).

Summary of Legal Admissibility Challenges

Our search identified 85 cases in which Static-99 or Static-99R faced an admissibility challenge. Cases were identified through searches of legal databases (i.e., LEXIS, Westlaw, Casetext, Justia, & CanLII) using all combinations of the keywords "admissibility," "Daubert," or "Frye" and "static 99" or "static-99," along with outreach to legal colleagues to identify impounded cases. Two cases were from Canada and the rest were from the United States. With few exceptions, our search generated cases in which the trial court's decision on admissibility was challenged at the Appellate or Supreme Court level. As such, unchallenged decisions at the trial court level, which may have an impact on how practices evolve in jurisdictions, are likely not captured in this review. Given the evolution of case law and practice in this area, additional relevant cases may exist, and still others may have been litigated since our review was completed. Lastly, given the methods of our search (e.g., reaching out to colleagues), we do not intend for this to be a fully comprehensive or replicable search; nonetheless, we believe it is a helpful resource and summary.

Online Supplement C presents a summary of the 83 U.S. cases, organized by standard (*Daubert*, *Frye*, then other), and then by jurisdiction and date. In organizing states by standards, we first sorted states based on internet searches with a heavy reliance on the existing categorization provided by the Expert Institute since it is updated regularly (Funk, 2021). We then cross-checked the accuracy of this information with local statutes, case law, and/or by consulting District Attorneys. When discrepancies emerged, we settled them by favoring the published statute, case law, and District Attorney opinion. We also accounted for changes in standards, as some jurisdictions moved from *Frye* to *Daubert* at some point

after the *Daubert* decision. That is, the standard listed for each case is the one that was applied to the specific case decision at the time.

Despite the frequency of challenge in the U.S., Static-99/R was ruled fully inadmissible in only four cases (5% of the time). These cases concerned Static-99 (not Static-99R) and were from 2003 or earlier, when there was little to no research on the scale. The scales were ruled partially admissible in another five cases. Additionally, Static-99 was admissible but Static-99R was inadmissible in one case, whereas the opposite decision was reached in another case. Three decisions deemed the scale admissible with some reservations, while two cases discounted the Static-99R results. Below we highlight key admissibility challenges to Static-99/R. Although the vast majority of court cases ruled the scale admissible, we highlight the cases where Static-99/R was ruled inadmissible or only partially admissible. For full coverage of the cases, please see Online Supplement C.

Admissibility Challenges under *Frye*

Across nine states, nineteen early admissibility challenges focused on the novelty of the instrument as evidence of a lack of general acceptance in the scientific community. Between 2000 and 2011, only one state court (Illinois, in *Re Det. Of Bolton*, 2003) ruled Static-99 inadmissible because it had not gained general acceptance. In eight cases (six in Illinois, and one each in California and Washington), courts ruled that *Frye* did not apply because actuarial instruments did not involve a scientific method or principle, admitting Static-99 results without an admissibility hearing and without scrutiny. In the same time frame, Static-99 was excluded in two cases from New Jersey and Illinois on the basis of improper scoring (*Re Detention of Hargett*, 2003) and improper application of the scale to someone under age 16 at time of their sexual offense (i.e., age 15; Re *Civil commitment of JP*, 2001). Early on in its use, eight additional challenges focused on the instrument's reliability and accuracy, or its inability to capture dynamic risk factors or therapeutic change, but none of these challenges were successful. One final challenge of the 2009 norms and sample types (i.e., Routine and Preselected high risk/need) resulted in a partial

admissibility decision, whereby categorial risk and Static-99 score were admissible in an SVP disposition hearing, but the new recidivism norms (from either sample type) were not (*State v. Rosado*, 2009).

As research accumulated, admissibility challenges under *Frye* became less frequent, with only 15 challenges found between 2012 and 2020. Seven focused on general acceptance and none were successful. Other challenges centered on Static-99R's reliability, accuracy, and predictive validity. In one case, *New Jersey v. CB* (2020), the defendant sought to admit Static-99R results, but the N.J. trial court (and affirmed by the Appellate court) rejected this information based on concerns around generalizability to individuals offending in New Jersey, and worries that Static-99R did not capture all the relevant risk factors. In contrast, another New Jersey trial court (and affirmed by the Appellate court) admitted Static-99R results over the defendant's objections, despite the court's own concerns over predictive validity (*In re Commitment of C.R.M*, 2014).

Admissibility Challenges under *Daubert*

In parallel, Static-99 was subject to 18 admissibility challenges under Daubert or similar standards between 2000 and 2011. The majority of these challenges (13) focused on reliability, accuracy, and predictive validity, with many centering on interrater reliability and "only-moderate" accuracy rates (e.g., *United States v. McIlrath,* 2008, p.7; *United States v. Ellis,* 2010, pp. 6-7). In most cases, the courts found that reliability and accuracy issues went to the weight of the evidence rather than its admissibility, or the scale was accepted because the results were presented in the context of a clinical evaluation, which included other factors. In an early Iowa case, novelty and lack of research resulted in inadmissibility (*In re Det. of Johnson*, 2000; affirmed by the Appellate Court). As described in *McIlrath v. United States* (2008), McIlrath petitioned to have his sentence vacated arguing that his attorney failed to demonstrate the admissibility of Static-99. However, the U.S. district court discounted Static-99 results, stating that "the test falls very short of reliably predicting a particular defendant's likelihood of re-offending" (*McIlrath v. United States,* 2008, II.

Ineffective Assistance of Counsel #8). An earlier appeal of the same criminal sentencing case (*United States v. McIlrath,* 2008) noted that the judge was entitled to discount Static-99 due to its limitations, and evidence does not need to be admissible at a trial in order to be considered in a sentencing hearing.

Later on, in two New Hampshire cases (*State v. Ragan*, 2011; *State v. Hurley*, 2010) the State Supreme Court found Static-99R's nominal risk categories and three of four reference groups[3] for recidivism data inadmissible. Other aspects of the instrument survived the Daubert analysis. Finally, two cases (*State v. Ploof*, 2009; *United States v. Carta*, 2011) challenged the admissibility of selecting normative groups for recidivism estimates. In *Ploof,* the State Supreme Court in New Hampshire ruled that experts could report the Static-99R total score, but recidivism results were restricted to the routine reference group. They could not select other reference groups or estimate the individual's risk along the range provided by the routine and high/risk need norms. Additionally, experts must use wording clearly denoting the frequentist interpretation of the estimates (e.g., "a group of individuals with a similar score reoffended at X rate"). In *Carta*, the U.S. District Court expressed reservations around empirical validation of "…placing offenders in bins [sample type reference groups] in order to calculate more precise re-offense rates" (*United States v. Carta,* 2011, Footnotes), but ultimately admitted Static-99 (and 99R) while placing little dispositive weight in the recidivism estimate.

As with *Frye*, fewer admissibility challenges were observed between 2012 and 2020. Of 10 challenges, six focused on the instrument's reliability and accuracy, with two decisions in Wisconsin producing conflicting decisions. In *re Commitment of Perren* (2015), the Appellate Court affirmed the lower's court's decision to exclude Static-99R but admit Static-99, finding more precision and reliability for the older instrument because the manual for

[3] The 2012 Evaluator Workbook had 4 reference groups for estimated recidivism rates (Phenix et al., 2012). In the later update to the recidivism norms (Hanson, Thornton et al., 2016), this was reduced to two: routine/complete and preselected high risk/needs.

Static-99R was not available at the time. In contrast, in *State v. Wortman* (2017), the trial court excluded Static-99 (but admitted Static-99R) based on the notion that, "as data changes, our use of evidence-based tools will have to change as well" (*State v. Wortman*, Oral Ruling, May 4, 2017, p. 5). Three other challenges centered around normative samples and risk categories, as well as misuses or misinterpretations (most commonly, the continued use of Static-99 norms). In Massachusetts, the State Supreme Court decided that classification of defendants into risk categories (i.e., "moderately-likely to reoffend") lacked probative value and should not be admitted (*Commonwealth v. George*, 2017). In contrast, the State Supreme Court in Arizona determined that even an unscored administration (i.e., where a total score was never calculated or compared to norms) of Static-99R was fully admissible (*In the matter of Martin*, 2013).

Challenges in Canada

We identified two cases in Canada where the applicability of Static-99R was challenged for Indigenous individuals (not part of the Online Supplement). The main concern in these cases was the possibility of cross-cultural bias when applying a scale developed on predominantly White individuals to Indigenous individuals. In *R.v.* Awasis (2016), the Provincial Court of British Columbia ruled that Static-99R was admissible because it has been sufficiently validated with Indigenous people. *Ewert v. Canada* (2018) was a Supreme Court of Canada ruling pertaining to grievances filed as far back as 2007 (for an overview of this case, see Hart, 2016), with Mr. Ewert alleging risk tools applied by the federal prison system were biased against Indigenous individuals. In the final ruling, the Supreme Court sided in part with Mr. Ewert, in that the federal prison system had breached their statutory obligations by not validating the tool with Indigenous individuals. However, they did not preclude the use of Static-99R or other risk scales with Indigenous individuals, nor did they review the research on its applicability; rather, they set the expectation that cross-cultural validity research was needed to apply the scales. These court decisions mirror

increasing momentum to study risk assessment with Indigenous peoples in Canada (e.g., Babchishin, Blais, & Helmus, 2012; Lee et al., 2020; Olver, Sowden, et al., 2018).

Admissibility Considerations and Summary

The cases in Online Supplement C demonstrate the breadth and complexity of considerations for the admissibility of Static-99/R, while also highlighting the instrument's increased acceptance within the courts over time. Increased acceptance may reflect the increased research on the scale, but it may also demonstrate the kind of unquestioning acceptance reported by Neal et al. (2019). Attorneys or judges may be reluctant to exclude Static-99R because of its widespread use but such unscrutinized acceptance is problematic when the scale is misused or scored incorrectly. On the other hand, there are examples of Static-99R excluded or discounted unnecessarily. For example, in several cases, a defendant sought to introduce Static-99/R evidence but was denied the opportunity to do so (e.g., *McIlrath v. United States*, 2008; *New Jersey v. CB*, 2020), although Static-99/R evidence had been previously admissible in the same courts when presented by prosecutors. Furthermore, some admissibility decisions may reflect misunderstandings of the research by the court. For example, in another case (*In re Commitment of C.R.M.*, 2014), the New Jersey trial court judge (later affirmed by the Appellate Court) discounted Static-99R results because they were only moderately predictive and instead privileged information about the individual's denial of their offenses, which has no predictive power (Mann et al., 2010).

Today, admissibility challenges centered on general acceptance are unlikely to succeed. In contrast, challenges surrounding the different normative groups and risk communication metrics have had higher success rates, although updates in research may change the outcome of future cases. The regular updates of these scales have resulted in attorneys investing substantial efforts in learning about the statistical and empirical basis of these practices. This highlights a continued need for accessibility of research and for collaboration between attorneys, researchers, and clinicians.

In the U.S., challenges around the suitability of Static-99R to specific subpopulations have been few and far between. Only six U.S. cases challenged the use of Static-99/R in individuals who, by some characteristic (e.g., young age at index offense, cognitive disability, or mental health diagnosis) were atypical of the samples used to develop and validate Static-99R. Although atypical and less commonly researched, the scales have been valdated for use with developmentally delayed (Hanson, Sheahan, & VanZuylen, 2013) and mentally disordered individuals (Baudin et al., 2021; Hanson et al., 2007). Less is known about validity for individuals with specific types of neurocognitive abnormalities such as brain damage or dementia, although there is no a priori reason to believe Static-99R would function differently with those groups. Only one such challenge was successful in excluding Static-99 evidence (where the defendant was 15; *Re Civil commitment of JP,* 2001), which is consistent with the coding manual's instructions not to use the scale for such cases. We also identified no similar U.S. challenges based on a defendant's race or ethnicity.

How Should I Prepare for Cross-Examination?

This section is intended for both practitioners preparing for testimony and attorneys preparing their cases. This review offers multiple insights for cross-examination. Expert testimony on Static-99R results can be highly persuasive to decision-makers (Elwood, 2019), yet flawed testimony and misuse are rarely challenged in courts (DeMatteo et al., 2019). Although expert witnesses have ethical obligations for training and continuing education to avoid misuse and error, attorneys have an important role in identifying scoring errors and misuses. As such, we encourage attorneys to attend Static-99R trainings to better understand scoring, stay abreast of research, and devise strategies for questioning.

One common reason for user error appears to be a lack of formal training or updated training by a certified Static-99R trainer. If training experience is not readily apparent on the witness' curriculum vitae, questions about training should be asked during voir dire or when questioning becomes focused on Static-99R results. Such questions might include:

- Where did the witness receive training? Who trained them? Was the trainer a certified trainer? Importantly, not all individuals who offer Static-99R trainings are certified trainers; evaluators may falsely presume their trainer was certified.
- When did the witness last receive training? Did the witness receive updated training since the revised coding manual was released in 2016?

It should be a red flag if the witness has not completed training since November 2016. The 2016 coding manual (Phenix, Fernandez, et al., 2016) has notable changes and the authors strongly encouraged users to receive a minimum half-day booster training before applying the new coding manual. Static-99R developers also required trainers to attain re-certification using the 2016 manual. Consequently, a non-trivial number of trainers lost their certification status. Re-trainings can also help evaluators stay up to date on research advances, prevent drift, and refresh on tricky details. The new Guidelines for Static-99R training suggest re-training every 2-5 years, with anything past 10 years not recommended (SAARNA, 2021). It is important to note, however, that these Guidelines are new, and it may take some time for these recommendations to be widely disseminated and adopted. Questions about the version of the coding manual the witness used (i.e., 2003 or 2016, if any at all) or the Evaluator Workbook (e.g., 2012, 2016, or 2021) may reveal important information about potential coding errors or misuses.

Furthermore, although the developers have disavowed use of Static-99 after the publication of Static-99R, some evaluators continue to use it, and recent attempts to exclude expert witness testimony based on Static-99 results have largely been unsuccessful (e.g., *Re Detention of Powell*, 2016; *State v. Jones*, 2018; for an exception, see *State v. Wortman,* 2017). More commonly, effective cross examination can highlight salient weaknesses in continued use of the original Static-99, including diminished acceptance within the professional community (Kelley et al., 2020), substantial over-estimation of risk for individuals aged 60 and older (Helmus, Thornton, et al., 2012; Raymond et al., 2020),

lack of current norms, and clear instructions from the developers against continued use of Static-99 (Hanson, 2016; Thornton, 2016; see also Phenix, Fernandez, et al., 2016, p. 4).

As many courts have held that issues of scoring and interpretation of Static-99R go to the weight of the evidence, rather than its admissibility, most concerns should be addressed with skillful cross examination. Broadly, attorneys should ask about potential deviations from the established Static-99R protocol and current recommendations from the scale developers to establish the reliability of the expert's methods (e.g., Re *Detention of Hargett*, 2003; *U.S. v. McIlrath*, 2008). Furthermore, the onus of establishing the reliability and validity of these deviations would lie with the evaluator (Joint Committee, 2014).

Although the use of Static-99R alone for screening decisions (e.g., assignment of cases to treatment intensity based on risk level) may be appropriate, forensic evaluations such as considerations for SVP commitment should use Static-99R as part of a comprehensive evaluation, because each individual's true risk varies over time and in context (ATSA, 2014). Comprehensive evaluations should consider dynamic risk and treatment change, which have been demonstrated to contribute to the prediction of sexual re-offense risk beyond Static-99R (Brankley et al., 2021; Olver, Mundt, et al., 2018; Thornton & Knight, 2015). Preliminary research also suggests that protective factors have promise as additional considerations (Nolan, 2021; Willis & Thornton, 2021). Attorneys should be wary of forensic reports that lack a measure of dynamic risk unless the expert provides a case-specific and cogent rationale for its exclusion. Although Static-99R should not be the sole information source for high stakes assessments, attorneys should also be mindful of some experts' tendency to artificially inflate risk estimates on Static-99R via clinical override. As discussed earlier, overrides consistently degrade predictive accuracy (for review, see Helmus, 2021) and merit critical questioning on the research basis and reliability of the expert's decision.

Recidivism Estimates

Absolute recidivism estimates are the least reliable risk communication metric across samples and settings (Helmus, 2018) and when experts report them, there are a number of

things for attorneys (and evaluators) to be mindful of (for a review paper on this topic, see Helmus, 2021). Firstly, in high stakes assessments such as SVP proceedings, absolute recidivism estimates should ideally be reported with a qualifying statement that the rates associated with each total score have been found to vary across diverse samples and settings (Hanson, Thornton, et al., 2016). Secondly, the decision process for choosing between the routine/complete versus preselected high risk/need norms can be fraught (Abbott, 2013). The current recommendations are to use the routine/complete norms as the default and only use the preselected high risk/need norms when there is a strong, case-specific justification based on the density of criminogenic needs (Hanson et al., 2016). Evaluators tend to make this choice in one of three ways: (1) only use the Routine/Complete group and largely ignore the high risk/need norms, (2) match their cases to the reference group by either using historical selection factors like administrative screening decisions or clinical judgment to infer the density of criminogenic needs, and (3) use of an actuarial measure of criminogenic need to guide the selection of the reference group (Kelley et al., 2020).

Inconsistency in reference group selection raises concerns about the reliability of the method as well as potential adversarial allegiance (Chevalier et al., 2014). Although the first method is appropriate in many samples and settings, it can ignore important variation in others (Hanson, Thornton, et al., 2016). The second method introduces a substantial amount of clinical judgment into an otherwise structured actuarial tool, is discouraged by the development team (Hanson, Thornton, et al., 2016), and runs counter to ATSA's (2014) standards for using structured assessments of dynamic risk (as opposed to unstructured). The third method has some early empirical support (Hanson & Thornton, 2012), albeit unpublished, and is consistent with the scale developers' recommendations (Hanson, Thornton, et al., 2016).

Using an actuarial measure of dynamic risk is not only the most defensible method but it can also allow evaluators to sidestep difficult questions about reference group

selection entirely. Specifically, some dynamic risk tools (e.g., STABLE-2007 & VRS-SO) offer recidivism estimates for dynamic risk combined with Static-99R (e.g., Brankley et al., 2017; Olver, Kelley, et al., 2020). This is a more precise assessment of what the normative groups are roughly approximating, while also capturing dynamic risk that may lower risk measured from static factors alone (given that there are no preselected low risk normative groups available). Decisions about dynamic instrument selection are beyond the scope of this paper, but consideration should be given to the pros and cons of the tools with regard to the available research, the nature of the evaluation, the population being assessed, and existing caselaw in their jurisdiction (where applicable).

Also important to consider is the degree to which the estimated recidivism rates associated with a Static-99R score change over time when an individual has lived in the community since release from the index sexual offense (Hanson et al., 2018; Thornton et al., 2021). Static-99R total scores do not change post-release (e.g., as the person ages), but the longer they remain in the community without further re-offenses, the more their risk decreases, and mechanical methods (i.e., Excel-based calculator allowing input of data for relevant variables important in modeling risk) have been developed to incorporate this into recidivism estimates. How long an individual needs to be offense-free in the community before reducing their estimated risk to below 2% (i.e., the statistical desistance threshold; Kahn et al., 2017) will vary based on the Static-99R score and the initial base rate. However, most individuals will have desisted from sexual offending by 10 to 15 years offense-free (Thornton et al., 2021).

Therefore, questions about whether the evaluator considered an appropriate reduction of risk based on time free in the community will be important for any cases involving community supervision violations as well as sexual offender registry cases. This is applicable not just for individuals supervised in the community for a sexual offense, but also for those serving a sentence for a non-sexual offense scored on Static-99R because of a past sexual offense (i.e., Static-99R is tied to their index sexual offense, which is not

necessarily the same as their current sentence). This may have several implications, such as appropriately scoring the age item and considering time free effects linked to the appropriate release date.

Discussion

What Future Research is Needed?

In the 23 years since Static-99 was first developed, there has been a proliferation of research and resources to refine and improve the scale. It is unlikely that this process will ever be 'done' unless the scale is fully replaced by an alternate process. Risk scales *should* be continually updated (Dawes et al., 1989); this is evidence-based practice. The scale has considerable research support (more than most other sexual recidivism risk tools) and general acceptance by the scientific and legal community. Nonetheless, there are still myriad opportunities for improvement of the scale and considerable unknowns.

Overall meta-analytic averages have found moderate predictive accuracy for the scale. However, subgroup analyses in field settings suggest that among properly trained evaluators who are committed to the process, accuracy could be substantially higher (e.g., AUCs of .80). In the current meta-analysis, simply mentioning the coding manual was associated with higher accuracy. Although there are Training Guidelines for Static-99R (SAARNA, 2021), these were mostly developed based on consultation among experienced trainers and represent educated hypotheses. Future research could explore training effectiveness under different formats (e.g., length of training, number of participants, single session vs. multiple sessions, types of practice cases and competency exams, training modalities, variations in content, prerequisite trainings, trainer characteristics) as well as different strategies to cultivate buy-in and motivation from evaluators and agencies.

As demonstrated by the improvements in calibration with the revised age weights (Helmus, Thornton, et al., 2012), it is possible that additional research will lead to further improvements in discrimination or calibration for the scale. This may come from exploring age in more detail (e.g., comparing older cases with recent versus historical sexual

offenses), considering time in custody, or examining other risk factors or characteristics. Research may also lead to expanded or improved application of the scale to more modern sex offenses, such as those committed over the internet.

Although larger and more recent datasets allow examination of the accuracy of Static-99R for more refined subgroups, it does not necessarily follow that Static-99R should be presumed invalid for new offenses or for subgroups not specifically examined. The Joint Committee (2014) Standards note that not every possible subgroup or unique case requires separate validation research. As examples, we often hear attorneys or psychologists asking if the Static-99R validation samples include individuals "like [insert this characteristic]." With enough specificity, each individual becomes entirely unique. But the question is, do those types of differences matter insofar as the purpose of the assessment? The Joint Committee Standards emphasize that test development and validation must consider "relevant subgroups" (see Chapter 3). It is often safe to presume broad generalization of risk scales, unless there is credible theory or research to suggest otherwise. However, where there is evidence or theory suggesting that a scale may predict differently for a subgroup (and sample sizes are sufficient), the Joint Committee Standards are clear that "test developers and/or users are responsible for evaluating the possibility of differential prediction for relevant subgroups" (Standard 3.7, p. 66). Consequently, as research advances, we may identify the need for further subgroup validation. Currently, the coding manual lists some subgroups that are sufficiently different that the scale should not be applied to them (e.g., women, individuals who committed their latest sexual offense at age 16 or younger, individuals whose sole sexual offense is possession of child pornography). This list may be expanded with future research; alternatively, subsequent research may find adjustments to the scale to increase its range of appropriate applications.

Cross-cultural validity of Static-99R represents one area where there is sufficient theory and/or research to warrant further empirical investigation. There is ample evidence of over-representation of certain groups in the criminal legal system, particularly Black and

Indigenous people (Nellis, 2016; Public Safety Canada, 2020), related to systemic racism and differential policing and sentencing practices (Ontario Human Rights Commission, 2020). Racism and discrimination have resulted in some non-White groups experiencing higher risk factors for crime and recidivism (Alexander, 2010; Truth and Reconciliation Commission, 2015). Consequently, it is important to examine the possibility of cross-cultural biases in the predictive accuracy of risk assessment scales (this is emphasized by the Supreme Court of Canada decision in *Ewert v. Canada*). Preliminary data suggest Static-99R can be used with Indigenous people in North America (Babchishin, Blais, & Helmus, 2012; Lee et al., 2020; Myer, 2019), as well as Black and Latino people in the United States (Boccaccini et al., 2017; Lee & Hanson, 2017). It is also important to examine the applicability of Static-99R in other countries, particularly in Asian, African, and Central and South American countries, where there has been very little research to date. Considerations in international applications of the scale were discussed earlier. In a recent exploratory study from Singapore with a very small sample size (n = 134), Static-99R approached statistical significance with an AUC value (AUC = .70) comparable to the meta-analytic average. More studies like this are needed.

Additionally, calibration of the recidivism estimates should be continually re-evaluated. Rates of sexual offenses have fluctuated over time (mostly declining; Mishra & Lalumière, 2009), and recidivism follow-up studies take time to accumulate. Factoring in study planning, analyses, and publication, any study with 10-year follow-up data typically examines individuals released at least 15 years earlier. Recidivism datasets should regularly be updated and evaluated, particularly after big cultural shifts that may impact rates of sexual offending or of reporting of sexual offenses (e.g., the #MeToo movement in 2017).

Risk Communication, Knowledge Translation, and Implementation Issues

The accuracy of Static-99R is irrelevant if evaluators, attorneys, and decision-makers do not understand and use the information. Research on how to effectively communicate the results of Static-99R (or other risk scales) has lagged behind research on predictive

accuracy (for review of risk communication research, see Hilton et al., 2015). Indeed, the court cases we reviewed were rife with examples of the scale and its research misunderstood or misrepresented by attorneys, judges, and even expert witnesses. Clearly, more work is needed to improve risk communication and knowledge translation.

Recent vignette-based studies examined how the communication of actuarial risk scale results can influence risk perception, which should ideally influence decision-making. Key findings from this line of research are that risk ratios are poorly understood (Varela et al., 2014), different risk communication metrics can make it harder or easier to differentiate between low and high risk cases (Krauss et al., 2018; Varela et al., 2014), some risk communication metrics contribute to higher or lower perceptions of risk (e.g., risk levels are associated with higher perceived risk, whereas recidivism estimates and multi-metric communication are associated with lower perceived risk; Helmus et al., 2018), and graphs generally seem to improve understanding of risk information (Hilton & Helmus, 2021).

Evidence-based risk scales continue to evolve and are rife with possibilities for misuse. This highlights the need for conscientiousness and careful risk communication, appropriate training, and staying up to date on the latest research and recommendations. This is no easy feat and gaps in knowledge translation contribute to subpar practice. Historically, the development team for Static-99R has contributed to the confusion by failing to regularly update the static99.org website. Part of this could be related to the non-proprietary nature of the scale. Although there are benefits to users for the scale being available free of charge, the absence of an income stream, coupled with the development team having diverse jobs which generally do not include time devoted to the scales, creates obstacles in providing timely research and resources to assist with implementation worldwide. Recently, SAARNA was launched as a non-profit designed to provide trainings and resources for users of actuarial risk scales such as Static-99R, Static-2002R, STABLE/ACUTE-2007, and Risk Matrix 2000 (see saarna.org). The scales themselves are

still free to use, but fees from training and subscription resources may help fund better resources and knowledge dissemination practices in the future.

Strengths and Limitations

In this review paper we attempted to provide a thorough, transparent, and balanced coverage of the relevant topics, with an international and multidisciplinary team of authors, some involved in Static-99R research and implementation (e.g., certified trainers) and some fully independent from the scale. Having some authors affiliated with the scales brings both strengths (e.g., deep knowledge of the available material and extensive experience with the scale) and limitations (e.g., potential for affiliation biases). Having co-authors familiar with but independent from the scale was intended to minimize biases. Nonetheless, biases (for or against the scale) are always a possibility.

The meta-analysis had several key strengths, including all studies being coded and consensified by two raters, a narrow scope (reducing the possibility of omitted research or important moderators), high statistical power (due to the large number of studies), and a fairly comprehensive search (e.g., contacting professionals in the field) that resulted in considerable unpublished material, minimizing possible publication biases. One notable limitation, however, is that we did not track the number of unique hits across the various search strategies and the reason for their exclusion (though Online Supplement A does list some studies that were excluded and why; e.g., a study that examined institutional recidivism as opposed to new sexual offending in the community). Such methods are recommended reporting practices (e.g., see the PRISMA guidelines; Page et al., 2021). Although they do not speak to the comprehensiveness of a search, they do index the transparency and replicability of the search.

Similarly, our case law review had similar strengths and limitations. We attempted to make the search as thorough as possible (e.g., multiple databases searched, reached out to colleagues), but it is not necessarily transparent or replicable. Additionally, many of the databases index primarily American cases and higher level court decisions. We do not

expect that Online Supplement C is an exhaustive list of admissibility challenges to the scale, but we do believe it is a helpful resource.

Conclusions

Given that human behavior is not perfectly predictable, and all risk scales have limitations and room for improvement, we believe Static-99R can be a useful and informative tool in legal decisions, with some important caveats. Where applicable, in addition to considering the evidence supporting the validity or utility of a risk assessment scale for a particular purpose, it is also helpful to consider the harms of *not* using the scale (Olver, Sowden, et al., 2018). No risk tool is without limitations, but where a decision must be made, we should consider the benefits and harms both of using the scale and of not using the scale. For example, in sexual recidivism risk assessment, unstructured clinical judgment is significantly worse than structured risk scales (Hanson & Morton-Bourgon, 2009) and humans tend to overestimate risk (Kahneman, 2011). For example, roughly half of jurors have opined that even a 1% risk of sexual recidivism would be grounds for SVP commitment (Knighton et al., 2014), despite this level of risk being consistent with individuals with a criminal record for nonsexual offending and no known sexual offenses (Kahn et al., 2017). Consequently, even a risk scale with moderate predictive accuracy (when we would prefer high accuracy) is often better than the alternative of not using a risk scale. Not using a scale will often result in higher and less accurate assessments of risk.

When there are numerous possible tools available, we must consider the advantages of using the scale compared to other commonly available options. In some cases, a scale such as Static-99R may have known limitations or error rates that have not been examined for other risk scales, so it may be preferable to use a scale where the limitations are at least known as opposed to one where they are unknown. Lastly, even if the scale is better than the alternatives, it does not mean that improvements to the scale are not needed.

Other possible alternatives include using SPJ scales or no scale at all but referring to local base rates of recidivism. Some SPJ scales have comparable accuracy (Hanson &

Morton-Bourgon, 2009) although they do not have the quantitative risk communication metrics available in actuarial scales (percentiles, risk ratios, recidivism estimates) and the increased room for professional discretion can also increase room for bias. Referring to local base rates of recidivism is not advisable because it does not incorporate any individualized assessment of risk, essentially treating all individuals as a homogeneous group (contrary to the principles of effective intervention and management with justice-involved individuals; Bonta & Andrews, 2017). Any risk scale that predicts significantly better than chance is going to be more informative than providing the overall base rate of recidivism.

Static-99R is a defensible scale to use, but it is not the be-all-end-all; it has limitations and room for improvement. It is routinely ruled legally admissible (see above) and performs at least as well as other scales (see Hanson & Morton-Bourgon, 2009). Generally, no risk assessment tool has consistently found to be empirically superior to other structured risk tools (with the exception that RRASOR is statistically inferior; Hanson & Morton-Bourgon, 2009). Overall, Static-99R is one of the best options we have available (but not the only option). Clear advantages in using Static-99R lie in it being the most widely used risk tool for sexual recidivism, and with the most extensive research base.

In a review of psychological assessment tools used in legal contexts, Neal et al. (2019) raised concerns that many instruments were not generally accepted in the field and/or did not have favorable psychometric properties. Moreover, admissibility of these instruments was not often scrutinized. Consistent with the findings of a large number of cases that have scrutinized Static-99R, we believe that Static-99R meets Daubert criteria for admissibility in court, when used appropriately. Its limitations should speak to the weight it is given in legal decisions (and this may vary depending on the jurisdiction, the referral question, and unique circumstances of the case), not its admissibility. We address this in terms of the specific prongs of the Daubert criteria:

1) *Whether the technique in question can be and has been tested.* Yes, risk scales can be validated and Static-99R has been.

2) *Whether it has been subject to peer review and publication.* Yes, the scale has been published in peer reviewed journal articles (Helmus, Thornton, et al., 2012), as well as each of the risk communication metrics (Hanson, Babchishin et al., 2016; Hanson et al., 2013; Hanson, Lloyd et al., 2012; Hanson, Thornton et al., 2016; Lee & Hanson, 2021) and most of the validation studies (see Online Supplement A).
3) *Its known or potential error rate.* Consistent with a previous meta-analysis (Helmus, Hanson et al., 2012), this meta-analysis found that Static-99R demonstrates moderate predictive accuracy in discriminating those who reoffend from those who do not (AUC of .68 to .69). Its accuracy is significantly better than chance and significantly better than unstructured clinical judgments of risk (Hanson & Morton-Bourgon, 2009). The SEM for Static-99R is roughly 1 point. The 5-year probability estimates of new charges and convictions for sexual recidivism, however, may overestimate *detected* 5-year recidivism rates in many samples; consequently these estimates should be interpreted with caution. This is likely true for all actuarial risk assessment scales (Helmus, 2018) and also does not consider other factors that could contribute to recidivism estimates either overestimating or underestimating the outcome of interest, which should be weighed together (Helmus, 2021).
4) *The existence and maintenance of standards controlling its operation.* Static-99R has a more detailed coding manual than any other scale we are familiar with (Phenix, Fernandez et al., 2016) and training standards (SAARNA, 2021). Remaining scoring questions can be submitted to the website.
5) *Whether it has attracted widespread acceptance within a relevant scientific community.* Yes. As noted, every survey we are aware of finds that Static-99R is the most frequently used risk scale worldwide to assess sexual recidivism risk.

STATIC-99R STRENGTHS AND LIMITATIONS

Static-99R provides an important baseline risk assessment based on static risk factors. It may be particularly useful in triaging large numbers of individuals for resource allocation. It should not be mistaken for a comprehensive risk assessment; it is but one piece of a larger puzzle. Highly consequential referral questions, such as SVP commitment qualification, should prompt a comprehensive assessment including a structured assessment of dynamic risk factors (ATSA, 2014) and consider other factors such as changes over time, environmental influences, protective factors, and unique case features (e.g., health conditions that may impact mobility). Such a decision should never be based on Static-99R alone (or any single risk scale).

Although Static-99R should not be the sole piece of information considered for legal decisions such as civil commitment, its results should not be overridden either. Although imperfect, it does have a strong research base, whereas studies examining overrides of sexual recidivism risk tools have consistently found that overrides degrade predictive accuracy. Practitioners are therefore in a predicament: they know Static-99R is not the only relevant information, but they also know that their attempts to adjust the results of it will make things worse. Where does this leave the conscientious evaluator? Our advice is that a comprehensive risk assessment includes humility. Evaluators should consider both static and dynamic risk scales, other sources of information, stick to the research (including an understanding of its limitations), and use their professional judgment sparingly. Results for adversarial allegiance effects, for example, suggest that the more room there is for subjectivity in risk scales, the more room for bias (e.g., Murrie et al., 2013).

A major research limitation for Static-99R is that absolute recidivism estimates are difficult to estimate with precision. This should be acknowledged when recidivism estimates are reported. Although they may be the risk communication metric of most importance in some legal decisions, such as SVP cases, their limits should not be glossed over. Importantly, however, variability in recidivism estimates is unlikely to be specific to Static-99R. Given the myriad factors that impact detected recidivism rates beyond an individual's

risk, this is likely a limitation of all actuarial risk scales (Helmus, 2018). Recidivism estimates from scales combining both static and dynamic risk factors will likely reduce this variability, but ultimately the research suggests that precise probabilities of recidivism associated with actuarial risk scores are not easy to derive. They are, however, preferable to unstructured judgment. Overall, humility and deference to research should be guiding principles (for guidance on commenting on recidivism probabilities, see Helmus, 2021).

As discussed in the introduction, it is also important to remember that Static-99R was not developed specifically for a legal context or any particular legal decision. It is designed to rank order individuals in their likelihood of reoffending, primarily to inform resource allocation. It does fairly well at that. Any legal question that is more specific involves some gap between the legal question and what the scale is intended to measure. The size and type of gap depend on the legal question. Evaluators should not treat Static-99R as if it provides a direct answer to the legal question. Those gaps need to be acknowledged, and some professional judgment is required to bridge that gap in the overall opinion. This limitation applies to all risk scales. As noted by Hanson (2009), "Scoring an actuarial risk tool is not a risk assessment. Evaluators will always need to make a separate judgment as to whether the risk scale score fairly represents the risk posed by the individual being assessed" (p. 174).

References

Abbott, B. R. (2013). The utility of assessing "external risk factors" when selecting Static-99R reference groups. *Open Access Journal of Forensic Psychology*, *5*, 89-118.

Alexander, M. (2010). *The new Jim Crow: Mass incarceration in the age of colorblindness*. New Press.

Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with Level of Service general risk and need. *Criminal Justice and Behavior, 38*(5), 413-432. https://doi.org/10.1177/0093854811401990

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*(1), 84-94. https://doi.org/10.1207/s15327752jpa8701_07

Association for the Treatment of Sexual Abusers (2014). *ATSA practice guidelines for assessment, treatment interventions, and management strategies for male adult sexual abusers.* Professional Issues Committee.

Association for the Treatment of Sexual Abusers and Sex Offender Civil Commitment Programs Network (2015). *Civil commitment of sexual offenders: Introduction and overview.* https://www.atsa.com/sites/default/files/%5bCivil%20Commitment%5d%20Overview.pdf

Austin, J., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the Static-99/RRASOR sex offender risk assessment instruments.* National Institute of Corrections website, http://nicic.gov/Library/022957

Babchishin, K. M., Blais, J., & Helmus, L. (2012). Do static risk factors predict differently for

Aboriginal sex offenders? A multi-site comparison using the original and revised Static-99 and Static-2002 scales. *Canadian Journal of Criminology and Criminal Justice*, *54*(1), 1-43. http://dx.doi.org/10.3138/cjccj.2010.E.40

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012). Even highly correlated measures can add incrementally to actuarial risk prediction. *Assessment*, *19*(4), 442-461. https://doi.org/10.1177/1073191112458312

Baudin, C., Nilsson, T., Sturup, J., Wallinius, M., & Andiné, P. (2021). A Static-99R validation study on individuals with mental disorders: 5 to 20 years of fixed follow-up after sexual offenses. *Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.625996

Blair, P. R., Marcus, D. K., Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15*(4), 346-360. https://doi.org/10.1111/j.1468-2850.2008.00147.x

Blais, J., & Forth, A. E. (2014). Prosecution-retained versus court-appointed experts: Comparing and contrasting risk assessment reports in preventative detention hearings. *Law and Human Behavior*, *38*(6), 531-543. https://doi.org/10.1037/lhb0000082

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A. K., & Jeglic, E. L. (2012). Implications of Static-99 field reliability findings for score use and reporting. *Criminal Justice and Behavior*, *39*(1), 42-58. http://dx.doi.org/10.1177/0093854811427131

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. http://dx.doi.org/10.1037/pas0000377

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct* (6th ed.) Routledge.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Wiley & Sons, Ltd.

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. https://doi.org/10.3138/cjccj.2016-0024

Brankley, A. E., Babchishin, K. M., & Hanson, R. K. (2021). STABLE-2007 demonstrates predictive and incremental validity in assessing risk-relevant propensities for sexual offending: A meta-analysis. *Sexual Abuse, 33*(1), 34-62. https://doi.org/10.1177/1079063219871572

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

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. https://doi.org/10.1177/1073191114568114

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 https://doi.org/10.1177/1079063217691965

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*(3), 209-218. https://doi.org/10.1037/lhb0000114

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, T. H., Lowenkamp, C. T., Bechtel, K., & Flores, A. W. (2020). Risk Assessment Overrides: Shuffling the Risk Deck Without Any Improvements in Prediction. *Criminal Justice and Behavior*, *47*(12), 1609–1629. https://doi.org/10.1177/0093854820953449

Commonwealth v. George, 477 Mass. 331, 76 N.E.3d 217, 2017 Mass. LEXIS 396 (Mass. S.J.C., Jun. 21, 2017)

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 113 S. Ct. 2786, 125 L. Ed. 2d 469 (1993).

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674. https://doi.org/10.1126/science.2648573

DeMatteo, D., Fishel, S., & Tansey, A. (2019). Expert evidence: The (unfulfilled) promise of Daubert. *Psychological Science in the Public Interest*, *20*(3), 129-134. https://doi.org/10.1177/1529100619894336

Doyle, D. J., Ogloff, J. R. P., & Thomas, S. D. M. (2010). An analysis of dangerous sexual offender assessment reports: Recommendations for best practice. *Psychiatry, Psychology and Law*, *18*(4), 537-556. https://doi.org/10.1080/13218719.2010.499159

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

Eher, R., Rettenberger, M., Etzler, S., Eberhaut, S. & Mokros, A. (2019). Eine gemeinsame Sprache für die Risikokommunikation bei Sexualstraftätern – Trenn- und Normwerte für das neue Fünf-Kategorienmodell des Static-99 [A common language for communicating risk in sexual offenders – the 5-category model of the Static-99]. *Recht & Psychiatrie, 37*(2), 91-99.

Elwood, R. W. (2019). Agreement between courts and SVP evaluators in the State of Wisconsin. *Criminal Justice and Behavior*, *46*(6), 853-865. https://doi.org/10.1177/0093854819839746

Ewert v. Canada, 2018 SCC 30 2 S.C.R. 165 Lexbox (SCC 2018)

Fernandez, Y. (2021, Winter). The history of Static-99 and the value of research-practitioner collaboration. *The Forum, 40*(1).

Fernandez, Y., Harris, A. J. R., Hanson, R. K., & Sparks, J. (2014). *STABLE-2007 coding manual – revised 2014* [Unpublished report]. 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).

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley.

Funk, C. (last updated Aug. 9, 2021). *Daubert versus Frye: A national look at expert evidentiary standards.* Expert Institute. Retrieved from https://www.expertinstitute.com/resources/insights/daubert-versus-frye-a-national-look-at-expert-evidentiary-standards/

Frye v. United States, 293 F. 1013, 34 A.L.R. 145 (D.C.Cir. 1923)

Gonçalves, L. C., Gerth, J., Rossegger, A., Noll, T., & Endrass, J. (2020). Predictive validity of the Static-99 and Static-99R in Switzerland. *Sexual Abuse*, *32*(2), 203-219. http://dx.doi.org/10.1177/1079063218821117

Guarnera, L. A., 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.

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology/Psychologie Canadienne, 50*(3), 172-182. https://doi.org/10.1037/a0015726

Hanson, R. K. (2016, January 15). *An affidavit*. http://www.static99.org/pdfdocs/Legal-Thorton-Hanson_Affidavits(2016).pdf

Hanson, R. K. (2021). *Prediction statistics for psychological assessment*. American Psychological Association.

Hanson, R. K., Babchishin, K. M., Helmus, L., & Thornton, D. (2013). Quantifying the relative risk of sex offenders: Risk ratios for Static-99R. *Sexual Abuse*, *25*(5), 482-515. https://doi.org/10.1177/1079063212469060

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. https://doi.org/10.1037/pas0000371

Hanson, R. K., Bourgon, G., McGrath, R., Kroner, D., D'Amora, D. A., Thomas, S. S., & Tavarez, L. P. (2017). *A five-level risk and needs system: Maximizing assessment results in corrections through the development of a common language.* The Council of State Governments Justice Center.

Hanson, R. K., Harris, A. J. R., Letourneau, E., Helmus, L. M., & Thornton, D. (2018). Reductions in risk based on time offense-free in the community: Once a sexual offender, not always a sexual offender. *Psychology, Public Policy, and Law, 24*(1), 48-63. http://dx.doi.org/10.1037/law0000135

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). 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. https://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, 11*(1), 9-23. https://doi.org/10.1080/14999013.2012.667511

Hanson, R.K., Lunetta, A., Phenix, A., Neely, J., & Epperson, D. (2014). The field validity of Static-99R Sex Offender Risk Assessment Tool in California. *Journal of Threat Assessment and Management, 1*(2), 102-117. https://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. https://doi.org/10.1037/a0014421

Hanson, R. K., Sheahan, C. L., & VanZuylen, H. (2013). Static-99 and RRASOR predict recidivism among developmentally delayed sexual offenders: A cumulative meta-analysis. *Sexual Offender Treatment, 8*(1), 1-14.

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. https://doi.org/10.1023/a:1005482921333

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of the Static-2002* (User Report 2003-01). Solicitor General Canada. Available from www.static99.org

Hanson, R. K., & Thornton, D. (2012, October). *Preselection effects can explain variability in sexual recidivism base rates in Static-99R and Static-2002R validation studies*. Presentation at the 31st Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Denver, CO.

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*(3), 218-252. https://doi.org/10.1177/1079063215574710

Harris, A. J. R., & Hanson, R. K. (2010). Clinical, actuarial, and dynamic risk assessment of

sexual offenders: Why do things keep changing? *Journal of Sexual Aggression, 16*(3), 296-310. https://doi.org/10.1080/13552600.2010.494772

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

Harris, G. T., Lowenkamp, C. T., & Hilton, N. Z. (2015). Evidence for risk estimate precision: Implications for individual risk communication. *Behavioral Sciences and the Law*, *33*(1), 111-127. https://doi.org/10.1002/bsl.2158

Hart, S. D. (2016). Culture and violence risk assessment: The case of *Ewert v. Canada. Journal of Threat Assessment and Management, 3*(2), 76-96. https://doi.org/10.1037/tam0000068

Hawes, S. W., Boccaccini, M. T., & Murrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the psychopathy Checklist—Revised. *Psychological Assessment, 25*(1), 233–243. https://doi.org/10.1037/a0030391

Heilbrun, K., Newsham, R., & Pietruszka, V. (2016). Risk communication: An international update. In Singh, J. P., Bjørkly, S., & Fazel, S. (Eds.), *International perspectives on violence risk assessment.* Oxford University Press.

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* [Master's thesis, Carleton University]. Carleton University Research Virtual Environment.

Helmus, L. M. (2018). Sex offender risk assessment: Where are we and where are we going*? Current Psychiatry Reports, 20*(6)*,* 46. https://doi.org/10.1007/s11920-018-0909-8

Helmus, L. M. (2021). Estimating the probability of sexual recidivism among men charged or convicted of sexual offences: Evidence-based guidance for applied evaluators. *Sexual Offending: Theory, Research, and Prevention, 16*, 1-24. https://doi.org/10.5964/sotrap.4283

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

Helmus, L. M., Cording, J., Murrie, D., & Hilton, N. Z. (2018, October). *Same score, different message? A replication/extension of Varela et al. (2014)* [Paper presentation]. Association for the Treatment of Sexual Abusers 37th Annual Research and Treatment Conference, Vancouver, BC, Canada.

Helmus, L. M., Hanson, R. K., Murrie, D. C., & Zabarauckas, C. L. (2021). 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. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0001010

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. https://doi.org/10.1177/0093854812443648

Helmus, L. M., Lee, S. C., Phenix, A., Hanson, R. K., & Thornton, D. (2021). *Static-99R & Static-2002R evaluators' workbook*. Society for the Advancement of Actuarial Risk and Needs Assessment (SAARNA). Available at www.saarna.org

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, 24*(1), 64-101. https://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. https://doi.org/10.1136/bmj.327.7414.557

Hilbert, J. (2018). The disappointing history of science in the courtroom: Frye, Daubert, and the ongoing crisis of junk science in criminal trials. *Oklahoma Law Review, 71*, 759-821.

Hilton, N. Z., & Helmus, L. M. (2021). Using graphs in sexual violence risk communication: Benefits may depend on the risk metric. *Sexual Abuse, 33*(6), 698-724. https://doi.org/10.1177/1079063220951191

Hilton, N. Z., Scurich, N., & Helmus, L. M. (2015). Communicating the risk of violent and offending behavior: Review and introduction to special issue. *Behavioral Sciences and the Law, 33*(1)*,* 1-18. https://doi.org/10.1002/bsl.2160

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

In re Civil Commitment of C.R.M., 2014 N.J. Super. Unpub. LEXIS 2956 (N.J. Sup. Ct. App. Div., Dec. 24, 2014).

in re Commitment of Perren, 2015 Wis. App. Ct. Unpub. 2013AP1195–NM, WL 13119689 (Wis. Ct. App., Feb. 18, 2015).

In re Det. of Johnson, 2002 Iowa App. LEXIS 1065, 2002 WL 31309172 (Iowa App. Ct., Oct. 16, 2002)

In re J.P., 339 N.J. Super. 443, 772 A.2d 54, 2001 N.J. Super. LEXIS 176 (N.J. Sup. Ct. App. Div., Apr. 24, 2001).

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. *Sexual Abuse, 19*(4)*,* 425-448. https://doi.org/10.1007/s11194-007-9062-3

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

Kahn, R. E., Ambroziak, G., Hanson, R. K., & Thornton, D. (2017). Release from the sex offender label. *Archives of Sexual Behavior, 46*(4), 861–864. https://doi.org/10.1007/s10508-017-0972-y

Kahneman, D. (2011). *Thinking fast and slow*. MacMillan.

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. https://doi.org/10.1177/1079063218800474

Knighton, J. C., Murrie, D. C., Boccaccini, M. T., & Turner, D. B. (2014). How likely is "likely to reoffend" in sex offender civil commitment trials? *Law and Human Behavior, 38*(3), 293–304. https://doi.org/10.1037/lhb0000079

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine,* 15(2), 155-163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk. *Archives of General Psychiatry, 54*(4), 337-343. https://doi.org/10.1001/archpsyc.1997.01830160065009

Krauss, D. A., Cook, G. I., & Klapatch, L. (2018). Risk assessment communication difficulties: An empirical examination of the effects of categorical versus probabilistic risk communication in sexually violent predator decisions. *Behavioral Sciences & the Law*, *36*(5), 532-553. https://doi.org/10.1002/bsl.2379

Landis, J.R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310.

Lee, S. C., & Hanson, R. K. (2017). Similar predictive accuracy of the Static-99R risk tool for White, Black, and Hispanic sex offenders in California. *Criminal Justice and Behavior, 44*(9), 1125-1140. https://doi.org/10.1177/0093854817711477

Lee, S. C., & Hanson, R. K. (2021). Updated 5-Year and new 10-Year sexual recidivism rate norms for Static-99R with routine/complete samples. *Law and Human Behavior*, *45*(1), 24-38. http://dx.doi.org/10.1037/lhb0000436

Lee, S. C., Hanson, R. K., & Blais, J. (2020). Predictive accuracy of the Static-99R and Static-2002R risk tools for identifying Indigenous and White individuals at high risk for sexual recidivism in Canada. *Canadian Psychology, 61*(1), 42-57. https://doi.org/10.1037/cap0000182

Leguízamo, A., Lee, S. C., Jeglic, E. L., & Calkins, C. (2017). Utility of the Static-99 and Static-99R with Latino sex offenders. *Sexual Abuse*, *29*(8), 765-785. http://dx.doi.org/10.1177/1079063215618377

Lehmann, R. J. B., Hanson, R. K., Babchishin, K. M., Gallasch-Nemitz, F., Biedermann, J., & Dahle, K.-P. (2013). Interpreting multiple risk scales for sex offenders: Evidence for averaging. *Psychological Assessment, 25*(3), 1019–1024. https://doi.org/10.1037/a0033098

Looman, J., Morphett, N. A. C., & Abracen, J. (2013). Does consideration of psychopathy and sexual deviance add to the predictive validity of the Static-99R? *International Journal of Offender Therapy and Comparative Criminology, 57*(8), 939-965. https://doi.org/10.1177/0306624X12444839

Lyon, D. R., & Welsh, A. (2017). *The psychology of criminal and violent behaviour*. Oxford University Press.

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: Journal of Research and Treatment*, *22*(2), 191–217. https://doi.org/10.1177/1079063210366039

Marshall, E., Miller, H. A., Cortoni, F., & Helmus, L. M. (2021). The Static-99R is not valid for women: Predictive validity in 739 females who have sexually offended. *Sexual Abuse, 33*(6), 631–653. https://doi.org/10.1177/1079063220940303

Mattek, R., & Hanson, R. K. (2018). Committed as a Violent Sexual Predator in his 10th decade: A case study. *Archives of Sexual Behavior,* 47(2), 543–550. https://doi.org/10.1007/s10508-017-1041-2

Matter of State of New York v. Rosado, 25 Misc. 3d 380, 889 N.Y.S.2d 369, 2009 N.Y. Misc. LEXIS 1741, 2009 NY Slip Op 29290, 242 N.Y.L.J. 8 (N.Y. Sup. Ct. Bronx, Jun. 29, 2009).

McCallum, K. E., Boccaccini, M. T., & Bryson, C. N. (2017). The influence of risk assessment instrument scores on evaluators' risk opinions and sexual offender containment recommendations. *Criminal Justice and Behavior, 44*(9), 1213-1235. https://doi.org/10.1177/0093854817707232

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

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS) psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse*, *24*(5), 431-458. http://dx.doi.org/10.1177/1079063211432475

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

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Mishra, S., & Lalumière, M. (2009). The big drop in sex crimes. *The Forum, 21*(2), 41-48.

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. https://doi.org/10.1177/0956797613481812

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator

proceedings. *Psychology, Public Policy, and the Law, 15*(1), 19-53. https://doi.org/10.1037/a0014897

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., Slobogin, C., Saks, M. J., Faigman, D. L., & Geisinger, K. F. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom?. *Psychological Science in the Public Interest, 20*(3), 135-164.

Nellis, A. (2016). *The color of justice: Racial and ethnic disparity in state prisons*. Washington: The Sentencing Project.

Nolan, T. (2021). *Attending to the positive: Retrospective validation of the SAPROF-SO* [Unpublished master's thesis]. University of Canterbury.

Olver, M., Kelley, S., Johnson, L., & Wong, S. (2020). *Violence Risk Scale – Sexual Offense Version (VRS-SO): User's workbook*. Retrieved from https://psynergy.ca/vrs-so

Olver, M. E., Mundt, J. C., Thornton, D., Beggs Christofferson, S. M., Kingston, D. A., Sowden, J. N., Nicholaichuk, T. P., Gordon, A., & Wong, S. C. P. (2018). Using the Violence Risk Scale-Sexual Offense version in sexual violence risk assessments: Updated risk categories and recidivism estimates from a multisite sample of treated sexual offenders. *Psychological Assessment, 30*(7), 941-955. http://dx.doi.org/1037/pas0000538

Olver, M. E., Sowden, J. N., Kingston, D. A., Nicholaichuk, T. P., Gordon, A., Beggs Christofferson, S. M., & Wong, S. C. P. (2018). Predictive accuracy of the Violence Risk Scale – Sexual Offender version: Risk and change scores in treated Canadian Aboriginal and Non-Aboriginal sexual offenders. *Sexual Abuse, 30*(3), 254-275. https://doi.org/10.1177/1079063216649594

Ontario Human Rights Commission (2020). *A disparate impact: Second interim report on the inquiry into racial profiling and racial discrimination of Black persons by the Toronto Police Service*.

Page, M. J., McKenzie, J. E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal, 372*:n71. https://doi.org/10.1136/bmj.n71

People v. Bolton (In re Bolton), 343 Ill. App. 3d 1223, 800 N.E.2d 128, 2003 Ill. App. LEXIS 1387, 279 Ill. Dec. 286 (Ill. App. Ct. 4 Dist., Nov. 13, 2003).

People v. Hargett (In re Hargett), 338 Ill. App. 3d 669, 786 N.E.2d 557, 2003 Ill. App. LEXIS 254, 272 Ill. Dec. 18 (Ill. App. Ct. 3 Dist., Feb. 21, 2003).

People v. Powell (In re Powell), 2016 IL App (1st) 130795-U, 2016 Ill. App. Unpub. LEXIS 608 (Ill. App. Ct. 1 Dist., Mar. 31, 2016

Phenix, A., & Epperson, D. L (2015). Overview of the Development, Reliability, Validity, Scoring, and Uses of the Static-99, Static-99R, Static-2002, and Static-2002R In A. Phenex and H. M. Hoberman (Eds.) *Sexual offending: Predisposing antecedents, assessments and management*. Springer. https://doi.org/10.1007/978-1-4939-2416-5_19

Phenix, A., Fernandez, Y., Harris, A. J. R., Helmus, M., Hanson, R. K., & Thornton, D. (2016). *Static-99R Coding Rules Revised-2016*. static99.org

Phenix, A., Helmus, L., & Hanson, R.K. (2016). *Static-99R & Static-2002R evaluators' workbook*. Available at www.static99.org

Public Safety Canada. (2020). *2019 Corrections and Conditional Release Statistical Overview.* Retrieved from: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ccrso-2018/index-en.aspx

Quesada, S. P., Calkins, C., & Jeglic, E. L. (2014). An examination of the interrater reliability between practitioners and researchers on the Static-99. *International Journal of Offender Therapy and Comparative Criminology*, *58*(11), 1364-1375. http://dx.doi.org/ 10.1177/0306624X13495504

Raymond, B. C., McEwan, T. E., Davis, M. R., Reeves, S. G., & Ogloff, J. R. (2020). Investigating the predictive validity of Static-99/99R scores in a sample of older sexual offenders. *Psychiatry, Psychology and Law*. Advance online publication.http://dx.doi.org/ 10.1080/13218719.2020.1767714

Reeves, S. G., Ogloff, J. R. P., & Simmons, M. (2018). The predictive validity of the Static-99, Static-99R, and Static-2002/R: Which one to use? *Sexual Abuse, 30*(8), 887-907. https://doi.org/10.1177/1079063217712216

Regina v. Awasis, 2016 BCPC 0219 CanLII (BCPC 2016)

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

Rice, A. K. (2016). *Predictive validity of Static-99 and Static-99R scores among offenders scored on multiple occasions*. Unpublished doctoral dissertation. Sam Houston State University.

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. 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. https://doi.org/10.1007/s10979-005-6832-7

Rogers, R. (2003). Forensic use and abuse of psychological tests: Multiscale inventories. *Journal of Psychiatric Practice, 9*(4), 316-320. https://doi.org/10.1097/00131746-200307000-00008

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schulze, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Zeitschrift für Psychologie/Journal of Psychology, 215*(2), 90-103. https://doi.org/10.1027/0044-3409.215.2.90

Scurich, N. (2018). The case against categorical risk estimates. *Behavioral Sciences & the Law, 36*(5), 554-564. https://doi.org/10.1002/bsl.2382

Skaar, C. L. (2013). Inter-rater Reliability of Sex Offender Risk Assessment Measures (RRASOR, MnSOST-R, STATIC-99, AND STATIC-99R) in a Clinical Forensic Setting. Doctoral Dissertation submitted to the faculty of the Wisconsin School of Professional Psychology.

Society for the Advancement of Actuarial Risk and Needs Assessment (SAARNA) (2021). *Static-99R and Static-2002R/BARR-2002R Training Guidelines*. www.saarna.org.

The State of New Hampshire v. Hurley, 2010 N.H. Sup Ct. Unpub. No. 07-E-0236 (N.H. Sup. Ct., Apr. 22, 2010)

The State of New Hampshire v. Ploof, 2009 NH Sup. Ct. Unpub. No. 07-E-0238 (N.H. Sup. Ct., Apr. 28, 2009).

State v. C.B., 2020 N.J. Super. Unpub. LEXIS 813, 2020 WL 2096147 (N.J. Sup. Ct. App. Div., May 1, 2020).

State v. Jones (In re Jones), 2018 WI 44, 381 Wis. 2d 284, 911 N.W.2d 97, 2018 Wisc. LEXIS 221, 2018 WL 2071928 (Wis. S. Ct., May 4, 2018)

State v. Martin, 2013 Ariz. App. Unpub. LEXIS 973, 2013 WL 4774143 (Ariz. Ct. App, Div. 1, Sept. 5, 2013

State v. Ragan, 2011 N.H. Sup. Ct. No. 10-E-64 Lexis 110 (N.H. Sup. Ct., Apr. 12, 2011)

State v. Wortman, Wis. Cir. Ct. Oral Ruling (Wis. Cir. Ct., May 4, 2017)

Stephens, S., Newman, J. E., Cantor, J. M., & Seto, M. C. (2018). The Static-99R predicts sexual and violent recidivism for individuals with low intellectual functioning. *Journal of sexual aggression*, *24*(1), 1-11. https://doi.org/10.1080/13552600.2017.1372936

Thornton, D. (1997). *A 16-year follow-up of 563 sexual offenders released from HM Prison Service in 1979.* Unpublished raw data.

Thornton, D. (2016, January 19). *An affidavit*. http://www.static99.org/pdfdocs/Legal-Thorton-Hanson_Affidavits(2016).pdf

Thornton, D., Hanson, R. K., Kelley, S. M., & Mundt, J. C. (2021). Estimating lifetime and residual risk for individuals who remain sexual offense free in the community: Practical applications. *Sexual Abuse, 33*(1), 3-33. doi:10.1177/1079063219871573

Thornton, D., & Helmus, L. M. (2021, September). *Using research to guide application of Static-99R.* Full-day conference workshop at the Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, virtual conference.

Thornton, D., & Knight, R. A. (2015). Construction and validation of SRA-FV need assessment. *Sexual Abuse*, *27*(4), 360-375. http://dx.doi.org/10.1177/1079063213511120

Truth, & Reconciliation Commission of Canada. (2015). *Canada's Residential Schools-Missing Children and Unmarked Burials: The Final Report of the Truth and Reconciliation Commission of Canada* (Vol. 4). McGill-Queen's Press-MQUP.

United States v. Carta, 2011 U.S. Dist. LEXIS 73007, 2011 WL 2680734 (U.S. D. Ct. Mass., Jul. 7, 2011).

United States v. McIlrath, 512 F.3d 421, 2008 U.S. App. LEXIS 442 (U.S. App. Ct. 7 Cir., Jan. 10, 2008)

United States v. McIlrath, *aff'd,* 512 F.3d 421 (7th Cir. 2008).

Varela, J. G., Boccaccini, M. T., Cuervo, V. A., Murrie, D. C., & Clark, J. W. (2014). Same score, different message: Perceptions of offender risk depend on Static-99R risk communication format. *Law and Human Behavior*, *38*(5)*,* 418-427. https://doi.org/10.1037/lhb0000073

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.

Wakefield, H., & Underwager, R. (1993). Misuse of psychological tests in forensic settings: Some horrible examples. *American Journal of Forensic Psychology, 11*(1), 55–75.

Willis, G., & Thornton, D. (2021, May). *Introducing the SAPROF-SO version 1*. Symposium conducted at NYS ATSA & The Alliance Virtual Joint Conference.

Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *International Journal of Forensic Mental Health, 16*(3), 227-238.

Zapf, P. A., & Grisso, T. (2012). Use and misuse of forensic assessment instruments. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony: Based on the original work by Jay Ziskin* (p. 488–508). Oxford University Press. https://doi.org/10.1093/med:psych/9780195174113.003.0020

Table 1
Meta-Analysis Results

	Fixed effect			Random effects						
	AUC	LL	UL	AUC	LL	UL	*k*	*n*	*Q*	I^2
All	.678	.669	.686	.689	.671	.706	56	71,515	182.0***	69.8
Country										
Canada	.693	.674	.711	.693	.674	.711	14	9,361	11.6	0.0
U.S.	.652	.641	.664	.668	.642	.693	24	53,887	75.3***	69.4
Europe	.731	.712	.751	.727	.693	.762	12	5,918	32.6***	66.2
Australia/New Zealand	.697	.663	.731	.695	.641	.748	5	2,215	9.0	55.3
Other	.700	.504	.896	.700	.480	.916	1	134	-	-
$Q_{between}$		=53.6***			=7.5					
Used in Norms										
No	.669	.659	.679	.686	.660	.710	29	56,882	123.0***	77.2
Yes	.694	.680	.708	.693	.667	.718	27	14,633	51.0**	49.0
$Q_{between}$		=8.1**			=0.2					
Used to develop Static-99R										
No	.674	.665	.684	.692	.667	.717	33	62,099	139.4***	77.0
Yes	.687	.671	.703	.685	.661	.710	23	9,416	40.9**	46.2
$Q_{between}$		=1.7			=0.1					
Developers are co-authors										
No	.689	.678	.700	.692	.672	.712	46	27,727	124.2***	63.8
Yes	.664	.651	.676	.679	.641	.717	10	43,788	48.9***	81.6
$Q_{between}$		=9.0**			=0.3					
Sample Type										
Routine	.679	.669	.689	.699	.676	.723	28	61,552	116.5***	76.8
Treatment	.691	.671	.711	.695	.661	.730	13	5,532	31.9**	62.4
High risk/need	.618	.585	.651	.618	.585	.651	7	1,933	4.6	0.0
Other	.692	.660	.723	.700	.650	.748	8	2,498	13.7	48.8
$Q_{between}$		=15.3**			=9.1*					
Mentioned coding manual										
No	.663	.647	.679	.663	.632	.693	19	16,312	33.1*	45.6
Yes	.683	.673	.693	.701	.680	.722	37	55,203	144.6***	75.1
$Q_{between}$		=4.3*			=4.2*					
Appropriate Training										
No/unknown	.672	.663	.681	.685	.665	.704	47	62,091	134.2***	65.7
Yes	.711	.689	.732	.708	.665	.750	9	9,424	37.3***	78.5
$Q_{between}$		=10.5**			=1.0					
Recidivism Criteria										
All[b]	.673	.665	.682	.687	.670	.704	54	69,767	141.3***	65.9
Arrests/charges	.666	.657	.676	.682	.661	.703	34	60,100	113.8***	71.0
Convictions	.697	.679	.714	.697	.668	.725	20	9,667	19.0	0.2
$Q_{between}$		=8.5**			=0.7					

*$p < .05$; **$p < .01$; ***$p < .001$
[a]$p = .054$
[b]Two studies were excluded because their recidivism criteria were something other than charges or convictions