American Psychological Association
American Psychology-Law Society

Law and Human Behavior

© 2023 American Psychological Association
ISSN: 0147-7307

2023, Vol. 47, No. 1, 275–291
https://doi.org/10.1037/lhb0000517

Predictive Accuracy of Static-99R Across Different Racial/Ethnic Groups: A Meta-Analysis

Simran Ahmed[1], Seung C. Lee[2], and L. Maaike Helmus[1]
[1] Department of Criminology, Simon Fraser University
[2] Society for the Advancement of Actuarial Risk Needs Assessment, Kingston, Ontario, Canada

Objective: The overrepresentation of numerous racial/ethnic groups in the criminal legal system warrants examination of the cross-cultural applicability of risk assessment tools. Static-99R is a tool used in diverse countries to assess sexual recidivism risk. We conducted a meta-analysis on the predictive accuracy of Static-99R across different racial/ethnic groups. ***Hypotheses:*** No hypotheses were made regarding discrimination, given that past research could support hypotheses of differential or equivalent accuracy. We hypothesized that Indigenous individuals would score higher on Static-99R than non-Indigenous or White individuals. ***Method:*** Our search identified 18 eligible documents (from 17 distinct studies) with 41 nonoverlapping effect sizes. These 17 studies examined the predictive accuracy of Static-99R with racially/ethnically diverse men charged with or convicted of sexually motivated offenses. We report analyses using both fixed-effect and random-effects meta-analysis. ***Results:*** Indigenous and Black individuals scored significantly higher on Static-99R than their non-Indigenous or White counterparts, with small effect sizes. For discrimination, area under the curve (AUC) values were generally moderate-to-large and statistically significant for all groups in both fixed-effect and random-effects analyses. Within-study subgroup analyses indicated significantly lower accuracy for Indigenous and Hispanic individuals compared with White/non-Indigenous samples (though for Hispanic individuals, this finding was significant only in the fixed-effect analyses). No statistically significant differences in accuracy were found between White and Black individuals. Static-99R significantly predicted recidivism with large effect sizes across two samples of Asian individuals. Two studies supported calibration across Black, White, and Hispanic individuals. Two studies examining calibration of Static-99R for Indigenous individuals had mixed findings. ***Conclusions:*** Given a small number of studies and limitations with both the fixed- and random-effects analyses, readers should interpret findings regarding Hispanic individuals with caution. The analyses clearly found significant but lower accuracy for Static-99R with Indigenous individuals. Potential reasons for this differential accuracy are discussed, along with limitations of the meta-analysis and suggestions for research and practice.

Public Significance Statement
Across numerous studies, results appear to support Static-99R's ability to accurately measure sexual recidivism risk among Black, White, and Asian individuals charged with or convicted of sexually motivated offenses. It also predicts recidivism for Indigenous and Hispanic individuals but with lower accuracy. Given that the scale is still predictive for these latter groups, its use should not be abandoned unless and until research has identified empirically superior alternatives.

Keywords: Static-99R, risk assessment, race, cross-cultural validity, Indigenous peoples

Supplemental materials: https://doi.org/10.1037/lhb0000517.supp

Stephane Shepherd served as Action Editor.

Seung C. Lee https://orcid.org/0000-0002-6705-4535
L. Maaike Helmus https://orcid.org/0000-0002-5032-2548

This work was completed on the traditional and unceded territories of the Coast Salish Peoples (where the Canadian city of Vancouver, British Columbia, is currently located), specifically the Squamish (Sḵwx̱wú7mesh Úxwumixw), Tsleil-Waututh (səl̓ilw̓ətaʔɬ), and Musqueam (xʷməθkʷəy̓əm) nations, and also on the traditional and unceded territory of the Algonquin nation (where the Canadian city of Ottawa, Ontario, is currently located).

The meta-analysis data sets (and syntax template) are publicly available on OSF (https://osf.io/bk84n/).

L. Maaike Helmus is a certified trainer on Static-99R. The government of Canada holds the copyright for this measure, and the scale authors do not receive royalties for its use.

The data are available at https://osf.io/bk84n/?view_only=c3e143125259450c845ca81b08f65ebf.

Correspondence concerning this article should be addressed to Simran Ahmed, Department of Criminology, Simon Fraser University, Saywell Hall, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. Email: Simran_ahmed@sfu.ca

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Risk assessments are a routine part of virtually all decisions in the criminal legal system and have important consequences for both public safety and the individual being assessed. The development, validation, and implementation of risk assessment tools to predict sexual recidivism have been an active area of research in recent decades (e.g., Hanson, 2009; Hanson & Morton-Bourgon, 2009; Kelley et al., 2020; Neal & Grisso, 2014). Using structured, evidence-based risk assessment tools is the best practice (e.g., see the professional guidelines from the Association for the Treatment of Sexual Abusers [ATSA], 2014). Importantly, however, using an evidence-based risk assessment tool implicitly or explicitly assumes that the individual being assessed is similar enough to those in the development and validation samples that the research is likely to generalize to them. Most often, this is a reasonable assumption. The field of psychology is predicated on the principle that research on groups of individuals is generally useful when making decisions about specific individuals (even though it will never perfectly speak to each unique consideration of the individual at hand). However, there is considerable debate over whether this assumption would apply to individuals across different racial, ethnic, or cultural groups.

First, we must note difficulties in defining and delineating race, ethnicity, and culture. Scientific research has been wildly inconsistent in their definitions of these terms, when they have bothered to define them at all (McKenzie & Crowcroft, 1994). The American Psychological Association (American Psychological Association [APA], 2020) publication guidelines note that race emphasizes physical differences that are socially defined as important to certain groups. This social construct was created and used to justify social, political, and economic oppression of non-White groups by White people (https://www.racepowerofanillusion.org/). Ethnicity, in contrast, emphasizes shared cultural characteristics, such as language or ancestry. APA notes the importance of considering both common definitions (e.g., applied by governments) and participants' self-designation. The American Medical Association has recently summarized the work of a committee to provide guidance for reporting race and ethnicity in medical research, noting limitations including a lack of clear consensus and shifting definitions over time (Flanagin et al., 2021). Given ambiguity, overlap, and the socially constructed nature of both race and ethnicity, there have been some calls for combining the terms into a single word (*race/ethnicity*; Flanagin et al., 2021). Although both race and ethnicity are related to shared culture among groups, culture itself is a much broader term. Race and ethnicity are often related to ancestry or national identities, whereas culture refers more broadly to shared values, beliefs, and customs of a group, which could be defined on the basis of race, ethnicity, or any other social categorization. For example, many workplaces have their own distinct culture.

For the purposes of this article, we will err on the side of using race/ethnicity as a single term to refer to groups defined on the basis of physical, national, or ancestral characteristics. This article largely defers to how race/ethnicity was categorized in the studies we summarize, noting that these definitions are likely to vary (and most often are not defined at all). We generally avoid the term "culture" given its broader connotations, but we do use the term "cross-cultural validity" given its common usage in psychology; however, here we are referring to cross-cultural validity based on race/ethnicity.

It is indisputable that race, culture, and ethnicity matter in understanding people's identities and how they experience the world (APA, 2021; DiAngelo, 2018; Tamatea, 2017). It is also clear that certain racial/ethnic groups are treated differently in the criminal legal system, but that does not necessarily mean that risk assessment scales will predict recidivism differently across these groups. Racial/ethnic differences can clearly exist, but they may or may not be risk relevant. In other words, the same risk factors and scales may be equally predictive across racial/ethnic groups, even if those groups are not treated equally in certain countries. Alternatively, they may be equally predictive but not adequately operationalized in current risk scales or are assessed with considerable evaluator bias, impacting their predictive accuracy across groups.

Importance of Examining Predictive Accuracy of Risk Scales Across Race/Ethnicity

Most risk assessment development and validation research is produced in countries with a history of slavery and colonization, the legacy of which is still evident today in multiple domains (not just in the criminal legal system; Alexander, 2010; DiAngelo, 2018; Monchalin, 2016; Truth and Reconciliation Commission [TRC] of Canada, 2015). In particular, the ways in which psychology has historically contributed to racism have been well documented (APA, 2021). One ongoing impact of systemic racism is the overrepresentation of certain racial/ethnic groups in correctional systems around the world, including Black, Hispanic, and Indigenous peoples in the United States (Nellis, 2016; Travis et al., 2014), Aotearoa (New Zealand), Māori (Tamatea, 2017), Indigenous Australians (Australian Bureau of Statistics, 2021), British Black citizens in the United Kingdom (Ministry of Justice, 2013), and Canadian Indigenous youth and adults, who are overrepresented by approximately five- to eightfold in their respective criminal legal systems (Department of Justice Canada, 2017; Public Safety Canada, 2020). Additionally, there is evidence in Canada that Indigenous peoples are overrepresented at rates exponentially higher than Black or Hispanic individuals (Brankley & Lee, 2016). However, not all racial/ethnic minority groups are overrepresented or equally overrepresented. For example, East Asian individuals tend to be underrepresented (Hu & Esthappan, 2017). This indicates that overrepresentation is more complex than simply delineating White and non-White groups.

Reasons for the overrepresentation of certain racial/ethnic groups in the criminal legal system are myriad and complex (Shepherd & Ilalio, 2016), with longstanding roots in colonialism and slavery (Alexander, 2010; Monchalin, 2016; TRC, 2015). Research has primarily focused on Black individuals in the United States and Indigenous peoples in Canada. Although these groups have dramatically different cultures and histories, what they share is that former British colonies dispossessed them of their lands and traditions (either through the slave trade or colonization) and subjected them to legal systems that were not compatible with their traditions and that were used to exploit them for the benefit of the White majority. More proximal factors contributing to this overrepresentation are higher rates of policing, prosecution, and use of force, which have been seen in Indigenous communities in the United States, Canada, Australia, and New Zealand (Fergusson et al., 2003; Nettelbeck & Smandych, 2010; Perry, 2006) and in Black communities in Canada and the United States (Bolger, 2015; Ontario Human Rights Commission, 2020).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Given these longstanding injustices, it is necessary to explore whether risk assessment scales function similarly across groups or whether they may perpetuate or exacerbate disparities. Concerns have been raised, given that most risk assessment scales have been developed and validated on samples that are predominantly White (Day et al., 2018; Shepherd & Lewis-Fernandez, 2016; Shepherd & Willis-Esqueda, 2018). This does not necessarily mean the scales are invalid or biased against these groups, but it does warrant examination. The standards promulgated by the Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association et al. (2014) provide guidance for the application of existing assessment tools to new or different populations. They do not require that every possible subgroup or unique case undergo separate validation research. Instead, the standards emphasize that test development and validation must consider *relevant* subgroups (see Chapter 3). In other words, we can often presume that risk scales will broadly generalize to diverse populations unless there is theory or research to suggest that a scale may predict different results for a relevant subgroup. The unique history of many non-White groups (particularly Indigenous and Black individuals)—colonial legal systems, the legacy of slavery and colonization, and its remnants in modern-day systemic racism—provides strong justification for cross-cultural validation (see also the Canadian Supreme Court decision in *Ewert v. Canada*, 2018).

Risk Assessment and Predictive Validity

Before examining the cross-cultural validity of risk assessment, it is important to consider what risk assessment is and how it should be evaluated. Risk assessment tools are used for diverse decisions, some dichotomous (e.g., whether an individual should be granted parole or civilly committed), and some related to triaging of services or sanctions (e.g., treatment or supervision intensity). Nonetheless, risk tools themselves are generally intended to measure risk as a continuum/dimension (Hanson et al., 2013) and should not be conflated with legal/policy decisions they are meant to inform.

The Joint Committee et al.'s (2014) Standards for Educational and Psychological Testing note that not all forms of reliability and validity (e.g., construct validity, concurrent validity, face validity, and predictive validity) from the psychometrics literature are relevant to all measures. Instead, they emphasize that the types of evidence needed to support the validity of a scale depend on the particular purpose, use, and interpretation of the scale. Risk assessment scales are criterion-referenced and prognostic measures (Hanson, 2022; Helmus & Babchishin, 2017). The primary implications of this are that risk should be communicated on a continuum as opposed to dichotomously (e.g., this person will or will not reoffend; ATSA, 2014), and the primary validity evidence to consider for risk scales is predictive accuracy.

Predictive accuracy can be further divided into two main types: discrimination and calibration (Hanson, 2022). Discrimination considers how well the scale rank orders individuals in terms of risk to reoffend. This speaks to the confidence with which we can say higher risk individuals are more likely to reoffend than lower risk individuals, regardless of what the absolute rate of recidivism is. Calibration, on the other hand, examines the extent to which the recidivism rate per risk score or category is similar between groups. Calibration can be examined in terms of whether recidivism estimates from the scale's normative data apply to new samples/groups or whether the recidivism estimates per score vary across racial/ethnic groups in the same sample (with the latter being of particular relevance to the issues in the present study).

Helmus and Babchishin (2017) outline statistics commonly used to report discrimination and calibration. The most common statistics for assessing discrimination and calibration of risk scales are area under the curve (AUC) and the expected/observed (E/O) index, respectively. Helmus and Babchishin (2017) also eschew the use of any diagnostic statistics that are predicated on artificially dichotomizing risk assessment scores as a proxy for classifying individuals as recidivists or nonrecidivists (e.g., sensitivity, specificity, positive predictive accuracy, and true positives), as these methods tend to be arbitrary and out of sync with the recommended use and interpretation of risk scales.

Cross-Cultural Validity of Risk Assessment

It is not fully clear how overrepresentation, systemic racism, and disparate opportunities may impact the accuracy of risk assessment tools across different racial/ethnic groups. As a result of systemic racism, we would hypothesize that generations of disparities have systematically exposed many non-White groups to higher rates of adverse circumstances and outcomes, many of which would overlap with criminogenic needs. Consequently, non-White groups would likely have a higher prevalence of risk factors for crime. This hypothesis is consistently supported in individual studies and meta-analyses (Gutierrez et al., 2016; Olver et al., 2014; Perley-Robertson et al., 2019; Shepherd et al., 2014).

There have been attempts to reconcile and summarize varying definitions of test bias or racial/ethnic fairness, both specific to risk assessment (e.g., Ashford et al., 2022) and more broadly (Warne et al., 2014). Given that mean group differences in scores (also referred to as *statistical parity*) are more a reflection of systemic inequality rather than test bias (Warne et al., 2014) and that risk scales are fundamentally intended to predict recidivism, what is of primary importance for evaluating cross-cultural validity is examining differential predictive accuracy of the scale (and ideally the individual items as well). Differential predictive accuracy can be defined and assessed in numerous ways (e.g., Ashford et al., 2022), but we follow the primary analyses of discrimination and calibration recommended by Hanson (2022) and Helmus and Babchishin (2017).

Regarding differential accuracy, it is not fully understood whether the underlying causes and predictors of crime and recidivism would differ across groups. However, their manifestation and measurement might. Operationalizations of risk factors in current assessment scales may not be culturally sensitive (e.g., evaluations of family dynamics may be strongly rooted in a Western focus on the nuclear family). Additionally, overpolicing and overprosecution could complicate the measurement of criminal history or recidivism, as both these variables are flawed proxies of actual criminal behavior, which may be detected and sanctioned differently across racial/ethnic groups (Hanson, 2020). This may explain why research has found criminal history risk factors to have lower predictive accuracy for Indigenous individuals (Gutierrez et al., 2013; Perley-Robertson et al., 2019).

Examining the Level of Service Inventory (LSI) family of risk assessment scales, which are widely used and based on the well-researched Central 8 risk factors for crime (i.e., criminal history, procriminal personality, procriminal attitudes, procriminal associates,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

family/marital problems, education/employment problems, substance abuse, and poor use of leisure/recreation time Bonta & Andrews, 2017), Olver et al. (2014) conducted an initial comprehensive meta-analysis. Broad analyses of all samples (mainly from Canada and the United States, followed by some Oceanic, European, and Asian countries) found some evidence for significant but lower predictive accuracy among racial/ethnic minorities in the fixed-effect meta-analyses but not the random-effects results. In their meta-analysis, the random-effects results are likely the more credible findings.[1] Olver et al. (2014) found good predictive accuracy for specific groups (i.e., Indigenous, Asian, and Hispanic). The fixed-effect results revealed lower accuracy for Black individuals, but this appeared to be attributable to one outlier study; its removal resulted in similar findings for Black individuals compared with the other groups. To further reduce the variability across studies that may be disguising patterns, Olver et al. (2014) restricted analyses to comparisons of racial/ethnic groups within the same study (removing between-study variability). These analyses similarly found comparable predictive accuracy in the random-effects analyses but somewhat lower accuracy for racial/ethnic minorities in the fixed-effect analyses (e.g., for general recidivism, a correlation of .23 for racial/ethnic minorities and .32 for nonminority individuals).

In contrast, a meta-analysis from the same year examined in more detail the LSI scales among Indigenous and non-Indigenous individuals (Wilson & Gutierrez, 2014). This meta-analysis found that the LSI scales and the subscales (based on the Central 8 risk factors) did predict recidivism for Indigenous individuals, but the total scores (and most subscales) had significantly lower predictive accuracy for non-Indigenous individuals. A possible reason that this meta-analysis found clearer evidence of disparate accuracy could be because they used Cohen's *d* as their effect size. Olver et al. (2014) used correlations as the effect size metric, which are strongly influenced by the base rate of recidivism (Babchishin & Helmus, 2016). Given higher base rates of recidivism for Indigenous individuals, this bias would minimize differences in predictive accuracy.

Static-99R and Cross-Cultural Validity

Static-99R (Hanson & Thornton, 2000; Helmus et al., 2012) is the most widely used tool to assess risk of sexual recidivism for men charged with or convicted of a sexually motivated offense (Bourgon et al., 2018; Kelley et al., 2020; Neal & Grisso, 2014; for a detailed overview of the scale, including its psychometric properties, see Helmus et al., 2022). It contains 10 static risk factors based on commonly available demographic and criminal history information. The total score can range from −3 to 12 and can be used to place individuals assessed in one of five risk categories based on the Justice Center Standardized Risk Level System (Hanson, Bourgon, et al., 2017): Level I, very low (−3 to −2); Level II, below average (−1 to 0); Level III, average (1–3); Level IVa, above average (4–5); and Level IVb, well above average (6+; Hanson, Babchishin, et al., 2017). A recent meta-analysis (from which the current studies were drawn) found Static-99R to be a moderate predictor of sexual recidivism (AUC = .69, 95% confidence interval [CI: .67, .71], $k = 56$, $n = 71{,}515$; Helmus et. al., 2022).

Although Static-99R is the most extensively researched sexual recidivism risk tool, scholars have raised concerns about the measure's cross-cultural validity and utility (e.g., Lee & Hanson, 2017; Leguízamo et al., 2017; Myer, 2019; Smallbone & Rallings, 2013; Varela et al., 2013). The original version of the scale (Static-99) was developed using three Canadian samples and an English/Welsh sample (Hanson & Thornton, 2000), which would have consisted of predominantly White individuals. Given the frequent use of the scale in diverse countries (for a review, see Helmus et al., 2022), there is an urgent need to ascertain its predictive validity across different racial and ethnic groups.

Two meta-analyses have been conducted to date, focusing solely on comparisons of the scale for Indigenous and non-Indigenous individuals. Across five Canadian samples ($n = 1{,}588$; 319 Indigenous and 1,269 non-Indigenous participants), Babchishin et al. (2012) found that the scale had high predictive accuracy for both groups (AUCs = .71 and .74, respectively), and the difference in accuracy was not statistically significant. Although these results are encouraging, this meta-analysis was preliminary; sample sizes were small and, at the time, only Canadian research was available.

More recently, Lee, Hanson, and Blais (2020) examined Static-99R among Indigenous and White individuals but restricted the analysis to samples that were preselected as unusually high risk/need. Previous research has identified that high risk/need samples are meaningfully different from routine correctional samples and tend to have lower predictive accuracy for Static-99R (Hanson et al., 2016), likely because of a restriction-of-range on measured and unmeasured risk factors. This meta-analysis also included five Canadian samples; two were from Babchishin et al. (2012), one was from the same setting but was larger than those used by Babchishin et al. (2012), and two were new. For 5-year follow-up analyses, Static-99R predicted sexual recidivism with significant but small accuracy for Indigenous individuals (AUC = .61, $n = 599$) and moderate accuracy for White individuals (AUC = .70, $n = 1{,}375$). However, logistic regression analyses indicated that the differences in accuracy between groups were not statistically significant.

The current meta-analysis incorporates the studies from these previous meta-analyses and other studies from the Helmus et al. (2022) meta-analysis where it was possible to disaggregate effect sizes for different racial/ethnic groups. This meta-analysis did not examine research on the original version of the scale (Static-99) given that the developers no longer recommend its use (Helmus et al., 2012) or update the normative data for it. Nonetheless, available research on the earlier version can still be informative about cross-cultural validity, even if not incorporated into the current analyses. Many Static-99 studies with race/ethnicity data or analyses were subsequently included in the Static-99R normative data and hence were included in the current meta-analysis, but there are two additional studies of Static-99 only that are worth mentioning.

Watanabe et al. (2007) examined an average 14-year follow-up of 406 individuals arrested for child rape in Japan. Static-99 demonstrated a significant but small predictive effect (AUC = .62). Långström (2004) analyzed the predictive accuracy of Static-99 among men released from prison in Sweden, with an average follow-up of 5.7 years. Participants were grouped on the basis of citizenship. The scale demonstrated large and significant predictive effect

[1] When there is considerable variability in findings across samples (which there was in this meta-analysis), fixed-effect results indicate unrealistically narrow confidence intervals; random-effects results are generally more plausible but are less stable when the number of studies is less than 30 (Schulze, 2007).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

sizes for Nordic and non-Nordic individuals but no accuracy for the combined group of individuals with citizenship from other countries (primarily in Africa, Asia, and the Middle East). These findings should be considered with caution given that the "other" group combined so many diverse continents and was based on a small sample ($n = 128$) with only four sexual recidivists (statistical power for AUCs is largely determined by the number of recidivists). Additionally, given that this group was defined on the basis of citizenship as opposed to race/ethnicity, it is possible that many of these individuals were deported after being released from prison, reducing the ability to detect recidivism and measure predictive accuracy.

The Present Study

Exploring the cross-cultural validity of risk assessment scales is imperative. Static-99R is one of the few scales for which sufficient research is available to conduct a meta-analysis. Two meta-analyses have examined the scale specifically with Indigenous individuals, but the first meta-analysis is already a decade old, and the second was restricted to a unique setting. Consequently, there is a need for an updated and comprehensive summary of the scale's discrimination accuracy across diverse racial/ethnic groups. This meta-analysis provides such a summary. No hypotheses were made so as not to conflict with hypotheses from previous research. For example, Babchishin et al. (2012) found similar predictive accuracy for Indigenous individuals, but larger meta-analyses of general recidivism risk assessment have found lower accuracy (e.g., Wilson & Gutierrez, 2014).

Although mean group differences are not a form of test bias (Warne et al., 2014), we analyzed them as well, given that they provide valuable information about risk factor prevalence between groups, which can guide policy and interventions. Given previous research, we hypothesized that Indigenous and Black individuals would have higher risk scores than their White or non-Indigenous counterparts. No hypotheses were made about comparisons between Hispanic and White individuals. Last, given limited research, we provided a narrative summary of research on calibration across different racial/ethnic groups.

Method

To be included in the current meta-analysis, studies had to contain a sample of adult males charged with or convicted of a sexually motivated offense (including both contact and noncontact sexual offenses), and men in the sample had to have scores on Static-99R, race/ethnicity information, and assessments of sexual recidivism at a follow-up period. We also required sufficient statistical information to calculate the number of recidivists, nonrecidivists, the effect size (AUC), and its variance for the predictive accuracy of Static-99R. All studies needed to have a total sample size of at least 10 with at least one recidivist. Overlapping samples were excluded so as to include only unique samples; where possible, we retained the largest study from that sample.

We included studies for which it was possible to code predictive accuracy for two or more racial/ethnic subgroups, as well as one study from Singapore (Tsao & Chu, 2021), which was coded as Asian (100% of the sample was of Asian descent, with 72% being Chinese; I. T. Tsao, personal communication, August 4, 2022). We did not include and separately analyze studies from specific countries that were predominantly White but would have cultural differences (e.g., single studies from different Western European countries). First, with typically only one study from each of those European countries, the value of such comparisons would be limited and more easily explained by idiosyncratic differences in methodology than cultural differences. Second, given the absence of information of prejudice/racism against subgroups of White individuals (e.g., Scandinavian) in the criminal legal system, we were not interested in exploring cross-cultural validity within White groups.

Procedure

The procedure for this meta-analysis was admittedly unorthodox. It was a subset and expansion of a larger meta-analysis on all Static-99R predictive accuracy studies (Helmus et al., 2022). In the original meta-analysis, one of the authors searched the keyword "Static-99*" or "Static 99*" in the document text for the following three academic databases: PsycINFO, Criminal Justice Abstracts, and ProQuest Dissertations and Theses. Additionally, Google Scholar was searched with the same keywords. Results were examined from 2009 onward because that was the earliest year in which the revised scale was disseminated. All studies of Static-99R were examined to determine whether there was sufficient information to code an effect size for predictive accuracy for any sexual recidivism in the community among males charged with or convicted of a sexually motivated offense. If it was clear from the abstract that there was no recidivism follow-up data, we did not necessarily retrieve the full text; the full text was retrieved in any case in which the title or abstract did not rule out the possibility of recidivism data being reported. After compiling results from this search, we sent our list to the ATSA LISTSERV to request any additional or upcoming studies we were not aware of. The search was completed on February 21, 2021.

Two authors from the original meta-analysis coded all studies for descriptive information and effect sizes for the overall groups (not for racial/ethnic subgroups). Interrater reliability was not practical (as the authors coded the studies at different times), but all studies were double-coded and consensus ratings were generated when there were disagreements (of which there were few). The second author of this article additionally coded the effect sizes for racial/ethnic subgroups.

Subsequently, the first author of this article undertook this meta-analysis as part of a graduate course on meta-analysis, taught by the third author of this article (and one of the coders of the Helmus et al., 2022, meta-analysis). The first author conducted a search and coded all studies blind to the larger meta-analysis. The first author searched Google Scholar, PsycINFO, Criminal Justice Abstracts, and ProQuest Dissertations and Theses for the combinations of "Static-99R" (including variations with and without the *R* or the hyphen) and the keywords "race," "ethnic," "Asian," "Indigenous," "Latino," "Hispanic," "Black," and "Aboriginal" in the title or abstract to target specific groups.

After the first author completed their coursework of this meta-analysis, they were given access to the larger project to amalgamate search results and study coding. Differences in coding from the larger meta-analysis were resolved with one or both raters from the larger meta-analysis. The author from the larger meta-analysis who

coded effect sizes for the combined sample (but not the racial/ethnic subgroups) verified those data before conducting analyses. Although formal interrater reliability analyses were not conducted, the study coding was generally reviewed by all three coauthors.

As of November 1, 2021, our search yielded 18 eligible documents from 17 settings (i.e., study samples) with 41 nonoverlapping groups of individuals. When possible, the most precise racial/ethnic group was coded (e.g., we coded White rather than non-Indigenous individuals if both were available). The category of "non-Indigenous" was used solely in Canadian studies that compared Indigenous with non-Indigenous individuals (i.e., a sample of Hispanic individuals was not also coded as non-Indigenous). On the basis of the demographic characteristics of correctional populations in Canada, we can assume that the non-Indigenous category is almost entirely White with a small proportion of non-Indigenous, non-White individuals. However, given that individuals in this category were not exclusively White, it was coded separately from analyses of White individuals.

Overview of Analyses

The meta-analysis data sets (and syntax template) have been made publicly on OSF (https://osf.io/bk84n/). The study design and analyses were not preregistered.

Effect Size

Cohen's *d* was used to summarize group differences in Static-99R scores, with values of 0.20, 0.50, and 0.80 considered small, moderate, and large, respectively (Cohen, 1992). The effect size indicator used to summarize the relative predictive accuracy of Static-99R was the AUC from receiver operating characteristic (ROC) analysis. The AUC value indicates the likelihood that a randomly selected recidivist will score higher on the Static-99R than a randomly selected nonrecidivist. AUC values range from 0 to 1, and Rice and Harris (2005) suggest that, as a heuristic for interpretation, an AUC of .56 is small, .64 is moderate, and .71 is large (these correspond to Cohen's *d* values of 0.20, 0.50, and 0.80, respectively). An AUC value is statistically significant if the 95% CI does not include .50.

Aggregation of Findings

Meta-analyses followed the formulae of Borenstein et al. (2021) in SPSS Version 27. Although random-effects analyses are often conceptually preferable because they have greater generalizability, they are unstable when the number of studies is small (<30; Schulze, 2007), which was the case in virtually all of the present analyses. In contrast, however, the greater the variability across studies, the more unrealistically narrow the CIs from fixed-effect analyses are. Consequently, we report both fixed- and random-effects analyses. Variability in findings across studies is reported using Cochran's Q statistic and the I^2 effect size statistic (Borenstein et al., 2021). 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). Studies were identified as outliers following the criteria used by Hanson and Bussière (1998; i.e., if variability across studies is significant and reduces by more than 50% with the potential outlier removed), and results are presented both with and without the potential outlier.

To examine differences in predictive accuracy between some racial/ethnic groups and a comparison group (either White or non-Indigenous offenders), the strongest test must recognize that samples are matched within a study (to remove between-study variability). To conduct these within-study comparisons, we calculated new effect sizes (difference in AUCs) by subtracting the effect size of one group (e.g., non-Indigenous) from the effect size of another group (e.g., Indigenous). The variance of this difference score was the sum of the variance of each AUC minus their covariance, which was defined as $2 \times r \times SD_{\text{Group1}} \times SD_{\text{Group2}}$, where r is the correlation between the two effect sizes (Ley, 1972). If the 95% CI for the difference between the AUCs does not include zero, then the difference is statistically significant. For comparisons of Black and Hispanic individuals with White individuals, we excluded one study (Swinburne Romine et al., 2008) when calculating the correlation between the effect sizes because the effects for Black and Hispanic individuals were based on such small *n*s that they produced improbable within-study correlations (i.e., slightly negative). Their removal resulted in large positive correlations between the effect sizes of the two groups, which would be expected given that they are nested within studies.

Results

Definition of Racial/Ethnic Subgroups

The included studies categorized participants in groups such as White, Black, Indigenous, or Hispanic. Definitions were generally not provided for these terms, nor did the articles indicate how individuals with mixed racial/ethnic identities were categorized. Articles did not provide information about whether race/ethnicity was based on self-reported identification or an official record. For the Canadian studies, on the basis of our knowledge of the correctional systems, we presume that Indigenous ancestry was typically self-reported. In other studies, it is not clear. It is likely that in many jurisdictions where race/ethnicity is recorded in correctional file information, it is sometimes based on the person's self-reported identification, and at other times correctional staff or police may have presumed it on the basis of a person's skin color or last name. It may also have been defined on the basis of country of birth, citizenship, or primary language. High rates of missing information on race/ethnicity (e.g., Boccaccini et al., 2017) are an additional problem.

Study Descriptors

Table 1 provides descriptive information on the 18 documents coded (reflecting 17 study settings and 41 effect sizes). For the meta-analysis from Lee, Hanson, and Blais (2020), we coded effect sizes from two of the samples in independent study settings (Brankley et al., 2017; Lee, Mularczyk, et al., 2018, as cited in Lee, Hanson, and Blais, 2020), as these were from unpublished data sets that did not overlap with any other studies in this meta-analysis. For two other study samples (Studies 8 and 16), there were two documents from that sample/setting, with different subgroup analyses available in different documents; they were treated as the same study setting but different documents (Studies 8.1 and 16.1, respectively). The total sample consisted of 47,139 males charged with or convicted of sexual offenses. Sample sizes for individual studies varied between

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Table 1
Descriptive Information for Studies Included in the Meta-Analysis (k = 18)

Study no.	Study	*N*	Document type	Country	Setting	Preselection	Average follow-up (in years)	Recidivism criteria	Mean age (in years)	Mean Static-99R score
1	Boer (2003)	284	Dissertation	Canada	Corrections	Routine	13.26	Convictions	41.4	2.77
2	Bonta and Yessine (2005)	128	Government or technical report	Canada	Corrections	Preselected high risk/need	5.54	Convictions	39.8	5.03
3	Haag (2005)	191	Dissertation	Canada	Corrections	Preselected high risk/need	7.00	Convictions	37.1	3.96
4	Spiranovic (2012)	822	Government or technical report	Australia	Treatment	Treatment sample	5.00	Police reports/ arrests	35.3	
5	Smallbone and Rallings (2013)	387	Journal article	Australia	Corrections	Routine	2.42	Police reports/ arrests	45.0	2.42
6	Hanson et al. (2015)	760	Journal article	Canada	Corrections	Routine	7.39	Charges	41.5	2.37
7	Lee et al. (2016)	1,535	Government or technical report	United States	Corrections	Routine	5.00	Police reports/ arrests	43.2	2.26
8	Lee, Hanson, et al. (2020)	573	Journal article	United States	Mixed/other	Routine	7.70	Convictions	38.2	2.74
8.1	Leguízamo et al. (2017)	483	Journal article	United States	Mixed/other	Routine	6.14	Convictions	39.7	1.64
9	Boccaccini et al. (2017)	34,505	Journal article	United States	Corrections	Routine	5.23	Police reports/ arrests	39.9	2.18
10	Lee, Hanson, Fullmer, et al. (2018)	343	Government or technical report	United States	Corrections	Routine	10.30	Police reports/ arrests	42.8	2.40
11	Olver, Sowden, et al. (2018)	1,044	Journal article	Canada	Treatment	Preselected high risk/need	5.00	Charges	n/a	3.92
12	Myer (2019)	986	Journal article	United States	Corrections	Routine	5.00	Charges	33.5	
13	Lee, Hanson, and Blais (2020)									
	(a) Dynamic Predictors Project sample	343	Journal article	Canada	Corrections	Preselected high risk/need	5.00	Police reports/ arrests	38.0	4.45
	(b) National Flagging System sample	259	Journal article	Canada	Corrections	Preselected high risk/need	5.00	Convictions	42.1	5.00
14	Swinburne Romine et al. (2008)	635	Conference presentation	United States	Treatment	Treatment sample	16.69	Convictions	38.2	1.69
15	Tsao & Chu (2021)	134	Journal article	Singapore	Corrections	Routine	3.70	Charges	23.5	4.78
16	Helmus, Lee, and Zabarauckas (2021)	3,605	Manuscript	Canada	Corrections	Routine	4.61	Charges	41.6	2.44
16.1	Lee, Hanson, and Zabarauckas (2018)	122	Journal article	Canada	Corrections	Routine	4.3	Charges	39.0	2.39

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

122 and 34,687 with a mean sample size of 1,150. Studies were produced/released between 2003 and 2021 (*Mdn* = 2017), reflecting the relative recency of this area of research. Ten of the documents were published journal articles, two were dissertations, four were government or technical reports, one was a conference presentation, and one was an unpublished article (although the larger project is from a published journal article; Helmus, Hanson et al., 2021). For six studies (Boer, 2003; Bonta & Yessine, 2005; Haag, 2005; Hanson et al., 2015; Lee, Hanson, & Blais, 2020; Swinburne Romine et al., 2008), at least some effect sizes were provided on the basis of the raw data set by one of the study authors. Eight study samples were from Canada, six were from the United States, two were from Australia, and one was from Singapore.

Most studies included individuals from correctional settings ($k = 12$), but three studies included those from treatment (prison/community) settings, and two samples were from mixed/unknown settings. Ten studies had samples that were classified as routine correctional/police contact samples, two had treatment samples, five had preselected high risk/need samples (including specialized high-intensity treatment), and two had mixed/other samples. Six studies used convictions as the recidivism outcome, six studies used police reports/arrests, and five studies used charges.

The average age of the samples was 38.9 years, and the average Static-99R score was in Risk Level III (Average Risk; $M = 2.97$), although averages varied substantially across studies, ranging from 1.6 (Leguízamo et al., 2017) to 5.0 (Bonta, & Yessine, 2005). The earliest year of release was 1976 and the latest was 2014. The average length of follow-up was 6.5 years, with sample-level average follow-ups ranging from 2.4 years (Smallbone & Rallings, 2013) to 16.7 years (Swinburne Romine et al., 2008).

Mean Group Differences in Static-99R Total Scores

Table 2 presents fixed-effect meta-analytic average Static-99R scores for the different groups. Generally, Indigenous and Asian individuals had the highest mean Static-99R scores (*M*s = 3.77 and 4.18, respectively), Hispanic individuals had the lowest ($M = 1.83$), and Black and White individuals had intermediate means (*M*s = 2.92 and 2.41, respectively). It is difficult to directly compare these results, however, because they come from different samples. For example, most of the studies of Indigenous individuals were from samples preselected as unusually high risk/need, and most of the studies of Black and Hispanic individuals came from relatively routine, unselected samples.

Consequently, the strongest analysis was to conduct within-study comparisons of group means. These were calculated using Cohen's *d* effect sizes to assess the magnitude of difference between a non-White racial/minority group and their White (or non-Indigenous) counterparts (see Table 3 for meta-analytic results). Indigenous and Black individuals scored significantly higher on Static-99R compared with White/non-Indigenous groups (Cohen's *d*s were small, ranging from 0.35 to 0.41, depending on the analysis and meta-analysis model). Hispanic individuals scored slightly lower on Static-99R than White individuals. This difference was significant in the fixed-effect model but not the random-effects model; in either model, the effect size was very small (*d*s = −0.09 and −0.10, respectively).

Predictive Accuracy Across Racial/Ethnic Groups

A list of the effect sizes for each racial/ethnic group for each study is available in Table S1 in the online Supplemental Materials. Table 2 presents the results of the core summary analyses. For the combined overall sample ($k = 41$, $n = 47{,}129$), Static-99R significantly predicted sexual recidivism with a moderate mean weighted effect size in both fixed-effect (AUC = .671) and random-effects (AUC = .694) analyses. Variability in accuracy across effect sizes was significant and moderate in magnitude ($I^2 = 61.3$). In other words, roughly 61% of the variability in effect sizes was beyond what would be expected on the basis of sampling error. Predictive accuracy was large for samples of non-Indigenous individuals (AUC = .732 in both models; $k = 5$, $n = 2{,}519$) and moderate for White individuals (fixed-effect model: AUC = .682, random-effects model: AUC = .705; $k = 11$, $n = 13{,}756$); this likely reflects between-study differences as opposed to meaningful differences in White samples versus almost-all-White samples. Significant variability in predictive accuracy was found among the White samples (moderate in magnitude; $I^2 = 68.0\%$) but not in the non-Indigenous sample ($I^2 = 0.0\%$).

Table 2
Predictive Accuracy of Static-99R Across Offender Racial Groups for Sexual Recidivism

	Static-99R					Fixed-effect model		Random-effects model				
Offender group	*M*	*SE*	*k*	Recidivists (*n*)	Nonrecidivists (*n*)	AUC	95% CI	AUC	95% CI	*Q*	I^2	*k*
Overall sample	2.73	0.021	34	2,293	44,836	.671	[.659, .682]	.694	[.671, .717]	103.3***	61.3	41
Indigenous	3.77	0.051	10	239	1,804	.635	[.598, .671]	.635	[.598, .671]	8.2	0	11
Non-Indigenous	2.51	0.052	5	253	2,266	.732	[.698, .765]	.732	[.698, .765]	4.3	0	5
White	2.41	0.035	9	846	12,910	.682	[.662, .702]	.705	[.662, .748]	31.2***	68.0	11
Black	2.92	0.079	4	485	10,027	.666	[.642, .691]	.762	[.641, .883]	31.3***	87.2	5
Study 14 removed				484	10,002	.652	[.627, .677]	.709	[.624, .793]	8.2*	63.5	4
Studies 14 and 9 removed				46	715	.757	[.680, .834]	.757	[.680, .834]	0.3	0	3
Hispanic	1.83	0.061	4	302	9,746	.642	[.611, .674]	.642	[.611, .674]	2.5	0	5
Asian	4.18	0.097	2	12	244	.750	[.610, .880]	.750	[.610, .880]	0.5	0	2
Other				156	7,849	.621	[.576, .666]	.621	[.576, .666]	0.01	0	2

Note. Mean Static-99R scores were computed using fixed-effect meta-analysis. AUC = area under the curve; CI = confidence interval; *SE* = standard error.
* $p < .05$. *** $p < .001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Table 3
Within-Study Static-99R Group-Mean Comparisons

Comparison	Non-White group (*n*)	Reference group (*n*)	Fixed-effect model		Random-effects model				
			d	95% CI	*d*	95% CI	*Q*	I^2	Studies
Indigenous versus White/non-Indigenous	1,920	5,903	0.407	[0.354, 0.459]	0.394	[0.271, 0.518]	33.2	72.9	1–6, 11, 13a, 13b, 16
Black versus White	787	1,673	0.370	[0.278, 0.462]	0.348	[0.192, 0.505]	6.8	56.0	7, 8, 10, 14
Hispanic versus White	1,109	1,673	−0.094	[−0.179, −0.010]	−0.099	[−0.300, 0.092]	10.3*	71.0	7, 8, 10, 14

Note. CI = confidence interval.
* $p < .05$.

Predictive accuracy for Indigenous individuals fell just below the cutoff to be considered moderate in magnitude (AUC = .635 in both models; $k = 11$, $n = 2{,}043$). AUCs were moderate-to-large for Black individuals (fixed-effect model: AUC = .666, random-effects model: AUC = .762; $k = 5$, $n = 10{,}512$) and Hispanic individuals (AUC = .642 in both models; $k = 5$, $n = 10{,}048$). Compared with Indigenous individuals, Black and Hispanic individuals were examined in half as many studies but represented five times or more the sample size; this is largely attributable to one very large study (Boccaccini et al., 2017), which found fairly similar predictive accuracy for Black, White, and Hispanic individuals. There was large and significant variability across studies in the findings for Black individuals. One outlier with a small sample size, only one recidivist, and an improbably large effect (AUC = .92) was removed (Swinburne Romine et al., 2008), which then resulted in the largest study (Boccaccini et al., 2017) also meeting criteria for an outlier. Results are presented with first one and then both studies removed (see Table 2). However, given the small number of studies overall in this group, it may be too soon to draw conclusions about outliers. Predictive accuracy was large for Asian individuals (AUC = .750 in both models), although this is based on only two studies with an overall *n* of 256 and only 12 recidivists.

Within-Study Subgroup Comparisons

Table 4 compares the predictive accuracy of Static-99R across racial/ethnic groups using within-study subgroup analyses. Here, the effect size being analyzed is a difference score for the AUCs, with 0 reflecting no difference in accuracy between groups. There were 11 comparisons available for Indigenous individuals. Six comparisons had White individuals as the reference group; the remaining comparisons used non-Indigenous as the reference group (Boer, 2003; Hanson et al., 2015; Olver, Snowden, et al., 2018; Smallbone & Rallings, 2013; Spiranovic, 2012). The average weighted difference in AUCs was roughly −.07 and was statistically significant for both models, with variability consistent with what would be expected by chance ($I^2 = 0$). In other words, within studies, on average, the AUC was .07 lower for Indigenous individuals (e.g., if the AUC for White/non-Indigenous individuals is .70, you would expect the AUC for Indigenous individuals to be .63).

The comparison of AUCs for Black versus White individuals was not statistically significant for both models, either including or excluding a potential outlier (Swinburne Romine et al., 2008). Comparing Hispanic with White individuals, however, we found that the difference in AUCs was significant in the fixed-effect model, with lower AUCs for Hispanic individuals (difference = −.0237); in the random-effects model, however, the magnitude of the difference was larger (difference = −.0484) but not statistically significant. In both models, the magnitude of difference was smaller than in the Indigenous comparisons.

Calibration

Calibration can be examined by comparing predicted recidivism rates with those from the normative data or by comparing predicted recidivism rates per group. For examining the extent to which the recidivism norms are applicable more generally across new studies, the larger meta-analysis found that Static-99R norms overestimated recidivism in most modern routine correctional samples

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Table 4
Within-Study Subgroup Analyses for Racial Groups

Comparison	Non-White group		Reference group		Fixed-effect model		Random-effects model				
	Recidivists (*n*)	Nonrecidivists (*n*)	Recidivists (*n*)	Nonrecidivists (*n*)	AUC difference	95% CI	AUC difference	95% CI	*Q*	I^2	Studies
Indigenous versus White/non-Indigenous	239	1,804	571	6,093	−.0669	[−.1073, −.0264]	−.0669	[−.1073, −.0264]	9.1	0	1–6, 11, 12, 13a, 13b, 16
Black versus White	485	10,027	528	9,083	.0040	[−.0128, .0208]	.0309	[−.0936, .1554]	76.5***	94.8	7, 8, 9, 10, 14
Removed outlier	484	10,002	446	8,568	−.0141	[−.0314, .0032]	−.0236	[−.0632, .0159]	4.6	35.3	7, 8, 9, 10
Hispanic versus White	302	9,746	528	9,083	−.0237	[−.0413, −.0060]	−.0484	[−.1169, .0201]	10.9*	63.3	7, 8, 9, 10, 14

Note. AUC = area under the curve; CI = confidence interval.
* $p < .05$. *** $p < .001$.

(Helmus et al., 2022), although limitations in estimating absolute recidivism rates are discussed elsewhere (Helmus, 2021). For the purposes of this meta-analysis, we were not concerned about whether the normative data would over- or underestimate recidivism. Instead, we wanted to know whether the predicted recidivism rates for Static-99R would differ across racial/ethnic groups within the same sample. This gives more direct information about whether the scale would systematically over- or underestimate recidivism for certain groups above and beyond the instability in recidivism norms across samples that has already been identified in the field (Helmus, 2018).

Only four studies provided information to assess calibration accuracy, with the first three examining 5-year logistic regression analyses. Lee et al. (2016) found that the predicted recidivism rate for a Static-99R score of 2 (the median score) was not significantly different for Black (2.6%), White (3.0%), or Hispanic (2.4%) individuals in California ($Q = 0.47$, $df = 2$, $p = .792$). In another (smaller) sample from California, Lee, Hanson, Fullmer, et al. (2018) similarly found no significant differences in the predicted recidivism rate for a score of 2 between Black (3.0%), White (2.4%), and Hispanic (3.9%) individuals ($Q = 0.33$, $df = 2$, $p = .847$). Spiranovic (2012) reported logistic regression results for Indigenous versus non-Indigenous individuals in Western Australia, presenting the predicted recidivism rate (in logits) for Static-99R scores of 0 and 4, allowing a meta-analysis to compare the two groups. For a Static-99R score of 0, the predicted recidivism rate for Indigenous individuals (15.4%) was higher than for non-Indigenous individuals (6.5%), and this difference approached statistical significance ($Q = 3.58$, $df = 1$, $p = .058$). For a Static-99R score of 4, the predicted recidivism rate for Indigenous individuals (24.1%) was significantly higher than for non-Indigenous individuals (11.4%; $Q = 15.58$, $df = 1$, $p < .001$). Last, Helmus, Lee, and Zabarauckas (2021) did not have sufficient cases with 5-year follow-up data to report logistic regression analyses in their large sample from British Columbia, Canada. However, in a conference presentation from that sample, Helmus and Zabarauckas (2016) reported results from a Cox regression analysis showing that Indigenous individuals did not have significantly different recidivism rates than non-Indigenous individuals, after controlling for Static-99R score (hazard ratio = 1.15, 95% CI [0.84, 1.59], $p = .374$).

Discussion

The current meta-analysis explored the total scores, discrimination, and calibration of Static-99R across different racial/ethnic groups. Black and Indigenous individuals had higher Static-99R scores on average than their White or non-Indigenous counterparts, which may be attributable to differential opportunities because of systemic racism or to increased policing and prosecution of certain groups, which would lead to lengthier criminal records. We generally found moderate-to-large predictive accuracy for Static-99R, similar to results from the larger meta-analysis of which this study is a part (Helmus et al., 2022). In within-study comparisons, no significant differences in accuracy were found between Black and White individuals. In contrast, predictive accuracy was significantly lower for Hispanic individuals (in fixed-effect but not random-effects analyses) and was consistently, significantly, and meaningfully lower for Indigenous individuals (with AUCs on

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

average being .07 lower). In contrast, the observed degradations in accuracy for Hispanic individuals were smaller.

There were limited analyses of calibration, but two studies from California suggested that predicted recidivism rates for the median Static-99R score were similar across Black, White, and Hispanic individuals. For Indigenous individuals, a large study from Canada found no significant differences in recidivism rates between Indigenous and non-Indigenous individuals after controlling for Static-99R scores, whereas a smaller study from Australia found a pattern of recidivism rates per Static-99R score that were higher for Indigenous compared with non-Indigenous individuals. These findings are limited by the small number of studies available. Additionally, existing analyses have typically examined calibration differences at one point (e.g., the median score). Calibration results may differ across the range of scores and should be examined in future research.

Black and Hispanic Individuals

It is difficult to interpret the disparity in predictive accuracy between Hispanic and White individuals. It was fairly small in the fixed-effect analyses (though statistically significant), and it was larger but nonsignificant in the random-effects analyses. Random-effects models were conceptually preferable given the goals of this study, and they accounted for between-study variability in the CIs (which is generally more realistic), but they are not particularly stable with less than 30 studies (Schulze, 2007). Only five studies were available, suggesting that the random-effects results are less likely to be reliable. In contrast, however, the fixed-effect results gave implausibly narrow CIs when there was meaningful variability across studies (Borenstein et al., 2021), which we observed in the White versus Hispanic comparisons, suggesting that the statistical significance in this model may not be fully plausible either. Unfortunately, this leaves us with the trite suggestion that more research is needed. In the meantime, we would say the results raise concerns about differential predictive accuracy for Hispanic versus White individuals, although the magnitude of differences did not appear to be large.

Interestingly, the same set of five studies was used to compare Black and Hispanic individuals with White individuals, so between-study differences cannot account for the differences in findings for the Hispanic and Black comparisons. Although the mixed findings leave us unsure about whether Static-99R has meaningfully lower accuracy for Hispanic individuals, it is unexpected and unclear why we might observe lower accuracy for Hispanic individuals but not Black individuals.

One possible explanation for the findings is higher rates of immigration and deportation among Hispanic individuals, which would both reduce the quality of information used in scoring the risk scale (e.g., criminal records may not be accessible) and the ability to detect recidivism. Indeed, Leguízamo et al. (2017) found evidence that Static-99R was less accurate for Latino individuals born outside the United States, pointing to the role of good-quality information for immigrants. In contrast, however, in a much larger study, Boccaccini et al. (2017) did not replicate those findings. Ideally, good-quality Static-99R assessments should be based on full and accurate criminal history information, with follow-up data including fairly comprehensive sources of recidivism information. To the extent that studies are not able to identify and remove individuals who have been deported, accuracy should be degraded. The impact of this issue could be large; in Lee et al.'s (2018) study from California, 40% of the Latino individuals in their sample were deported after release, compared with only 2% of Black and White individuals. After the deportees from the calibration analyses were excluded, the initial finding of lower predictive accuracy in Hispanic individuals (from Hanson et al.'s, 2014, study) vanished. In other words, the observed low sexual recidivism rates among Hispanic individuals could largely be attributed to the high rates of deportation.

Varela et al. (2013) provide other possible explanations for the possible differential accuracy of Static-99R among Hispanic individuals. They point to cultural differences (e.g., higher traditional family values, greater rape myth acceptance) that could impact the meaning and predictiveness of traditional risk factors (e.g., unrelated or stranger victim) as well as the likelihood of victims reporting offenses, reducing the reliability of recidivism information. They also point out that higher rates of immigration among Hispanic populations could reduce trust in the authorities and decrease the chances that sexual offenses are reported, particularly if the victim fears deportation.

Indigenous Individuals

In some ways, the finding that Static-99R has less predictive accuracy among Indigenous individuals is not surprising; it is consistent with numerous other studies finding lower accuracy of risk factors and risk scales for Indigenous individuals (Gutierrez et al., 2013, 2016; Helmus, Lee, & Zabarauckas, 2021; Helmus & Forrester, 2014b; Perley-Robertson et al., 2019; Wilson & Gutierrez, 2014). Exceptions to this pattern include the Psychopathy Checklist–Revised (although it is not a risk assessment scale; Olver, Neumann, et al., 2018) and the Violence Risk Scale-Sexual Offender version (VRS-SO; Olver, Sowden, et al., 2018). On the other hand, this is a new finding for Static-99R: two previous (smaller) meta-analyses did not find a significant difference (Babchishin et al., 2012; Lee, Hanson, & Blais, 2020). The improved power and comprehensiveness of this meta-analysis suggest a clear and meaningful pattern of lower accuracy for Indigenous individuals.

Numerous studies and meta-analyses have found that justice-involved Indigenous individuals have higher recidivism rates and score higher on numerous risk factors, particularly those related to criminal history (Gutierrez et al., 2013; Perley-Robertson et al., 2019; Shepherd et al., 2014). As discussed in the introduction, much of this can be attributed to the harmful effects of colonization, which have systematically exposed Indigenous individuals to social and economic disadvantage. The risk factors that the literature highlights as most persistent for Indigenous individuals include a history of residential school attendance, poverty, domestic violence, physical or sexual abuse, substance abuse, homelessness, mental health issues, and lack of education/employment (Wilk et al., 2017). One study in Aoteoroa (New Zealand) found evidence for both higher involvement in crime among Māori as well as higher policing and prosecution rates (Fergusson et al., 2003), indicating that both disadvantage and ongoing systemic discrimination contribute to overrepresentation.

As noted by Warne et al. (2014), however, racial/ethnic group differences in scores on an assessment measure are not a form of test bias but rather an indication of inequality in society. Once someone is already in the criminal legal system and case management decisions need to be made, risk assessment is an essential task,

and it is of primary importance to examine differential predictive accuracy (i.e., whether risk scales make different predictions across groups). Unfortunately, differential accuracy was found for Indigenous individuals, which is an indicator of test bias. This raises the important question of why the scale would be less accurate for Indigenous groups.

Wilson and Gutierrez (2014) offered four possible explanations for differences in discrimination and calibration for the LSI family of scales in their meta-analysis. First, overpolicing and overprosecution may inflate detected crime and recidivism rates for Indigenous individuals. This could mean, for example, that the same number of prior convictions may not be measuring antisociality for Indigenous people in the same way as for White people. The second explanation is that higher rates of risk factors overall may be muddying the ability to detect predictive accuracy of any risk factor in particular. This would be more of an issue for risk factors than risk scales. Or, as applied to risk scales, it could suggest a restriction-of-range effect. Third, existing risk scales may not be operationalizing the risk factors in a manner that is culturally relevant to Indigenous peoples. This is more likely to be the case for dynamic risk factors (e.g., family/marital variables) as opposed to static factors. Last, current risk scales may be omitting culturally specific risk factors.

There are other possible reasons for these findings that are not mentioned by Wilson and Gutierrez (2014). Non-White victims (who are most likely to be victimized by non-White perpetrators) may be less likely to report offenses to police because of distrust in the system or previous negative experiences, which would reduce the quality of official recidivism data, which is already a flawed proxy of criminal behavior. Additionally, given the elevated risk factors in Indigenous communities cited above, it is possible that Indigenous samples exhibit a restriction-of-range on unmeasured risk factors, which could attenuate the accuracy of Static-99R, akin to the finding of lower accuracy among samples preselected as unusually high risk/need (Hanson et al., 2016). It is likely that several or all of these factors contribute to the observed differential prediction. Further research is needed to better understand the contribution of these potential explanations and others as well (discussed further below).

Additionally, many of the explanations above and the disadvantages faced by Indigenous peoples would seemingly be applicable to Black and Hispanic individuals, but similar disparities in predictive accuracy were not observed. The current meta-analysis suggests that there is something unique about justice-involved Indigenous people in terms of risk assessment applicability that may be distinct from other groups known to be negatively impacted by systemic racism. Further research to understand the disparities in accuracy for Indigenous peoples may also inform why we did not observe similarly strong degradations in accuracy for other groups.

Strengths and Limitations

Knowledge is cumulative and replication is essential. The primary strength of the present study is that it is a meta-analysis and provides a comprehensive summary of the research to date on the applicability of Static-99R across different racial/ethnic groups. Especially for several Canadian studies of Indigenous individuals, statistical power is often low for subgroups. Aggregating the results allows higher statistical power, greater confidence in the findings, and the ability to quantify variability in findings across studies.

Meta-analyses, however, are limited by the quality and quantity of the available research studies. There is much that is unknown or was inadequately measured in the present study. Importantly, we have virtually no idea how the authors of the individual studies defined race/ethnicity. It could have been defined on the basis of self-report information, census data, criminal justice system staff presumptions (e.g., skin color or last name), citizenship, country of birth, or primary language. These definitions differ in validity, and not only are we unable to examine the definitions but also there is insufficient research to compare various definitions. Studies were also limited in that they all used official records to examine recidivism. This necessarily excludes unreported offenses. Given that most sex offenses are perpetrated against same-race victims (Bureau of Justice Statistics, 2011, as cited in Varela et al., 2013), if race/ethnicity impacts reporting rates, this limitation may differentially affect predictive accuracy across groups.

There was also a considerable restriction-of-range in the racial/ethnic groups examined in the current meta-analysis. Almost all studies were of racial/ethnic subgroups within Canada or the United States. Two studies were from Australia, and one was from Singapore. Apart from the latter, these are fairly Western, educated, industrialized, rich, and democratic (WEIRD) countries and, more specifically, former British colonies with common law systems based on British jurisprudence. Only one study was available from Asia, and none are available from Central or South America, Africa, or the Middle East.

Additionally, this study focused narrowly on race/ethnicity and did not examine other cultural distinctions (e.g., between European nations, within linguistic groups such as Francophones vs. Anglophones). We also grouped racial/ethnic groups quite broadly, missing important distinctions. For example, Indigenous peoples were combined for all studies in this meta-analysis, including studies from Canada, Australia, and the United States. There are obviously important differences in Indigenous language, culture, and experiences both between and within the countries examined. However, it is unknown whether those differences are risk relevant. In fact, despite this lack of differentiation, the variability in accuracy of Static-99R was not significant across the Indigenous samples, nor was the observed differential accuracy. This suggests that in terms of the issue of predictive accuracy, specific variations in Indigenous culture may be less risk relevant than the common harms and consequences of colonization that have been shared by Indigenous groups across many former British colonies. Nonetheless, future research (particularly more studies from Australia and the United States and any research from Aotearoa/New Zealand) would be helpful to better explore variations between Indigenous cultures and experiences.

Another limitation is that this meta-analysis did not have adequate statistical power for meaningful exploration of moderators across some racial/ethnic groups. However, this limitation is mitigated given the low amounts of observed variability in results that need to be explained. There was significant variability in accuracy only for the subgroups of Black and White individuals. For White individuals, who comprise the majority in most validation studies, larger meta-analyses of Static-99R (which include the current studies) have already identified interesting moderators of predictive accuracy, including training, referencing the coding manual, and preselection of the sample (Helmus, Hanson, et al., 2021; Helmus et al., 2022). The analysis of Static-99R for Black individuals found

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

significant and large variability across the five studies examined. The results from Swinburne Romine et al. (2008) were based on a small sample size and only one recidivist; consequently, they are not expected to be stable (AUC = .92). Additionally, two studies were from California (Lee, Hanson, Fullmer, et al., 2018; Lee et al., 2016) and one was from Texas (Boccaccini et al., 2017); these three were all field implementations of the scale, as scored by parole and probation officers as part of routine practice. Interestingly, among field studies of Static-99R, California has tended to observe the highest predictive accuracy and Texas the lowest. One major hypothesis for this discrepancy relates to the substantive training and implementation differences across these two states (Helmus, 2015).

One factor not addressed in the current meta-analysis is differential impact of risk assessment scales. This is not a form of test bias but is nonetheless an important consideration (Warne et al., 2014). For example, even if a risk scale is equally predictive across groups, if one group has more risk factors and higher risk scores, then using that risk scale (even if it is empirically valid) has implications across groups. For example, higher risk groups are likely to require more intensive treatment, which may slow down progress toward parole readiness (and may also reduce likelihood of being granted parole). Consequently, some groups may face disadvantages even if the most accurate risk assessment methods are being used. This is a policy issue that needs to be addressed to remedy disparities in the criminal legal system (fast-tracking treatment options for certain groups, better educating parole board members about these issues).

Implications for Future Research

The limited number of studies, as well as the limitations and omissions in the meta-analysis as cited above, all point to the need for further and better-quality research. Research needs to be more explicit and nuanced about how race/ethnicity is defined and measured. Where sample sizes are available, analyses of subgroups are warranted (e.g., examining differences between Indigenous groups). More research is needed in non-WEIRD countries. Additionally, more research is needed on calibration, particularly patterns of calibration across different risk levels.

In particular, there is a need for research to improve risk assessment options for Indigenous individuals. Existing risk scales could be improved with research more clearly identifying which items are more robustly predictive for Indigenous individuals and which items may need to be omitted. For example, criminal history is normally one of the most powerful risk domains (Bonta & Andrews, 2017) but tends to show the largest differential accuracy (Gutierrez et al., 2013). However, Perley-Robertson et al. (2019) found that roughly one-third of criminal history items had equivalent accuracy for Indigenous and non-Indigenous individuals, suggesting that it is possible to construct a criminal history subscale that would demonstrate equivalent accuracy. Other risk constructs could be individually and thoroughly explored to identify and compare more culturally responsive operationalizations of those constructs. Such an approach would benefit from both quantitative research as well as qualitative research and consultations to identify new avenues to understand and measure known risk factors. Over time, risk scales could be modified with coding rules developed from this stream of research.

Alternately, given the insufficient amount of research on the possibility of culturally specific risk factors (for a review and discussion, see, e.g., Gutierrez, 2018), we should not rule out the necessity of separate risk tools for Indigenous individuals. Further research is needed on potentially culturally specific or culturally salient risk or protective factors such as cultural embeddedness (Fox et al., 2018), participation in cultural activities/ceremonies, foster care experiences, personal or familial residential or boarding school experiences, or experiences of racism/discrimination. Whether the path forward involves improving or replacing existing risk scales, as part of efforts to decolonize research and practice, such advances should involve consultation and collaboration with Indigenous scholars and communities (Hanson, 2020).

Implications for Practice

The results of the current research support the use of Static-99R for Black, White, and Asian people charged with or convicted of sexually motivated offenses, although more research on Black and Asian individuals would be helpful. Combined predictive accuracy was quite high for the two studies of Asian individuals (AUC = .750), but neither had a particularly large sample size. Additionally, an earlier study of Static-99 among Japanese men who had raped children demonstrated much more modest accuracy. Consequently, it is possible that further validation studies of Asian individuals could result in lower aggregate effects.

There was some evidence for lower predictive accuracy for Hispanic individuals; however, more research is needed to provide reliable recidivism information, including deportation data, for this group. In the meantime, given that the scale was significantly predictive for this group and the disparity was not large (and the absence of alternatives that have well-demonstrated cross-cultural applicability), we believe that evaluators can continue to use the scale with Hispanic individuals. This recommendation may change, however, as more research becomes available.

There was clear evidence of significant and meaningfully lower accuracy when the scale was used for Indigenous individuals. What does this mean for evaluators conducting risk assessments for Indigenous peoples? Unfortunately, the answer is not simple. In short, our current summary is that Static-99R is better than nothing, but it is not good enough. The reduced predictive accuracy with Indigenous individuals is problematic and should be recognized when conducting assessments with this population; evaluators should be encouraged to exercise greater caution. In particular, evaluators should be aware of issues uniquely and disproportionately facing justice-involved Indigenous individuals and exercise cultural humility in their assessments. Given the severe overrepresentation of Indigenous individuals, it may be prudent and defensible to veer toward lower risk assessments or less severe/restrictive recommendations when there is ambiguity or for individuals at the lower end of a risk category.

We do not support abandoning structured risk assessment scales with Indigenous individuals given the troubling differential predictive accuracy observed. We agree with Olver, Sowden, et al. (2018) that there needs to be a consideration of potential harms of both using or not using structured risk scales with Indigenous individuals. Although Static-99R is problematically demonstrating lower accuracy, we believe the harms of not using structured risk scales are greater. People tend to overestimate risk, especially when the risk is for something scary or rare (Kahneman, 2011; Mills et al., 2011). Additionally, the only study we are aware of that has examined a

Structured Professional Judgment (SPJ) risk scale with Indigenous individuals found that professional ratings of risk magnified score differences compared with total scores (in the direction of staff being more likely to rate Indigenous individuals as high risk compared with non-Indigenous individuals who have the same number of risk factors), and predictive accuracy was higher for total scores compared with the SPJ risk rating (Helmus & Forrester, 2014a, 2014b).

In an ideal world, evaluators would be unbiased and culturally sensitive, and therefore their judgment of risk could meaningfully incorporate risk-relevant differences across racial/ethnic groups. However, the current literature suggests that this is not the case. Consequently, not using risk assessment scales is likely to be even more harmful to Indigenous peoples than using scales with differential accuracy. In fact, even scales with differential accuracy may be protective of decisions biased against Indigenous individuals; Wilson and Gutierrez (2014) found that the LSI may underpredict detected recidivism for low risk Indigenous peoples, so the scale may be particularly useful in counteracting potential biases from overpolicing of this population. Similarly, Spiranovic (2012) found higher recidivism rates for Indigenous Australians compared with non-Indigenous Australians with the same Static-99R risk score. This suggests that Static-99R may be more likely to underestimate recidivism for Indigenous individuals, who may have a higher density of unmeasured risk factors (or greater prejudice in the criminal justice system, inflating recidivism rates). Consequently, use of existing actuarial measures and their norms may be protective and result in more lenient decisions compared with developing Indigenous-specific normative data or scales.

Conversely, however, although current scales such as Static-99R may be better than nothing for Indigenous individuals, this does not mean we should be complacent in continuing to use them. We must not lose sight of the other side of the coin: They are not good enough. They should not be abandoned without suitable alternatives, but it is necessary to do the research on these alternatives. This could include modifying the coding rules, altering the scale itself, or developing an entirely new scale for Indigenous individuals (either a scale unique to this group or one i.e., more culturally universal).

For sexual recidivism risk assessment, the VRS-SO (Olver et al., 2007) may be a preferable alternative, although further research is needed. Roughly one-third of the development sample for that scale was Indigenous, and research supports equivalent predictive accuracy for Indigenous and non-Indigenous groups based on three Canadian samples and one Aotearoa/New Zealand sample (Olver, Sowden, et al., 2018). That being said, however, there is less research with Indigenous individuals on the VRS-SO compared with Static-99R, and it is possible that future research may uncover differential accuracy as well. Although group differences in predictive accuracy were not statistically significant, they were similar or only slightly smaller for the VRS-SO (total posttreatment scores at 5- and 10-year follow-ups had AUC differences of .07 and .05, which is similar to the .07 difference observed in the current meta-analysis).

One key difference between VRS-SO and Static-99R is that the VRS-SO examines both static and dynamic risk factors. This has advantages and disadvantages. The VRS-SO is more comprehensive, and dynamic risk factors are necessary to identify treatment targets and to measure change. Conversely, however, it is more time consuming to administer, and it is not necessarily feasible to triage large correctional populations, so there are some contexts where a simple static scale is still needed. In terms of relative comparisons, one large study with 910 Indigenous individuals found that combining Static-99R with STABLE-2007 produced a Harrell's *C* value of .65 (this can be interpreted in the same way as AUCs, but Harrell's *C* accounts for varying follow-up times; Helmus, Lee, & Zabarauckas, 2021). In contrast, Olver, Sowden, et al. (2018) found AUCs of .68 and .69 for the VRS-SO after 5- and 10-year follow-up periods. This suggests that the VRS-SO has a fairly small advantage, but if it does have comparable accuracy for Indigenous and non-Indigenous individuals, it may be a practical alternative. It is difficult to compare the scales on different samples, however. More research is needed overall, particularly on within-sample comparisons.

Conclusion

Risk assessments are pervasive in the criminal legal system and are highly consequential for civil liberties and public safety. Consequently, it is paramount to advance evidence-based practice so that risk scales are as reliable, accurate, transparent, and unbiased as possible. Systemic racism is apparent in numerous countries and has particularly devastating consequences at every stage of the criminal legal system. Exploring the cross-cultural validity of risk assessment scales is a high priority for research. In the current meta-analysis, we found support for Static-99R for sexual recidivism risk assessment with Black and White individuals. The scale significantly predicted sexual recidivism for Hispanic and Indigenous people but with some evidence of lower accuracy for Hispanic people and clear evidence of lower accuracy for Indigenous people. More validation studies are still needed, particularly with other groups such as Asians. Also, Static-99R is unique because there are sufficient studies available to meta-analyze. Further research on other risk scales is needed; cross-cultural validity should not simply be presumed. This work should be contextualized within broader theoretical and empirical efforts to understand how culture and race/ethnicity impact risk and perceptions of risk. In doing so, it is imperative that we acknowledge how colonial racist policies and practices intersect at various points of the criminal legal system.

References

References marked with an asterisk indicate studies included in the meta-analysis.

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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

American Psychological Association. (2021). *Apology to people of color for APA's role in promoting, perpetuating, and failing to challenge racism, racial discrimination, and human hierarchy in U.S.* https://www.apa.org/about/policy/resolution-racism-apology.pdf

Ashford, L. J., Spivak, B. L., & Shepherd, S. M. (2022). Racial fairness in violence risk instruments: A review of the literature. *Psychology, Crime & Law*, *28*(9), 911–941. https://doi.org/10.1080/1068316X.2021.1972108

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.

Australian Bureau of Statistics. (2021, September 16). *Corrective services, Australia, June quarter 2021*. https://www.abs.gov.au/statistics/people/crime-and-justice/corrective-services-australia/latest-release

Babchishin, K. M., Blais, J., & Helmus, L. (2012). Do static risk factors predict differently for Aboriginal sex offenders? A multi-site comparison

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

using the original and revised Static-99 and Static-2002 scales. *Canadian Journal of Criminology and Criminal Justice*, *54*(1), 1–43. https://doi.org/10.3138/cjccj.2010.E.40

Babchishin, K. M., & Helmus, L.-M. (2016). The influence of base rates on correlations: An evaluation of proposed alternative effect sizes with real-world data. *Behavior Research Methods*, *48*(3), 1021–1031. https://doi.org/10.3758/s13428-015-0627-7

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

*Boer, A. (2003). *Evaluating the Static-99 and Static-2002 risk scales using Canadian sexual offenders* [Unpublished master's thesis]. University of Leicester.

Bolger, P. C. (2015). Just following orders: A meta-analysis of the correlates of American police officer use of force decisions. *American Journal of Criminal Justice*, *40*(3), 466–492. https://doi.org/10.1007/s12103-014-9278-y

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

*Bonta, J., & Yessine, A. K. (2005). *The national flagging system: Identifying and responding to high-risk, violent offenders* (User Report 2005-04). Public Safety and Emergency Preparedness Canada.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley. https://doi.org/10.1002/9781119558378

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., Lee, S. C., Babchishin, K. M., Hanson, R. K., & Harris, J. R. (2017). [Unpublished raw data on recidivism for 409 individuals from the Dynamic Predictors Project (1998)].

Brankley, A. E., & Lee, S. C. (2016, November). *The utility of Static-99R and STABLE-2007 across different ethnic groups: A prospective field study* [Paper presentation]. Association for the Treatment of Sexual Abusers 35th Annual Research and Treatment Conference, Orlando, FL, United States.

Bureau of Justice Statistics. (2011). *Race of incarcerated males by state* [Unpublished raw data].

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

Day, A., Tamatea, A. J., Casey, S., & Geia, L. (2018). Assessing violence risk with Aboriginal and Torres Strait Islander offenders: Considerations for forensic practice. *Psychiatry, Psychology, and Law*, *25*(3), 452–464. https://doi.org/10.1080/13218719.2018.1467804

Department of Justice Canada. (2017). *Spotlight on Gladue: Challenges, experiences, and possibilities in Canada's criminal justice system*. https://www.justice.gc.ca/eng/rp-pr/jr/gladue/toc-tdm.html

DiAngelo, R. (2018). *White fragility: Why it's so hard for White people to talk about racism*. Beacon Press.

Ewert v. Canada, SCC 30, 2 S.C.R. 165 (2018). https://www.canlii.org/en/ca/scc/doc/2018/2018scc30/2018scc30.html

Fergusson, D. M., Horwood, L. J., & Swain-Campbell, N. (2003). Ethnicity and criminal convictions: Results of a 21-year longitudinal study. *Australian and New Zealand Journal of Criminology*, *36*(3), 354–367. https://doi.org/10.1375/acri.36.3.354

Flanagin, A., Frey, T., Christiansen, S. L., & Bauchner, H. (2021). The reporting of race and ethnicity in medical and science journals. *JAMA*, *325*(11), 1049–1052. https://doi.org/10.1001/jama.2021.2104

Fox, R., Neha, T., & Jose, P. E. (2018). Tū Māori Mai: Māori cultural embeddedness improves adaptive coping and wellbeing for Māori adolescents. *New Zealand Journal of Psychology*, *47*(2), 13–23. https://www.psychology.org.nz/journal-archive/Cultural-Emeddeness-Improves-Coping-and-Wellbeing.pdf

Gutierrez, L., Helmus, L. M., & Hanson, R. K. (2016). What we know and don't know about risk assessment with offenders of Indigenous heritage. *Journal of Threat Assessment and Management*, *3*(2), 97–106. https://doi.org/10.1037/tam0000064

Gutierrez, L., Wilson, H. A., Rugge, T., & Bonta, J. (2013). The prediction of recidivism with Aboriginal offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology and Criminal Justice*, *55*(1), 55–99. https://doi.org/10.3138/cjccj.2011.E.51

Gutierrez, L. L. (2018). *Walls of Red Wing: An examination of culturally-informed sentencing, risk/need factors, and treatment for peoples of Indigenous heritage in Canada's criminal justice system* [Unpublished doctoral dissertation]. Carleton University.

*Haag, A. M. (2005). Do psychological interventions impact on actuarial measures: An analysis of the predictive validity of the Static-99 and Static-2002 on a re-conviction measure of sexual recidivism. *Dissertations Abstracts International*, *66*(8–B), Article 4531. https://doi.org/10.11575/PRISM/19654

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

Hanson, R. K. (2020). *Affidavit: R. v. Monsieur Joe Kritik (Salowatseak)*. Cour du Quebec, Court File No. 635-01-014601-166

Hanson, R. K. (2022). *Prediction statistics for psychological assessment*. American Psychological Association. https://doi.org/10.1037/0000275-000

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

*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., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management*, *1*(2), 102–117. 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., & 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., Helmus, L. M., & Babchishin, K. M. (2016). What sexual recidivism rates are associated with Static-99R and Static-2002R scores? *Sexual Abuse*, *28*(3), 218–252. https://doi.org/10.1177/1079063215574710

Helmus, L., & Forrester, T. (2014b). *The Static Factors Assessment (SFA) of the offender intake assessment process: Relationship to release and community outcomes* (Research Report No R-339). Correctional Service of Canada.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Helmus, L., & Forrester, T. K. (2014a). *Construct validity of the Static Factors Assessment in the offender intake assessment process*. Research Branch, Correctional Service of Canada.

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

Helmus, L. M. (2015). *What makes a "good" risk assessment? A note on the importance of quality control* [post on blog for Sexual Abuse: A Journal of Research and Treatment]. http://sajrt.blogspot.ca/2015/11/what-makes-good-risk-assessment-note-on.html

Helmus, L. M. (2018). Sex offender risk assessment: Where are we and where are we going*? Current Psychiatry Reports*, *20*(6), Article 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*, Article 1686. 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., 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*, *33*(7), 581–595. https://doi.org/10.1037/pas0001010

Helmus, L. M., Kelley, S. M., Frazier, A., Fernandez, Y. M., Lee, S. C., Rettenberger, M., & Boccaccini, M. T. (2022). Static-99R: Strengths, limitations, predictive accuracy meta-analysis, and legal admissibility review. *Psychology, Public Policy, and Law*, *28*(3), 307–331. https://doi.org/10.1037/law0000351

Helmus, L. M., Lee, S. C., & Zabarauckas, C. L. (2021). *Do Static-99R and STABLE-2007 work with Indigenous people charged or convicted of sexual offences? A prospective field validity study* [Unpublished manuscript]. Simon Fraser University.

Helmus, L. M., & Zabarauckas, C. (2016, November). *The utility of Static-99R and STABLE-2007 for Aboriginal sex offenders: A field validity study* [Paper presentation]. Association for the Treatment of Sexual Abusers 35th Annual Research and Treatment Conference, Orlando, FL, United States.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *The BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hu, C., & Esthappan, S. (2017). *Asian Americans and Pacific Islanders, a missing minority in criminal justice data*. Urban Wire: Crime and Justice. https://www.urban.org/urban-wire/asian-americans-and-pacific-islanders-missing-minority-criminal-justice-data

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

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

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse*, *16*(2), 107–120. https://doi.org/10.1177/107906320401600202

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., & 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

*Lee, S. C., Hanson, R. K., Calkins, C., & Jeglic, E. (2020). Paraphilia and antisociality: Motivations for sexual offending may differ for American Whites and Blacks. *Sexual Abuse*, *32*(3), 335–365. https://doi.org/10.1177/1079063219828779

*Lee, S. C., Hanson, R. K., Fullmer, N., Neeley, J., & Ramos, K. (2018). *The predictive validity of Static-99R over 10 years for sexual offenders in California: 2018 update*. State Authorized Risk Assessment Tools for Sex Offenders. http://saratso.org/pdf/Lee_Hanson_Fullmer_Neeley_Ramos_2018_The_Predictive_Validity_of_S_.pdf

*Lee, S. C., Hanson, R. K., & Zabarauckas, C. L. (2018). Sex offenders of East Asian heritage resemble other Canadian sex offenders. *Asian Journal of Criminology*, *13*(1), 1–15. https://doi.org/10.1007/s11417-017-9252-y

Lee, S. C., Mularczyk, K., Babchishin, K. M., Blais, J., & Bonta, J. (2018). [Unpublished raw data on recidivism for 360 individuals from the National Flagging System Project (2015)].

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

*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. https://doi.org/10.1177/1079063215618377

Ley, P. (1972). *Quantitative aspects of psychological assessment: An introduction*. Duckworth.

McKenzie, K. J., & Crowcroft, N. S. (1994). Race, ethnicity, culture, and science. *The BMJ*, *309*(6950), 286–287. https://doi.org/10.1136/bmj.309.6950.286

Mills, J. F., Kroner, D. G., & Morgan, R. D. (2011). *Clinician's guide to violence risk assessment*. Guilford Press.

Ministry of Justice. (2013). *Statistics on race and the criminal justice system 2012: A Ministry of Justice publication of the Criminal Justice Act 1991*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/269399/Race-and-cjs-2012.pdf

Monchalin, L. (2016). *The colonial problem: An Indigenous perspective on crime and injustice in Canada*. University of Toronto Press.

*Myer, A. J. (2019). Examining the predictive validity of the Static-99R on Native American sex offenders. *Justice Evaluation Journal*, *2*(2), 181–195. https://doi.org/10.1080/24751979.2019.1636614

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry. *Criminal Justice and Behavior*, *41*(12), 1406–1421. https://doi.org/10.1177/0093854814548449

Nellis, A. (2016). *The color of justice: Racial and ethnic disparity in state prisons*. The Sentencing Project. https://www.sentencingproject.org/publications/color-of-justice-racial-and-ethnic-disparity-in-state-prisons/

Nettelbeck, A., & Smandych, R. (2010). Policing Indigenous peoples on two colonial frontiers: Australia's mounted police and Canada's north-west mounted police. *Australian and New Zealand Journal of Criminology*, *43*(2), 356–375. https://doi.org/10.1375/acri.43.2.356

Olver, M. E., Neumann, C. S., Sewall, L. A., Lewis, K., Hare, R. D., & Wong, S. C. P. (2018). A comprehensive examination of the psychometric properties of the Hare Psychopathy Checklist-Revised in a Canadian multisite sample of indigenous and non-indigenous offenders. *Psychological Assessment*, *30*(6), 779–792. https://doi.org/10.1037/pas0000533

*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 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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*(1), 156–176. https://doi.org/10.1037/a0035080

Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment*, *19*(3), 318–329. https://doi.org/10.1037/1040-3590.19.3.318

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*.

Perley-Robertson, B., Helmus, L. M., & Forth, A. (2019). Predictive accuracy of static risk factors for Canadian Indigenous offenders compared to non-Indigenous offenders: Implications for risk assessment scales. *Psychology, Crime & Law*, *25*(3), 248–278. https://doi.org/10.1080/1068316X.2018.1519827

Perry, B. (2006). Nobody trusts them! Under- and over-policing Native American communities. *Critical Criminology*, *14*(4), 411–444. https://doi.org/10.1007/s10612-006-9007-z

Public Safety Canada. (2020). *2019 corrections and conditional release statistical overview*. https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ccrso-2019/index-en.aspx#c2

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

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

Shepherd, S. M., Adams, Y., McEntyre, E., & Walker, R. (2014). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law*, *20*(3), 281–293. https://doi.org/10.1037/law0000017

Shepherd, S. M., & Ilalio, T. (2016). Maori and Pacific Islander overrepresentation in the Australian criminal justice system—What are the determinants? *Journal of Offender Rehabilitation*, *55*(2), 113–128. https://doi.org/10.1080/10509674.2015.1124959

Shepherd, S. M., & Lewis-Fernandez, R. (2016). Forensic risk assessment and cultural diversity: Contemporary challenges and future directions. *Psychology, Public Policy, and Law*, *22*(4), 427–438. https://doi.org/10.1037/law0000102

Shepherd, S. M., & Willis-Esqueda, C. (2018). Indigenous perspectives on violence risk assessment: A thematic analysis. *Punishment & Society*, *20*(5), 599–627. https://doi.org/10.1177/1462474517721485

*Smallbone, S., & Rallings, M. (2013). Short-term predictive validity of the static-99 and static-99-R for indigenous and nonindigenous Australian sexual offenders. *Sexual Abuse*, *25*(3), 302–316. https://doi.org/10.1177/1079063212472937

Spiranovic, C. (2012). *The Static-99 and Static-99-R norms project: Developing norms based on Western Australian sex offenders*. University of Western Australia.

*Swinburne Romine, R., Dwyer, S. M., Mathiowetz, C., & Thomas, M. (2008, October). *Thirty years of sex offender specific treatment: A follow-up study* [Poster presentation]. Association for the Treatment of Sexual Abusers Conference, Atlanta, GA, United States.

Tamatea, A. J. (2017). Culture is our business: Issues and challenges for forensic and correctional psychologists. *Australian Journal of Forensic Sciences*, *49*(5), 564–578. https://doi.org/10.1080/00450618.2016.1237549

Travis, J., Western, B., & Redburn, F. S. (2014). *The growth of incarceration in the United States: Exploring causes and consequences*. Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. National Academies Press.

Truth and 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.

*Tsao, I. T., & Chu, C. M. (2021). An exploratory study of recidivism risk assessment instruments for individuals convicted of sexual offenses in Singapore. *Sexual Abuse*, *33*(2), 157–175. https://doi.org/10.1177/1079063219884575

Varela, J. G., Boccaccini, M. T., Murrie, D. C., Caperton, J. D., & Gonzalez, E., Jr. (2013). Do the Static-99 and Static-99R perform similarly for White, Black, and Latino sexual offenders? *International Journal of Forensic Mental Health*, *12*(4), 231–243. https://doi.org/10.1080/14999013.2013.846950

Warne, R. T., Yoon, M., & Price, C. J. (2014). Exploring the various interpretations of "test bias." *Cultural Diversity & Ethnic Minority Psychology*, *20*(4), 570–582. https://doi.org/10.1037/a0036503

Watanabe, K., Yokota, K., Yoshimoto, K., Ihara, N., & Fujita, G. (2007). *Recidivism in child rapists: Identifying high risk factors* [Unpublished manuscript]. National Research Institute of Police Science.

Wilk, P., Maltby, A., & Cooke, M. (2017). Residential schools and the effects on Indigenous health and well-being in Canada—a scoping review. *Public Health Reviews*, *38*(1), Article 8. https://doi.org/10.1186/s40985-017-0055-6

Wilson, H. A., & Gutierrez, L. (2014). Does one size fit all?: A meta-analysis examining the predictive ability of the Level of Service Inventory (LSI) with Aboriginal offenders. *Criminal Justice and Behavior*, *41*(2), 196–219. https://doi.org/10.1177/0093854813500958

Received April 22, 2022
Revision received September 18, 2022
Accepted November 11, 2022 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
Content may be shared at no cost, but any requests to reuse this content in part or whole must go through the American Psychological Association.