Open Access Journal of Forensic Psychology

http://www.forensicpsychologyunbound.ws/ -- 2010.2: 337-346

Maximizing Accuracy and Welcoming Scrutiny in SVP Evaluations: An Additional Response to Wilson and Looman

Terence W. Campbell, Ph.D., ABPP (forensic)
Sterling Heights, Michigan

Abstract

Using binary methodology to assess the accuracy of an actuarial instrument such as the Static-99R is entirely consistent with how those instruments are used in practice. Despite Wilson and Looman's position to the contrary, the high-risk norms of the Static-99R appear less than representative of U.S. sex offenders subject to civil commitment. Unfortunately, the predictive accuracy of the Static-99R never exceeds that obtained by relying on the base rate alone. Attempts at protecting the Static-99R from scrutiny are ill advised, inviting skepticism and disrespect from triers of fact.

DeClue and Campbell (2010) clearly detailed how binary methodology can be appropriately applied to the various Static instruments. This additional response emphasizes that progress in the field of sex-offender risk assessment, and basic ethical obligations, necessitate closely scrutinizing these instruments.

Keywords: Static-99, Static99R, SVP, risk assessment, recidivism

Actuarial Instruments and Binary Methodology

In their response to our (2010a) paper, Wilson and Looman (2010) rely extensively on a 2010 article by Hanson and Howard. Hanson and Howard (2010) contend: "In contrast to diagnoses, risk assessments estimate the likelihood of an event that has not yet happened, and may never happen. They are inherently stochastic, and the future outcome can only be estimated with a certain probability" (p. 276). For the various Static instruments, the only source of information approximating a manual (www.static99.org) reports recidivist-nonrecidivist outcomes for each score for the Static-2002, Static-2002R, and the Static-99R. Quite clearly, those outcomes are known and well identified; and as a result, the accuracy of those instruments can be assessed via binary methodology.

Using binary methodology to assess the accuracy of an actuarial instrument such as the Static-99R is also entirely consistent with how those instruments are used in practice. Over the course of my more than 12 years of experience with SVP matters, each and every assessment report I have reviewed reaches one of two conclusions regarding risk of reoffending, i.e., the offender will, or will not, reoffend. Moreover, SVP proceedings allow for one - and only one - of two decisions: Commit or do not commit. In other words, SVP commitments do not allow for a range of commitment options; and

as a result, binary methodology is entirely appropriate for assessing the accuracy of these risk assessments.

Wilson and Looman also contend: "Simply put, the Static-99 and its progeny were not designed to facilitate answering the 'more likely than not' [to sexually reoffend] question faced by SVP evaluators, no matter how much they may want them to" (p. 12). I recommend that Wilson and Looman reconsider this position in view of Hanson's 1998 recommendations regarding 2 X 2 prediction tables. Citing a September 2009 ATSA presentation by Hanson, Phenix, and Helmus, and a December 2009 training presentation by Thornton and Helmus, Sreenivasan, Weinberger, Frances, and Cuswoth-Walker (2010) characterized them as "declarations that this process of rating alone [Static-99R] is sufficient to describe risk" (p. 403). In other words, no matter how much Wilson and Looman may criticize the practice of using the Static-99R to assess reoffending risk, it is nonetheless done with great regularity. If Wilson and Looman persist in their position, I would ask: *If rejecting actuarial assessment as the centerpiece of risk assessment, then how should an SVP evaluator proceed?* Surely, any response must amount to more than a poorly disguised retreat into the vagaries of clinical judgment.

High-Risk Norms

Wilson and Looman advocate using the high-risk/need normative sample of the Static-99R for assessing risk of sexual reoffending. In their 2009 Power Point presentation, Hanson et al. advise that using the high-risk/need norms necessitates justifying that the routine norms do not apply to a particular offender. In particular, they recommend identifying offenders who are a "Member of small minority selected on risk/need factors external to Static-99R/Static-2002R: use high-risk/need norms" (p. 20). This recommendation necessitates asking, *Selected on what risk/need factors external to the Static-99R/Static-2002R?* The most compelling feature of this recommendation is the unavailability of a standardized procedure for following it. Too often, such risk/need factors are found more frequently in samples of offenders who do not reoffend compared to those who do (Campbell & DeClue, 2010b). Relying on these risk/need factors therefore results in false-positive classifications of sex offenders. Hanson et al. (2009) also indicated that using the "Routine norms" is "sufficient in most circumstances" (p. 20).

Hanson, Phenix, and Helmus (2009) advocated SVP evaluators to rely on local norms when possible. Examining the sources for the high-risk/needs norms identifies them as far less than "local" when evaluating U.S. offenders. The total high-risk/needs sample for the five-year follow-up is 1,313 offenders drawn from six separate samples. The high-risk/needs sample for the 10-year follow-up is 722 offenders drawn from five separate samples. These samples include the following (Phenix, Helmus, & Hanson, 2009, p. 12-13).

Bengston, 2008: This sample included Danish sex offenders undergoing pre-trial forensic assessment. These offenders presented histories of possible mental disorder

or mental retardation. The specific number of offenders from this sample is not identified.

Bonta & Yessine, 2005

This sample included Canadian offenders flagged as potentially Dangerous Offenders by Canada's National Flagging System. The specific number of offenders from this sample is not identified.

Haag, 2005

This sample included all male Canadian federal sex offenders with Warrant Expiry dates in 1995. The specific number of offenders from this sample is not identified.

Knight and Thornton, 2007

This sample included sex offenders assessed or treated at the Massachusetts Treatment Center between 1959 and 1984. In the original study, corresponding to this sample, Prentky, Lee, Knight, and Cerce (1997) advised: "The obvious, marked heterogeneity of sexual offenders precludes automatic generalization of the rates of reported here to other samples" (p. 656). The specific number of offenders from this sample is not identified.

Nicholaichuk 2001

This sample consists of sex offenders undergoing treatment at a federal maximum-security facility located in Saskatchewan. The specific number of offenders from this sample is not identified.

Wilson and colleagues, 2007

This sample consists of Canadian sex offenders detained in prison until their Warrant Expiry Date. Half of this group participated in Circles of Support and Accountability as part of their treatment. The specific number of offenders from this sample is not identified.

To say the least, it is unlikely that four Canadian samples, and one idiosyncratic U.S. sample can serve as comparison groups for contemporary U.S. offenders. In her 2009 thesis, Helmus addressed this issue emphasizing: "These norms should not be adopted in all contexts without caution. Given that base rates showed such large variability across samples, evaluators cannot take for granted that any set of norms is going to apply to the context in which they are assessing risk" (p. 131). It also becomes necessary to ask whether Wilson and Looman actually want to rely on the high-risk sample of 703 previously convicted sex offenders. While referring to how insurance companies use actuarial tables for rate decisions, Wilson and Looman might want to consider the many thousands of individuals included in those tables.

Boccaccini, Murrie, Capterton, and Hawes (2009) recommended that large jurisdictions (e.g., state correctional systems) develop their own local base rates of sexual recidivism. Boccaccini et al. cited Standard 3.13 of the 1999 Standards for Educational and Psychological Testing (SEPT) advising: "Before using an empirically derived measure, clinicians should ensure that supportive data is available from a sample 'sufficiently large and representative of the population for which the test is intended'" (p. 282) (SEPT, 1999, p. 46). It seems clear and evident that this sample of 703 previously convicted offenders is neither large enough, nor sufficiently representative, to support decision making in high-stakes legal proceedings such as SVP matters.

Identifying Predictive Accuracy

I would also point out that somehow Wilson and Looman have misread and/or misunderstood a basic issue in our 2010 paper. We do not—nor did we ever—advocate relying on a Static-99 score of 6 to maximize predictive accuracy. Nothing could be further from the truth. Like Wilson and Looman, I consider the original Static-99 as obsolete. When using the Static-99R, for example, I recommend obtaining the offender's score, and clearly identifying each of the 2 X 2 outcomes associated with that score for all four norm groups. In particular, consider an offender whose Static-99R score is a 6. Tables 1-8 report the outcomes for that score across each of the four norm groups.

Table 1: Outcomes obtained if relying on the "Routine" norms for a Static-99R score of 6 and above when ruling in recidivism risk

	True	False
Positive	41	193
Negative	2068	104

Positive Predictive Value (PPV) = .18 (41/234).

Table 2: Outcomes obtained if relying on the "Routine" norms for a Static-99R score of 6 and below when ruling out recidivism risk

	True	False
Positive	27	97
Negative	2164	118

Negative Predictive Value (NPV) = .95 (2164/2282).

Table 3: Outcomes obtained if relying on the "Selected for treatment" norms for a Static-99R score of 6 and above when ruling in recidivism risk

	True	False	
Positive	27	51	
Negative	697	9	PPV = .35 (27/78)

Table 4: Outcomes obtained if relying on the "Selected for Treatment" norms for a Static99R score of 6 and below when ruling out recidivism risk

	True	False	
Positive	15	26	
Negative	722	103	NPV = .90 (722/825)

Table 5: Outcomes obtained if relying on the "Non-routine" norms for a Static-99 score of 6 and above when ruling in recidivism risk

	True	False	
Positive	118	185	
Negative	1109	1214	PPV = .39 (118/303)

Table 6: Outcomes obtained if relying on the "Non-routine" norms for a Static-99 score of 6 and below when ruling out recidivism risk

	True	False	
Positive	70	103	
Negative	1191	262	NPV = .82 (1191/1453)

Table 7: Outcomes obtained if relying on the "High-risk" norms for a Static-99 score of 6 and above when ruling in recidivism risk

	True	False	
Positive	85	118	
Negative	381	119	NPV = .42 (85/203)

Table 8: Outcomes obtained if relying on the "High-risk" norms for a Static-99 score of 6 and below when ruling out recidivism risk

	True	False	
Positive	53	67	
Negative	432	151	NPV = .74 (432/583)

Tables 9 and 10 summarize the outcomes obtained for ruling in, or ruling out, recidivism risk for each of the four norm groups. PPV refers to Positive Predictive Value, or how accurately a given cutoff score identifies those who will reoffend. NPV refers to Negative Predictive Value, or how accurately a given cutoff score identifies those who do not reoffend.

Table 9: Sample Ns, Base Rates, Overall Accuracy, and PPV for four Static-99R norm groups when ruling in recidivism risk

Ruling-In Recidivism Risk

	Routine	Tx Need	Non-Routine	High-Risk
N of Group	2406	866	1626	703
Base Rate	6%	14%	20%	29%
Overall Accuracy	88%	84%	75%	66%
PPV	18%	35%	39%	42%

Table 10: Sample Ns, Base Rates, Overall Accuracy, and NPV for four Static-99R norm groups when ruling out recidivism risk

Ruling-Out Recidivism Risk

	Routine	Tx Need	Non-Routine	High-Risk
N of Group	2406	866	1626	703
Base Rate	6%	14%	20%	29%
Overall Accuracy	91%	85%	78%	69%
PPV	95%	90%	82%	74%

Given the data found in Tables 9 and 10, an evaluator can explain that Positive Predictive Value (PPV) means that, if ruling in recidivism risk, that decision will be correct somewhere between 18% and 42% of the time, depending on which norm group is considered. Ruling in recidivism risk will also be mistaken somewhere between 58% and 82% of the time. An evaluator can also explain that, if ruling out recidivism risk, that decision will be correct somewhere between 74% and 95% of the time depending

on which norm group is considered. Ruling out recidivism will also be mistaken somewhere between 5% and 26% of the time. Obviously, then, ruling out recidivism risk is typically more accurate than ruling it in.

What is particularly sobering about these data is the predictive accuracy obtained via the Static-99R never exceeds that obtained by relying on the base rate alone. For example, if using the Routine norms and concluding that no one will reoffend, an evaluator would be correct in 94% of cases. The overall accuracy for ruling in recidivism risk, at a cutoff of 6 and above, via the Routine norms is 88%. The overall accuracy for ruling out recidivism risk, at a cutoff of 6 and below, via the routine norms is 91%.

For Vrieze and Grove (2008), base-rate predictions provide a rational, objective index of predictive accuracy. They point out how, "Comparing the CF [Correct Fractions] of an instrument to the CF of betting the base rate is, quite frankly, a not very demanding validity hurdle" (p. 275). Vrieze and Grove further indicated that an actuarial instrument outperforming base-rate predictions demonstrates incremental validity. If an actuarial instrument does not outperform the base rate, Vrieze and Grove insist it "really has not much going for it" (p. 275).

Welcoming Scrutiny

Doren (2000) previously argued that the 1999 Standards for Educational and Psychological Testing (SEPT) were not applicable to actuarial instruments such as the Static-99. His arguments were premised on a misunderstanding of the term "test," and an excessively broad definition of who could use these instruments (Campbell, 2001). If Doren's arguments had gained any traction, the Static-99 would have avoided scrutiny via the SEPT. In 2006, Doren argued that SVP evaluators engaged not in "risk prediction," but rather in "risk assessment." A review of the relevant literature (Campbell, 2007), including Doren's own work, demonstrated that "risk prediction" has long been generally recognized and accepted by the community of SVP evaluators. If Doren's 2006 position had been generally accepted, evaluating actuarial instruments with indices of predictive accuracy would be inappropriate.

Whether they intend it or not (and most likely they do not), Wilson and Looman (2010) appear engaged in attempts at avoiding scrutiny for the Static instruments. Avoiding this kind of scrutiny is inconsistent with the ethical obligation of forensic psychologists to engage in transparent procedures. Reporting the kinds of data found in Tables 9 and 10 responds to this ethical obligation. Dismissing or attempting to circumvent these data will only arouse the skepticism and disrespect of triers of fact in SVP proceedings.

Though it represented progress as a third-generation risk assessment instrument, numerous problems are now inundating the Static-99 and the other Static instruments (Bani-Yaghoub, Feeroff, Curry, & Amundsen, in press). Linear additive instruments (Farrington & Tarling, 1985), such as the Static-99R, neglect the intercorrelations (or

overlap) among the factors they use. Additionally, these instruments rely on a "one-size-fits-all" approach in which all kinds of sex offenders (rapists, child molesters targeting males, child molesters targeting females, child pornography offenders, exhibitionists, etc.) are assessed in the same manner by the Static-99R. Progress in actuarial risk assessment necessitates that we learn from the Static-99 instruments, and then move on beyond them. Protecting the Static-99 instruments from scrutiny will only delay this needed progress.

Ethical Considerations

In their review of future dangerousness assessments related to capital sentencing, Edens, Buffington-Vollum, Keilen, Roskamp, and Anthony (2005) cited Standard 9.01 (a) - addressing "Bases for Assessments" - of the 2002 APA Ethical Code. Psychologists base their opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings." (p. 1071).

Accordingly, Edens and his colleagues advised: "It is incumbent upon mental health experts to eschew unreliable testimony that lacks a meaningful scientific foundation" (p. 77). To belabor the obvious, the comments of Edens et al. are equally applicable to SVP commitment proceedings. Disregarding the PPV and NPV of any actuarial instrument entirely undermines whatever scientific foundation that instrument can claim.

Additionally, Standard 9.06 of the American Psychological Association's 2002 Ethical Code - addressing "Interpreting Assessment Results" - states:

> When interpreting assessment results, including automated interpretations, psychologists take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists' judgments or reduce the accuracy of their interpretations. **They indicate any significant limitations of their interpretations** (p. 1072, emphasis added).

The predictive accuracy data reported by Campbell and DeClue (2010a), and the data found in this article, obligate SVP evaluators, using the Static instruments, to identify their limitations. To do otherwise falls short of the relevant ethical obligations.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060-1073.

Bani-Yaghoub, M., Federoff, J. P., Curry, S., & Amundsen, D. E. (in press). A time series modeling approach to risk appraisal of violent and sexual recidivism. *Law and Human Behavior*.

Boccaccini, M. T., Murrie, D. C., Capterton, J. D., & Hawes, S. W. (2009). Field validity of the Static-99 and the MnSOST-R among sex offenders evaluated for civil commitment as Sexually Violent Predators. *Psychology, Public Policy and Law, 15*, 278-314.

Campbell, T. W. (February/March 2001). Actuarial scales and evidentiary reliability: Clarifying the issues. *Sex Offender Law Report*, *2(2)*, 19-20, 30-31.

Campbell, T. W. (2007). When prophecy fails: Retreating from prediction. *Journal of Sexual Offender Civil Commitment: Science and the Law, 2*, 1-11.

Campbell, T. W., & DeClue, G. (2010a). Maximizing predictive accuracy in sexually violent predator evaluations. *Open Access Journal of Forensic Psychology, 2*, 148-232.

Campbell, T. W., & DeClue, G. (2010b). Flying blind with naked factors: Problems and pitfalls in adjusted actuarial sex-offender risk assessment. *Open Access Journal of Forensic Psychology, 2*, 75-101.

DeClue, G., & Campbell, T. W. (2010). Still maximizing accuracy and accepting scrutiny in Sexually Violent Predator evaluations. *Open Access Journal of Forensic Psychology, 2*, 322-336.

Doren, D. M. (2000 Aug/Sept). Evidentiary issues, actuarial scales, and sex offender civil commitments. *Sex offender Law Report, 1(5)*, 65-66, 78-79.

Doren, D. M. (2006). Inaccurate arguments in sex offender civil commitment proceedings. In A. Schlank (Ed.), *The sexual predator - law and public policy - clinical practice*. Kingston, NJ: Civic Research Institute, Inc.

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.

Farrington, D., & Tarlington, R. (1985). *Prediction and criminology*. Albany, NY: State University of New York Press.

Hanson, R. K., Phenix, A., Helmus, L. (2009 Sept 30). *Static-99R and Static-2002R: How to interpret and report in light of recent research*. Pre-Conference workshop at the 28th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Dallas, TX (www.static99.org).

Hanson, R. K., & Howard, P. D. (2010) Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior, 34*, 275-281.

Helmus, L. (2009 September). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* (unpublished master's thesis). Department of Psychology, Carleton University, Ottawa, Ontario.

Phenix, A., Helmus, L., & Hanson, R. K. (2009 Nov 1). *Static-99R: Evaluators' Workbook*. www.static99.org

Prentky, R. A., Lee, A. F. S., Knight, R. A., & Cerce, D. (1997). Recidivism rates among child molesters and rapists: A methodological analysis. *Law and Human Behavior, 21*, 635-659.

Sreenivasan, S., Weinberger, L. E., Frances, A., & Cuswoth-Walker, S. (2010). Alice in actuarial-land: Through the looking glass of changing Static-99 norms. *Journal of the American Academy of Psychiatry and the Law, 38*, 400-406.

Thornton, D., & Helmus, L. (2009 Dec 7). *Static risk assessment: Understanding and applying the 2009 norms*. Presentation to the California Department of Mental Health.

Vrieze, S. I. & Grove, W. M. (2008). Predicting sex offender recidivism: I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior, 32*, 266-278.

Wilson, R. J., & Looman, J. (2010). What can we reasonably expect to accomplish in conducting actuarial risk assessments with sexual offenders in civil commitment settings? A response to Campbell and DeClue: "Maximizing predictive accuracy in sexually violent predator evaluations." *Open Access Journal of Forensic Psychology, 2,* 306-321.