An Examination of the Interrater Reliability Between Practitioners and Researchers on the Static-99

International Journal of
Offender Therapy and
Comparative Criminology
2014, Vol. 58(11) 1364–1375

DOI: 10.1177/0306624X13495504
ijo.sagepub.com
SAGE

Stephen P. Quesada[1], Cynthia Calkins[1], and Elizabeth L. Jeglic[1]

Abstract

Many studies have validated the psychometric properties of the Static-99, the most widely used measure of sexual offender recidivism risk. However much of this research relied on instrument coding completed by well-trained researchers. This study is the first to examine the interrater reliability (IRR) of the Static-99 between practitioners in the field and researchers. Using archival data from a sample of 1,973 formerly incarcerated sex offenders, field raters' scores on the Static-99 were compared with those of researchers. Overall, clinicians and researchers had excellent IRR on Static-99 total scores, with IRR coefficients ranging from "substantial" to "outstanding" for the individual 10 items of the scale. The most common causes of discrepancies were coding manual errors, followed by item subjectivity, inaccurate item scoring, and calculation errors. These results offer important data with regard to the frequency and perceived nature of scoring errors.

Keywords

Static-99, interrater reliability, sex offender, risk, sexual offense

An Examination of the Interrater Reliability (IRR) Between Practitioners and Researchers on the Static-99

Rising public concern over sexual recidivism has helped to fuel the development, and continued refinement, of risk assessment tools. These tools, which aim to identify sex offenders who pose the greatest risk to the community, are regularly used to assist in

[1]John Jay College of Criminal Justice, Mastic, NY, USA

Corresponding Author:
Stephen P. Quesada, John Jay College of Criminal Justice, 123 Monroe Street, Mastic, NY 11950, USA.
Email: stephen.quesada@gmail.com

decision making regarding sentencing, treatment, parole, and sexually violent predator (SVP) classifications (Archer, Buffington-Vollum, Stredny, & Handel, 2006). The outcomes of such assessments can have significant impact on the offenders and their communities. Indeed, given that the civil liberties of the offender and the safety of our communities may hinge, at least in part, on the results from risk assessment measures, it is vital that the psychometric properties of these instruments be well understood. This may be particularly important in light of the economic downturn and subsequent limits on public funds available to manage offenders in the community, as fewer resources dictate that even more careful decisions are made about the allocation of treatment or risk management services.

The Static-99 (Hanson & Thornton, 2000), an actuarial scale that includes 10 items pertaining to past criminal history and victim characteristics, is currently the most widely used sexual offender risk assessment tool in the world (Static-99.org). In meta-analyses designed to identify factors associated with sexual reoffense, the 10 items included in the scale were found to be the most robust predictors of sexual recidivism (Hanson & Bussiere, 1998; Hanson & Thornton, 1999; 2000). Its psychometric properties have been studied across four continents (Alan, Dawson, & Allan, 2006, Australia; Ducro & Pham, 2006, Belgium; Baltieri & de Andrade, 2008b, Brazil; Kingston, Yates, Firestone, Babchishin, & Bradford, 2008, Canada; Bengston, 2008, Denmark; Stadtland et al., 2006, Germany; Skelton, Riley, Wales, & Vess, 2006, New Zealand; Craig, Beech & Browne, 2006, the United Kingdom) with sex offenders of varying demographics and there have been more than 60 replication studies supporting the Static-99's ability to predict future sex offending behavior.

Research that has examined the IRR of the Static-99 total score has found high levels of rater consistency (Hanson, 2001, intraclass correlation coefficient [ICC = .87]; Barbaree, Seto, Langton, & Peacock, 2001 [ICC = .90]; Harris et al., 2003 [K = .96]; de Vogel & de Ruiter, 2004 [K = .90]; Rettenberger, Matthes, Boer, & Eher, 2009 [ICC =.98]). Not only has the IRR of the Static-99 generally been found to be quite high, but it has also been found to be higher than that of comparable measures. For example, Barbaree et al. (2001) reported the IRR of the Static-99 total score to be demonstrably higher (r = .90) than that of the Minnesota Sex Offender Screening Tool–Revised (Mn-SOST-R; Epperson et al., 1998; r = .80). Harris and colleagues (2003) also found the individual items on the Static-99 to have higher IRR than items on the Psychopathy Checklist–Revised (PCL-R; Hare, 1991) and the Sex Offender Risk Appraisal Guide (SORAG; Quinsey, Harris, Rice, & Cormier, 1998). Doren (2004), who highlighted eight studies pertaining to the IRR of the Static-99 and its predecessor, the Rapid Risk Assessment for Sex Offence Recidivism (RRASOR), concluded, "Neither instrument has ever to date failed to show a high degree of IRR when empirically studied" (p. 26).

Like other actuarial risk assessment tools, the Static-99 attempts to reduce the variability of unstructured clinical judgment. Although generally considered a marked improvement over unstructured risk assessments, rater variance remains an issue even with these more structured tools. Despite accompanying coding manuals with well-operationalized terms and the historical or static nature of the tool's items, which tend to be invariant once assigned and therefore more reliably coded, actuarial risk

assessment tools such as the Static-99 may still be vulnerable to coding-manual interpretation errors and rater subjectivity (Glancy, 2006).

Much of what we know about the Static-99's reliability has derived from studies conducted under tightly controlled research conditions. Invariably, conditions nearer to tool validation research will produce higher validity and reliability estimates. Research on the predictive efficacy of the Violent Risk Appraisal Guide (VRAG; Quinsey et al., 1998), SORAG (Quinsey et al., 1998) and Static-99 (Hanson & Thornton, 1999; 2000) has revealed an enhanced ability to predict risk in studies conducted by instrument authors as opposed to studies conducted by instrument nonauthors (Blair, Marcus, & Boccaccini, 2008). This enhanced predictive accuracy, which may be due to the test authors' presumed ability to score the tool more accurately than nonauthors (Blair et al., 2008) or may reflect conditions closer to original tool validation in more tightly controlled research contexts, highlights the need for examination of tool reliability and validity in field settings.

Indeed, we know far less about the performance of risk measures in the field, where coding is likely to be completed by clinical or correctional staff who may have received less formal training on risk measures. The predictive validity of the Static-99 hinges on accurate scoring, and thus deviations from those scores may negatively impact the scale's performance. Considering the substantial impact that risk assessment results can have on sex offenders and the community, further research is needed on the reliability of actuarial methods in assessing sex offenders in "real world" settings. Still very little research has examined the reliability of the Static-99 in routine practice and among practitioners in the field, though findings from two field studies highlight some important concerns regarding IRR. Despite finding overall high levels of IRR (ICC values between .79 and .88) among field evaluators in correctional settings in Texas and New Jersey; Boccaccini et al. (2012) observed that raters arrived at different total scores in approximately 45% of the cases. These score differences warrant attention because even a 1-point score difference may affect decision-making about the management of sex offenders. Murrie and colleagues (2009), who examined the performance of the Static-99 within the high-stakes SVP context, found strong evidence for an allegiance effect. Score differences on the Static-99 and other risk measures varied systematically depending on the context; that is, whether evaluators were working for the petitioner or the respondent. IRR estimates between respondent and petitioner on the Static-99 were still quite high (ICC = .64), with raters showing less disagreement in comparison to other tools. Murrie and colleagues speculated that the enhanced reliability of the Static-99 was likely due to its greater emphasis on historic/static factors, suggesting that the Static-99 may be less vulnerable to rater bias.

To date, the only study to specifically compare risk assessment scores *between researchers and clinicians* focused on the Historical-Clinical-Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997), an adjusted actuarial risk assessment tool (de Vogel & de Ruiter, 2004). Using a Dutch version of the HCR-20, de Vogel & de Ruiter (2004) found that some clinicians (treatment group leaders) gave lower scores on the HCR-20 than did researchers. While other clinicians (treatment

supervisors) did not show mean score differences compared with researchers, they did show differences in their overall risk judgments; with treatment supervisors being more apt to interpret these scores as "low risk" than researchers. This is in contrast with other research which suggests that practitioners are more likely to score higher, leading to false positives or Type I errors (Edens, Buffington-Vollum, Keilen, Roskamp, & Anthony, 2005; Monahan, 1981; Mossman, 1994). While some independent studies have examined IRR of actuarial risk assessment tools among researchers (Harris et al., 2003), to date, no study has compared scores between researchers and field practitioners on an actuarial risk tool such as the Static-99.

Current Study

The current study examined IRR between field practitioners, who scored the Static-99 in a correctional setting, and researchers, who blindly coded the Static-99 at a later point using archival data. This project sought to examine the psychometric properties of the Static-99 outside of the more artificial research conditions, from where IRR estimates are typically derived, and to assess whether scores in the field are comparable to those derived by researchers. The goals of this study were threefold: (a) to assess the IRR between researchers and field practitioners on the Static-99, (b) to examine the frequency of inconsistencies between researchers and practitioners (i.e., whether practitioners tend to rate offenders as lower or higher risk than do researchers), and (c) to determine the nature of the disagreement in item ratings, as perceived by the researcher.

Method

Procedures

Data were collected as part of a larger study (Mercado, Jeglic, & Markus, 2008) that examined sex offender risk, selection for treatment and SVP commitment, and recidivism. The archival files of offenders who had an index sex offense and who were released from a state correctional facility between the years of 1996 and 2007 ($N = 3{,}175$) were reviewed by a team of graduate level research assistants. While the precise job titles of the practitioners were not available, level of education/professional training was coded and included individuals holding MA, MSW, LMSW, PhD, PsyD, and MD degrees. All researchers were trained on Static-99 administration, which included coding manual review and the completion of practice tests found on the Static-99 website (www.Static99.org). IRR was tabulated between researchers on a subsample of 30 cases and the ICC between total scores was "excellent" (ICC = .890) with individual item IRR ranging from "fair" ("prior sentencing dates," $k = .266$ and "prior nonsex violence," $k = .344$) to "outstanding" ("male victims," $k = .869$).

Using information in the file, research assistants completed blind Static-99 ratings (i.e., rating the Static-99 before viewing the Static-99 already in the file). After their

own review of the file and subsequent coding of their Static-99, research assistants then transcribed practitioner Static-99 scores to the data collection tool. When there were scoring discrepancies, researchers were asked to indicate what they perceived to be the nature of the error using the following categories:

a. *Total score calculation error*, operationalized as an error that resulted from a miscalculation leading to an incorrect total score.
b. *Discrepancy between file data and score given*, identified as when a coder gave a score to an item that does not coincide with the information available in the file. For example, scoring prior sex offenses with a 2 where it should have been marked as a 1 given information in the file about the number of prior offenses.
c. *Coding manual error*, or an error identified when a coder does not score an item according to the rules in the manual. For example, scoring a "stepchild" victim as unrelated despite an explicit coding manual rule to score a stepchild as a related victim.
d. *Discrepancy due to subjectivity of item*, identified as a discrepancy resulting from subjective judgment required in scoring the item. Hanson, Morton, and Harris (2003) suggested that the only subjectively scored item is whether the offender is "single" (has the offender lived with a partner for more than 2 years). As this information is not always clear from a historical account of relationship history, scoring may result in more subjective judgment.

Measures

Static-99. It is a 10-item actuarial risk assessment tool designed to predict sexual recidivism. The static risk variables include young age at release, ever lived with a lover for at least 2 years, prior convictions for nonsexual violence, index offence having nonsexual violence, number of prior sentencing dates, history of offending against unrelated victims, history of offending against male victims, history of offending against stranger victims, and number of prior offenses (see Hanson & Thornton, 1999, 2000). Each of the items, except number of prior offenses, is scored dichotomously as 1 (*present*) or 0 (*absent*). Scores for the number prior offenses item range from 0 (*none*) to 3 (*4 or more convictions, or 6 or more charges*). Scores on the Static-99 range from 0 to 12, with scores of 1 or below indicating low risk, scores of 2 or 3 indicating moderate-low risk, scores of 4 or 5 indicating moderate-high risk, and scores of 6 or above indicating high risk (Harris et al., 2003).

Archival records. Research assistants scored the Static-99 based on information available in the offender's prison file and clinical record. Available information typically included demographics, developmental history, criminal history, psychiatric and treatment history, educational and employment history, incarceration review, victim statements, and victim demographics.

Results

Of the 3,175 files reviewed, only 53% (n = 1,973) had comparable Static-99 total scores between researcher and practitioner. This sample consisted of offenders whose index offense included a contact sexual offense against an adult (21.6%, n = 424), a contact offense against a child (67.4%, n = 1,326), or a noncontact sexual crime (2.9%, n = 55), 8.1% (n = 161) of offenders had multiple types of offenses. Of those 1,973 offenders for whom data on race were available, 37.3% (n = 735) were identified as White, 40.8% (n = 804) as African American, 20.2 % (n = 398) as Latino, and less than 1% (n = 25) as Asian or Pacific Islander.

Interrater Agreement: Field and Research

Kappa correlations were used to identify levels of rater agreement across all items in the Static-99, while the ICC was used to identify levels of interrater consistency on the total score. Cohen's Kappa statistic is the primary measure used to analyze the degree of consensus between independent raters on dichotomous items, whereas the ICC statistic measures reliabilities between raters on scaled items, such as a total score. The current study used Landis and Koch's (1977) Kappa interpretations, with K = 0.01 to 0.20 indicating slight agreement, K = 0.21 to 0.40 indicating fair agreement, K = 0.41 to 0.60 moderate agreement, K = 0.61 to 0.80 as substantial agreement, and K = 0.81 to 1.00 indicating outstanding agreement. For the ICC interpretations, we used Fleiss's (1986) critical values for single measures whereby ICC ≥ .75 = excellent; .60 < ICC < .75 good; .40 < ICC < .60 moderate; and ICC < .40 poor.

When examining the 10 items of the scale individually, there was "substantial" consistency between researcher and practitioner ratings on eight items (k = .621-.788) and "outstanding" consistency on the remaining two items, including "male victims" (k = .941) and "stranger victims" (k = .804; see Table 1). The two items demonstrating the lowest consistency between clinician and researchers were "index offense nonsexual violence" (k = .621) and "conviction for a noncontact sex offense" (k = .671).

With regard to total Static-99 total scores, practitioners and researchers demonstrated an "excellent" level of agreement (ICC = .924) more than 55% of the time. Within the sample, 40.9% (n = 807) indicated some form of discrepancy between practitioner and researcher scoring, whereas 59.1% (n = 1,165) indicated no difference. Of the 807 cases where differences occurred, raters still arrived at the same total score (despite item coding differences) in 11.2% (n = 90) of the cases. 75.0% (n = 557) had a 1-point difference in total score, which accounts for 28.2% of the total sample. Last, 18% (n = 135), had a 2-point difference in total score, which accounts for 6.84% of the total sample. The final 7% (n = 57) of Static-99 score discrepancies had cases where the difference between researcher and practitioner score was more than 3 points, which accounts for 2.8% of the total original sample.

In general, however, researchers were more likely (M = 2.96, SD = 2.02) to arrive at higher scores than were practitioners (M = 2.77, SD = 1.98), $t(4668)$= 3.2, p = .04. In more than half of the cases where there was a total score discrepancy, this resulted

Table 1. Correlations Between Researcher and Practitioner Scores on Individual Items.

Static-99 item	*n*	Kappa/ICC	% agree	Interpretation[a]
Young	1,590	.761	94.3	Substantial
Ever lived with lover	1,569	.788	89.8	Substantial
Index nonsex violence	1,583	.621	91.5	Substantial
Prior nonsex violence	1,584	.693	86.9	Substantial
Prior sex offences	1,547	.679	87.0	Substantial
Prior sentencing dates	1,589	.758	89.1	Substantial
Noncontact sex offences	1,582	.671	95.7	Substantial
Unrelated victims	1,587	.733	91.1	Substantial
Stranger victims	1,573	.804	87.0	Outstanding
Male victims	1,580	.941	98.5	Outstanding
Total score	1,700	.924	55.8	Excellent[b]

Note. ICC = intraclass correlation coefficient.
[a]Interpretation of values are found in Landis and Koch (1977).
[b]Interpretation as cited by Fleiss (1986).

in the scores falling into different risk categories. In just more than a quarter of the cases (26.8%, $n = 210$) this resulted in researchers indicating a higher risk category than practitioners. In a nearly equivalent number of cases, this resulted in (26.6%, $n = 208$) practitioner's scores falling into a higher risk category than those of researchers. The remaining 46.6% ($n = 365$) indicated a discrepancy in total score but no change in risk category.

Nature of Differences Between Practitioner and Researcher Scores

Researchers judged the most common cause of discrepancy to be *coding manual errors* ($n = 196$). Other identified sources of discrepancy included *discrepancy due to item subjectivity* ($n = 165$), *discrepancy between file data and score given* ($n = 118$), and *total score calculation error* ($n = 27$). An additional 197 files contained discrepancies in more than one of the above stated error categories (see Figure 1).

Discussion

The current study is the first to examine the IRR of Static-99 between scores from the field and those of researchers. Similar to other studies examining the psychometric properties of the Static-99, results generally showed high IRR between practitioners and researchers on the Static-99, with total score IRR estimates between practitioners and researchers being "excellent." Kappa estimates on individual items ranged from "substantial" (on eight items) to "outstanding" (on two items).

Estimates of total score IRR from this study were within the range (ICC = .924) of that found in previous research (Barbaree et al., 2001; Hanson, 2001; Rettenberger

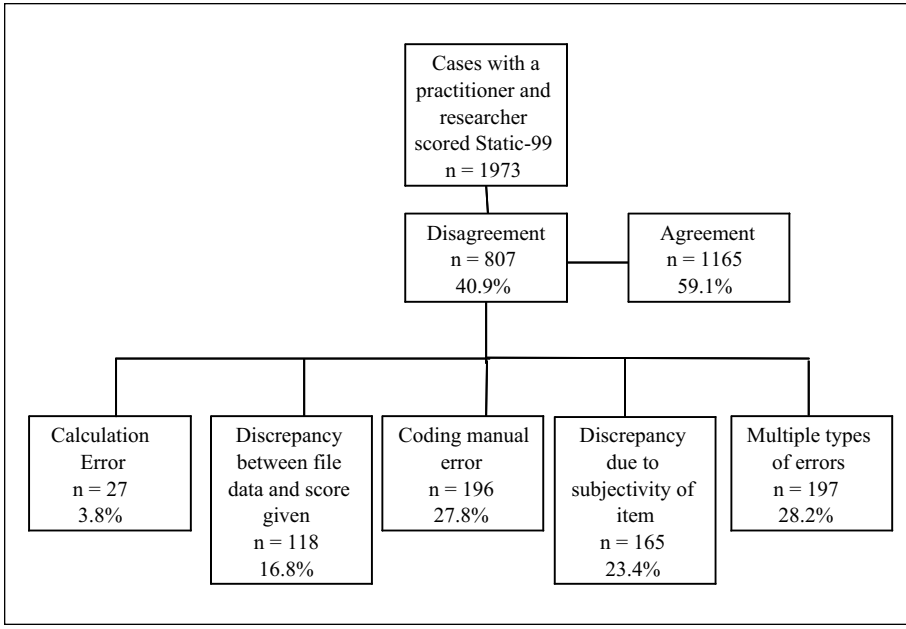

Figure 1. Static-99 error type.

et al., 2009). Although IRR estimates were generally very high, this study found somewhat lower IRR than did Harris et al. (2003), who observed "outstanding" agreement for all 10 items. In the current study, eight items were found to have a more modest "substantial" level of agreement. Most notably, the field practitioner–researcher IRR for the items "Any convictions for a noncontact sex offense" and "Index nonsexual violence" was .671 and .621, respectively. The lower level of IRR for these two items was unexpected, as item scoring is based solely on official records, which should, theoretically, make scoring of these items relatively straightforward. The fact that practitioners tended to rate this item lower than researchers may suggest that contact with the offender affected scoring, though not in the expected direction. Indeed, contact with the offender might be expected to have resulted in practitioners having knowledge of offenses or information not recorded in the file. The IRR for these items is below what is generally considered to be acceptable (Landis & Koch, 1977).

This study also examined differences in total scores between practitioners and researchers. Overall it was found that practitioners and researchers arrived at an identical total score more than half (55%) of the time, regardless of whether there were different scores on individual items. This rate is consistent with that found by Boccaccini and colleagues (2012), who also observed that evaluators in the field arrived at an identical score approximately 55% of the time. Because even a 1-point score difference can have important impact on the disposition of an offender, these findings about

the frequency of rater disagreement have important implications for the safety of the community. Boccaccini et al. suggested that enhanced training, the reporting of confidence intervals for Static-99 total scores, and the use of multiple raters as possible procedures that might be used to enhance reliability.

Notably, of the files in which there were rater discrepancies, practitioners were more likely to arrive at a lower total score than were researchers. Given that practitioners might have access to information outside of the file that researchers were not privy to, this finding was somewhat surprising. Indeed, practitioners in the field might have knowledge about other victims or other charges that are not in the file, whereas the only information available to the researchers was the file information. It is still unclear to what lead to the discrepancies in scoring between the researchers and practitioners, however future research that examines the differential validity of practitioner versus researcher ratings might help to elucidate whether additional information (i.e., file data plus nonfile information) or other aspects of the practitioner enhances or limits its predictive capacity.

For any item for which there was a score difference between practitioner and researcher, the cause of the discrepancy was examined. The most common explanation for scoring discrepancies was judged to be "error in coding from manual" (e.g., clinician failed to score "stepchild" as a "related victim" because a stepchild is not related by blood, despite instructions in the coding manual that such a person should be coded as related). One cause of these errors could be ambiguity or literal interpretation of the item name could have falsely mislead a rater to indicate this item as present in these cases. For example, "any unrelated victims" was sometimes rated as present in cases in which the victim and the offender were of no blood relation (e.g., stepchild, spouse) but should be coded as being related given the Static-99 operational definition of relationship (Hanson & Harris, 2001). Given that identified discrepancies were most frequently judged to result from coding manual error, these findings suggest the importance of additional (or more thorough) training in use of the Static-99 tool and its accompanying manual. Calculation and mismarking of the tool errors were relatively infrequent, as judged by the researchers.

There are several limitations of the current study. First, practitioners and researchers in this study may have had access to different levels of information. Although researchers and practitioners had access to the same file information, practitioners may also have had contact with offenders or access to other clinical or crime-related information not reported in the file. While the Static-99 codebook does allow for self-report for items pertaining to demographic or victim information, the authors caution that confirmation with official records is always preferable. In addition, the Static-99 authors stipulate that only official records are permissible for criminal history (www.static99.org). Second, although all field practitioners presumably received the training necessary to score the Static-99, we had incomplete information about the qualifications and training of most of the practitioners who scored these measures. Finally, the cause of errors in practitioner Static-99 scores were inferred by the researchers based on the available information; however, except for obvious instances (i.e., calculation error, error in following instructions of the manual), the cause of error was based on

the judgment of the researcher and, as such, is prone to error. Future research should incorporate reliability checks on the researchers' scoring of the errors.

The current study provides evidence for general consistency in the way researchers and practitioners score the Static-99, while also suggesting that scores in purely clinical forensic settings may not be quite as closely comparable to those in research settings as previously assumed. These findings have some important implications. While few, if any, jurisdictions rely solely on one measure to make decisions about sex offender placement, it is clear that in many instances Static-99 scores factor heavily in the decision-making process (Murrie et al., 2009). In this study, more than one third of Static-99 scores differed. Even a one-point difference can make the difference between risk levels. While more high stakes decisions (such as civil commitment decision making in which actuarial risk assessment play an important role) typically rely on the findings of two or more evaluators, who each typically score a risk measure as part of their evaluation, this is typically not the case in other circumstances such as decisions pertaining to treatment or parole. The findings from this study suggest that score differences are relatively common, and it is certainly the case that these differences could result in more or less intensive management services based on small, even 1 point, score differences. What remains unclear, however, is how decision makers should and do deal with score discrepancies between evaluators.

Given the potential ramifications of such discrepancies, replication of these findings may help determine whether certain trends (e.g., practitioners tending to score the Static-99 lower than researchers) exist in the coding of actuarial tools. Moreover, future research should gather recidivism data to determine whether comparative differences exist in the predictive accuracy of practitioners and researchers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Allan, A., Dawson, D., & Allan, M. (2006). Prediction of the risk of male sexual reoffending in Australia. *Australian Psychologist*, *41*, 60-68. doi:10.1080/00050060500391886

Archer, R., Buffington-Vollum, J., Stredny, R., & Handel, R. (2006). A survey of psychological test use patterns amongst forensic psychologists. *Journal of Personality Assessment*, *87*, 84-94.

Baltieri, D. A., & de Andrade, A. (2008b). Drug consumption among sexual offenders against females. *International Journal of Offender Therapy and Comparative Criminology*, *52*, 62-80. doi:10.1177/0306624X07299345

Barbaree, H., Seto, M., Langton, C., & Peacock, E. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior*, *28*, 490-521.

Bengston, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime and Law*, *14*, 85-106.

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

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

Craig, L., Beech, A., & Browne, K. (2006). Cross validation of the Risk Matrix 2000 sexual and violent scales. *Journal of Interpersonal Violence*, *21*, 612-633.

de Vogel, V., & de Ruiter, C. (2004). Differences between clinicians and researchers in assessing risk of violence in forensic psychiatric patients. *The Journal of Forensic Psychiatry & Psychology*, *15*, 145-164.

Doren, D. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 25-36.

Ducro, C., & Pham, T. (2006). Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: Journal of Research and Treatment*, *18*, 15-26.

Edens, J. F., Buffington-Vollum, J. K., Keilen, A., Roskamp, P., & Anthony, C. (2005). Predictions of future dangerousness in capital murder trials: Is it time to "Disinvent the Wheel?" *Law and Human Behavior*, *29*, 55-86.

Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Alexander, W., & Goldman, R. (1998). *Minnesota Sex Offender Screening Tool–Revised (MnSOST-R)*. St. Paul: Minnesota Department of Corrections.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York, NY: John Wiley.

Glancy, G. (2006). Caveat Usare: Actuarial schemes in real life. *Journal of American Academy of Psychiatry and Law*, *34*, 272-275. Available from http://www.jaapl.org/

Hanson, K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119-136.

Hanson, R. K. (2001). *Age and sexual recidivism: A comparison of rapists and child molesters* (User Report 2001-01). Ottawa, Canada: Department of the Solicitor General of Canada. Available from www.sgc.gc.ca

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, *66*, 348-362.

Hanson, R. K., & Harris, A. J. (2001). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *13*, 105-122.

Hanson, R. K., Morton, K. E., & Harris, A. J. (2003). Sexual offender recidivism risk: What we know and what we need to know. *Annals of the New York Academy of Science*, *989*, 154-166.

Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex offenders* (User Report 1999-02). Ottawa, Canada: Department of the Solicitor General of Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist–Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Harris, G., Rice, M., Quinsey, V., Lalumie`re, M., Boer, D., & Lang, C. (2003). A multi-site comparison of actuarial risk instruments for sex offenders. *Psychological Assessment*, *15*, 413-425.

Kingston, D., Yates, P., Firestone, P., Babchishin, K., & Bradford, J. (2008). Long-term predictive validity of the Risk Matrix 2000: A comparison with the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: Journal of Research and Treatment*, *20*, 466-484.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-33,174.

Mercado, C. C., Jeglic, E., & Markus, K. (2008). *Sex offender management, treatment, and civil commitment: An evidence based analysis aimed at reducing sexual violence* (2007-IJ-CX-0037). Washington, DC: National Institute of Justice.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques*. Beverly Hills, CA: Sage.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783-792.

Murrie, D., Boccaccini, M., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis) agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law*, *15*, 19-53.

Quinsey, L. V., Harris, T. G., Rice, E. M., & Cormier, A. C. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Rettenberger, M., Matthes, A., Boer, D., & Eher, R. (2009). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology*, *2*, 169-186.

Skelton, A., Riley, D., Wales, D., & Vess, J. (2006). Assessing risk for sexual offenders in New Zealand: Development and validation of a computer-scored risk measure. *Journal of Sexual Aggression*, *12*, 277-286.

Stadtland, C., Hollweg, M., Kleindienst, N., Dietl, J., Reich, U., & Nedopil, N. (2006). Evaluation of risk assessment instruments for sex offenders. *The Neurologist*, *77*, 587-595.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing the risk for violence* (Version 2). Vancouver, British Columbia, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.