Rethinking the Indefinite Detention of Sex Offenders

Fredrick E. Vars

Forthcoming, Volume 44 of the Connecticut Law Review

DRAFT---Forthcoming, Volume 44 of the *Connecticut Law Review*

RETHINKING THE INDEFINITE DETENTION OF SEX OFFENDERS

Fredrick E. Vars[†]

[†] Assistant Professor, University of Alabama School of Law (J.D., Yale; A.B., Princeton). I would like to thank Ian Ayres, David Becker, Montré Carodine, Judge Colquitt, Adam Cox, Shahar Dillbary, Ted Donaldson, Caroline Harada, David Kaye, Ron Krotoszynski, Grace Lee, Andy Morriss, Mike Pardo, Hong Min Park, David Patton, Pam Pierson, Hal Stern, Charlie Vars, and Richard Wollert for their help at various stages of this project. Michael Reeves and Jonathan Kolodziej provided excellent research assistance, as did Penny Gibson and the rest of the staff at the University of Alabama law library. Thanks also to Dean Ken Randall and the University of Alabama Law School Foundation for generous research support.

DRAFT---Forthcoming, Volume 44 of the *Connecticut Law Review*

RETHINKING THE INDEFINITE DETENTION OF SEX OFFENDERS

Abstract:

Thousands of sex offenders in the United States are being held indefinitely under civil commitment programs. The analysis in this Article suggests that none (or precious few) belong there. Specifically, in a large dataset, an instrument as good as the one most widely used by experts (the "Static-99") could not identify even one sex offender who met the legal standards for commitment. Supplementing such instruments with additional information appears not to improve matters, so the failure of the instrument is profoundly disturbing.

There are three possible responses: (1) improve instruments to meet existing standards; (2) lower the standards; and (3) abandon sex offender civil commitment. This Article focuses on the first response, identifying and correcting flaws in the most widely-used instrument. But the greater significance of the Article is to reframe the debate around the other two potential responses. Can we predict the future well enough to justify the indefinite detention of "dangerous" people?

TABLE OF CONTENTS

> [F]rom a legal point of view there is nothing inherently unattainable about a prediction of future criminal conduct.
> -*Schall v. Martin, 467 U.S. 253, 278 (1984)*
>
> Prediction is very difficult, especially about the future.
> -*Neils Bohr (Danish Physicist)*

Introduction

Preventive detention to protect public safety is an old idea, which has recently expanded in scope.[1] Two well-established examples are mental illness and contagious disease.[2] A more dramatic example is the now widely discredited internment of Japanese Americans during World War II.[3] As the internment case well illustrates, the great difficulty in preventive detention is accurately identifying truly dangerous individuals. This Article focuses on that problem in an area where the assessment of future dangerousness is particularly rigorous: sex offender civil commitment. Even here the prediction problem may be intractable.

Thousands of convicted sex offenders remain in custody after their prison terms expire.[4] As of March 2010, twenty states and the federal government had laws authorizing the civil commitment of certain sex offenders.[5] Committed individuals are rarely released.[6] The United States Supreme Court has upheld such laws against various constitutional challenges.[7] A

[1] Paul H. Robinson, *Punishing Dangerousness: Cloaking Preventive Detention as Criminal Justice*, 114 HARV. L. REV. 1429, 1429-31, 1447-49 (2001).

[2] *Id.* at 1444.

[3] *See* Hamdi v. Rumsfeld, 542 U.S. 507, 542 (2004) (Souter, J., concurring) ("[T]he Emergency Detention Act of 1950 . . . was repealed in 1971 out of fear that it could authorize a repetition of the World War II internment of citizens of Japanese ancestry; Congress meant to preclude another episode like the one described in Korematsu v. United States, 323 U.S. 214 (1944).").

[4] *See* Monica Davey & Abby Goodnough, *Doubts Rise as States Hold Sex Offenders After Prison*, N.Y. TIMES, Mar. 4, 2007, at A1 ("About 2,700 pedophiles, rapists and other sexual offenders are already being held indefinitely, mostly in special treatment centers, under so-called civil commitment programs . . . .").

[5] Keith Matheny, *Releases of Sexually Violent Predators Anger Local Areas*, USA TODAY (Mar. 4, 2010), *available at* http://www.usatoday.com/news/nation/2010-03-03-predator-housing_N.htm (accessed July 14, 2010).

[6] John Q. La Fond, *The Costs of Enacting a Sexual Predator Law and Recommendations for Keeping Them From Skyrocketing*, *in* PROTECTING SOCIETY FROM SEXUALLY DANGEROUS OFFENDERS: LAW, JUSTICE, AND THERAPY 288 (Bruce J. Winick & John Q. LaFond, eds., 2003).

[7] Kansas v. Crane, 534 U.S. 407 (2002); Kansas v. Hendricks, 521 U.S. 346 (1997).

common requirement for civil commitment is future dangerousness.[8] The two primary methods of determining the likelihood of recidivism are clinical judgment and so-called actuarial risk assessment instruments (“ARAIs”).[9] Studies have shown ARAIs to be more accurate.[10] The most widely used ARAI is the Static-99,[11] which will be the focus of this Article.

The Static-99 is a ten-item instrument. The coding form is shown in Table 1:[12]

[8] *E.g.*, *Hendricks*, 521 U.S. at 357 (“Commitment proceedings can be initiated only when a person ‘has been convicted of or charged with a sexually violent offense,’ and ‘suffers from a mental abnormality or personality disorder which makes the person likely to engage in the predatory acts of sexual violence.’”) (quoting Kan. Stat. Ann. § 59-29a02(a) (1994)); *see generally* Robert Prentky et al., *Sexually Violent Predators in the Courtroom*, 12 Psych., Pub. Pol’y & L. 357, 358 (2006).

[9] Debra A. Pinals, Chad E. Tillbrook, & Denise L. Mumley, *Violent Risk Assessment*, *in* Sex Offenders: Indentification, Risk Assessment, Treatment, and Legal Issues (Fabian M. Saleh et al., eds., 2009). Some advocate clinically adjusted actuarial assessments. *Id.* at 55. Others strongly disagree. Stephen D. Hart, Christine Michie, & David J. Cooke, *Precision of Actuarial Risk Assessment Instruments: Evaluating the ‘Margins of Error’ of Group v. Individual Predictions of Violence*, 190 (Supp. 49) British J. Psychiatry s60, s64 (2007).

[10] Pinals, Tillbrook, & Mumley, *supra* note 9, at 54; Marcus T. Boccaccini et al., *Field Validity of the Static-99 and MnSOST-R among Sex Offenders Evaluated for Civil Commitment as Sexually Violent Predators*, 15 Psych., Pub. Pol’y, & L. (2009) (“ARAIs designed to predict sexual reoffense (d = .67) clearly outperformed unstructured professional judgment (d = .42).”). *But see* Thomas R. Litwack, *Actuarial Versus Clinical Assessments of Dangerousness*, 7 Psych., Pub. Pol’y & L. 409 (2001).

[11] Jacqueline Waggoner, Richard Wollert, & Elliot Cramer, *A Respecification of Hanson’s Updated Static-99 Experience Table That Controls for the Effects of Age on Sexual Recidivism Among Young Offenders*, 7 L., Probability & Risk 305, 305-06 (2008); Rebecca L. Jackson & Derek T. Hess, *Evaluation for Civil Commitment of Sex Offenders: A Survey of Experts*, 19 Sex Abuse 425, 434, 438, 440 (2007).

[12] From R. Karl Hanson & David Thornton, *Improving Risk Assessment for Sex Offenders: A Comparison of Three Actuarial Scales*, 24 L. & Human Behav. 119, 133-35 appdx.I (2000) and www.static99.org (downloaded Apr. 2010).

Table 1. Static-99 Coding Form

Question	*Risk Factor*	*Codes*		*Score*
1	Young	Aged 25 or older		0
		Aged 18 – 24.99		1
2	Ever Lived With	Ever lived with lover for at least two years?		
		Yes		0
		No		1
3	Index Non-Sexual Violence-Any Convictions	No		0
		Yes		1
4	Prior Non-Sexual Violence-Any Convictions	No		0
		Yes		1
5	Prior Sex Offenses	Charges	Convictions	
		None	None	0
		1-2	1	1
		3-5	2-3	2
		6+	4+	3
6	Prior Sentencing Dates (Excluding Index)	3 or less		0
		4 or more		1
7	Any Convictions for Non-Contact Sex Offenses	No		0
		Yes		1
8	Any Unrelated Victims	No		0
		Yes		1
9	Any Stranger Victims	No		0
		Yes		1
10	Any Male Victims[13]	No		0
		Yes		1
	Total Score	Add up scores from individual risk factors		

Translating Static-99 Scores into Risk Categories

Score	Label for Risk Category
0, 1	Low
2, 3	Moderate-Low
4, 5	Moderate-High
6 plus	High

Risk categories can be further translated into recidivism rates based on figures from the sample used to develop the instrument:[14]

[13] The Static-99 is designed for male offenders only. The disparate impact of the Static-99 on homosexual and bisexual offenders is beyond the scope of this Article.

[14] Estimates based on Hanson & Thornton, *supra* note 12, at 129 tbl.5.

Table 2. 15-Year Recidivism By Static-99 Risk Category

Risk Category	*Sample Size*	*Sexual*	*Violent*
Low (0, 1)	257	0.09	0.16
Medium-Low (2, 3)	410	0.18	0.32
Medium-High (4, 5)	290	0.37	0.52
High (6+)	129	0.52	0.59
Total (avg = 3.2)	1086	0.26	0.37

For example, an individual with a score of six or higher on the Static-99 would have a predicted 15-year sexual recidivism rate of 52%.

Notably rare in the vast literature examining the Static-99 are studies addressing the fundamental, bottom-line question: can the Static-99 identify individuals who meet the legal standards for commitment?[15] The tests presented in this Article help fill that important gap. The profoundly disturbing answer—given the central role the Static-99 has played in the commitment of thousands of individuals—is essentially no. That answer calls into serious question the entire enterprise of sex offender commitment. And, at a minimum, it demands improvement or replacement of the Static-99.

The general approach of this Article is to develop an instrument that predicts recidivism roughly as well as the Static-99 (and is better in other respects), then to ask whether that instrument can identify individuals who qualify for sex offender civil commitment under the existing legal standards. Part I situates the present study within the literature and outlines three

[15] *But see* Richard Wollert, *Low Base Rates Limit Expert Certainty When Current Actuarials Are Used to Identify Sexually Violent Predators: An Application of Bayes's Theorem*, 12 PSYCH., PUB. POL'Y, & L. 56 (2006) ("[T]he best available risk-assessment method (i.e., actuarial testing) eventually points to the conclusion that the recidivism rate for each detainee . . . does not meet the commitment standard."). Two other studies that come closest are: Hart, Michie, & Cooke, *supra* note 9 (approximation using social science, not legal, standard), and Eric S. Janus & Paul E. Meehl, *Assessing the Legal Standard for Prediction of Dangerousness in Sex Offender Commitment Proceedings*, 3 PSYCH., PUB. POL'Y, & L. 33, 40, 60 (1997) (relying on assumptions rather than individual-level data).

problems with the Static-99.[16] Part II reports the results of a new model created and tested in a large dataset.

First, as the creators of the Static-99 have come to recognize, "the original Static-99 did not sufficiently account for age at release."[17] The creators have proposed a fix. This Article in Part II outlines a better one using a 15-state dataset: allowing for age effects throughout the range rather than using arbitrary cut-offs.

The second shortcoming, also acknowledged by the originators, is that sexual recidivism rates have fallen since the norms were established. New, lower norms are needed (Part I). In Part II, this Article provides one more data point in support of that conclusion and suggests an approach that would more seamlessly adjust to changing crime rates: updating the instrument as soon as new data become available.

Third, and most fundamentally, the Static-99, even as modified, fails to report uncertainty in predicted recidivism rates that is essential to determine whether an individual sex offender meets the commitment threshold according to the applicable standard of proof. Part I explains this problem and classifies each jurisdiction with sex offender civil commitment according to commitment standard and standard of proof. This Article's alternative prediction model in Part II quantifies the effect of failing to account for prediction error and demonstrates how this shortcoming may (or may not) be overcome. The technical solution is relatively straightforward; the resulting problem is that essentially no one qualifies for commitment. The most important finding of this Article is that an instrument as good as the Static-99 largely fails to identify any individuals who met the standards for commitment. There may be nothing "inherently

[16] There are others. *E.g.*, Melissa Hamilton, *Public Safety, Individual Liberty, and Suspect Science: Future Dangerousness Assessments and Sex Offender Laws*, __ TEMPLE L. REV. ___ (forthcoming), at http://ssrn.com/abstract=1580016. *See generally* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE (2007).

[17] Leslie Helmus et al., Static-99R: Revised Age Weights 6 (Oct. 5, 2009) (downloaded from www.static99.org).

unattainable" about predicting future behavior, but in this corner of the real world at least it is more difficult than present practice admits.

Part III discusses limitations of the present study—most notably, a short follow-up period—and implications beyond sex offender civil commitment. Actuarial risk assessment is pervasive, particularly in the area of criminal justice.[18] The Static-99 has perhaps more empirical grounding than other widely used instruments, and still it appears to fall short. Similarities to one other very popular instrument, the LSI-R, used for parole and other purposes, are highlighted.

I. THREE PROBLEMS WITH THE STATIC-99

A. Age

Older people commit less crime. There is a long-recognized inverse relationship between age and recidivism generally.[19] The review of sexual recidivism that led to the Static-99 listed young age as a factor.[20] And the Static-99 does account for young age. As shown in Table 1 above, individuals less than 25 years old receive an additional point. However, many have criticized the Static-99 for failing to account for age throughout the lifespan.[21] Substantial empirical evidence has accumulated showing that the risk of sexual recidivism declines with age well above 25.[22] The general conclusion has been that "recidivism rates decrease in a linear fashion with age-at-release."[23]

[18] HARCOURT, *supra* note 16, at 2.

[19] JOHN MONAHAN, PREDICTING VIOLENT BEHAVIOR: AN ASSESSMENT OF CLINICAL TECHNIQUES 32, 107-08 (1981); Robinson, *supra* note 1, 1451.

[20] R. Karl Hanson & Monique T. Bussière, *Predicting Relapse: A Meta-Analysis of Sexual Offender Recidivism Studies*, 66 J. CONSULTING & CLINICAL PSYCHOLOGY 348, 351 (1998).

[21] *E.g.*, TERENCE W. CAMPBELL, ASSESSING SEX OFFENDERS: PROBLEMS AND PITFALLS 76 (2004). This criticism applies to all of the five most commonly used ARAIs. Howard E. Barbaree & Ray Blanchard, *Sexual Deviance Over the Lifespan: Reductions in Deviant Sexual Behavior in the Aging Sex Offender*, *in* SEXUAL DEVIANCE: THEORY, ASSESSMENT, AND TREATMENT 38 (D. Richard Lewis & William T. O'Donohue, eds. 2008).

[22] Howard E. Barbaree & Ray Blanchard, *Sexual Deviance Over the Lifespan: Reductions in Deviant Sexual Behavior in the Aging Sex Offender*, *in* SEXUAL DEVIANCE: THEORY, ASSESSMENT, AND TREATMENT 37-60 (D.

The largest field validity test to date found that the Static-99 was not a significant predictor of violent or sexually violent recidivism after controlling for age at release, prior arrests, and release type (mandatory supervision versus discharge).[24] Of particular relevance for present purposes, age was a highly significant predictor—better than the Static-99.[25] On the other hand, when the analysis was limited to sexually violent recidivism, age was not a significant predictor and Static-99 score was.[26] It should be noted that release type, which was highly significant in both analyses, was apparently determined in part by Static-99 score.[27] Static-99 score also appears to have been used as a screening device elsewhere in the process.[28] Using the Static-99 in both these ways would tend to artificially reduce its observed predictive power. Still, the overall finding is that age tends to add predictive power beyond the Static-99.

Hanson and Thornton (with two co-authors) responded to this growing evidence on the Static-99 website.[29] First, they confirmed through regression analysis that age was a significant predictor of recidivism even while controlling for Static-99 score.[30] Second, they formulated a new scoring system for age—specifically, ages 18-34.9, +1 point; 35-39.9, 0 points; 40-59.9, -1 point; and 60 or older, -3 points.[31] Third, they reran the regressions using the modified Static-99 scores. Age was no longer statistically significant,[32] leading Hanson and Thornton to conclude

Richard Lewis & William T. O'Donohue, eds. 2008); LEAM A. CRAIG, KEVIN D. BROWNE, & ANTHONY R. BEECH, ASSESSING RISK IN SEX OFFENDERS: A PRACTITIONER'S GUIDE 62-67 (2008); Wollert, *supra* note 15, at 70 tbl.1.

[23] Howard E. Barbaree, Ray Blanchard, & Calvin M. Langton, *The Development of Sexual Aggression through the Life Span: The Effect of Age on Sexual Arousal and Recidivism among Sex Offenders*, 989 ANN. N.Y. ACAD. SCI. 59, 67 (2003). *Accord* Leam A. Craig, *The Effect of Age on Sexual and Violent Recidivism*, XX(X) INT'L J. OFFENDER THERAPY & COMPARATIVE CRIMINOLOGY 1, 10 (2009).

[24] Boccaccini et al., *supra* note 10, at 298 tbl.4.

[25] *Id.*

[26] *Id.*

[27] *Id.* at 292 & tbl.2.

[28] *Id.* at 305.

[29] Helmus et al., *supra* note 17.

[30] *Id.* at 2.

[31] *Id.* at 4.

[32] *Id.* at 4.

that "the original Static-99 did not sufficiently account for age at release, whereas the revised scale [Static-99R] does."[33]

This response is unsatisfactory. To be sure, four age categories are better than two, but the real question is why categorize at all? Why not just include age as a continuous variable and let the regression equation assign it the weight that optimizes the model's predictive power? That is the approach of this Article: to propose a methodology rather than a universal solution. Presumably, the reason Hanson and Thornton resist this approach is attachment to the notion that their instrument needs to be simple enough to be performed with pencil, paper, and no calculator. But Hanson and Thornton have already elsewhere suggested movement away from this model: a computerized coding form could eliminate logical and arithmetic errors, they observe.[34] Computerization could just as easily eliminate the conceptual error of applying crude actuarial methods rather than more powerful statistical techniques like logistic regression.

Failing to use age reasonably is arguably unconstitutional.[35] Due process dictates that a police "officer may not choose to ignore information that has been offered to him or her."[36] This does not translate into a constitutional duty to investigate, but it does entail a duty not to turn a blind eye to relevant evidence.[37] The same principle should apply to adjudicating sex offender civil commitments. "[R]equirements of notice and hearings are of little significance if the

[33] *Id.* at 6. A more sensitive four-age-category approach is outlined in Richard Wollert et al., *Recent Research (N=9,305) Underscores the Importance of Using Age-Stratified Actuarial Tables in Sex Offender Risk Assessments*, 22 SEXUAL ABUSE: A JOURNAL OF TREATMENT & RESEARCH 471 (2010).

[34] R. Karl Hanson, Leslie Helmus, & David Thornton, *Predicting Recidivism Amongst Sexual Offenders: A Multi-site Study of Static-2002*, 34 L. & HUMAN BEHAV. 198, 208 (2010).

[35] Others have argued that use of instruments like the Static-99 with older offenders "could be considered to be discriminatory." Prentky et al., *supra* note 8, at 376.

[36] Kingsland v. City of Miami, 382 F.3d 1220, 1229 (11th Cir. 2004).

[37] Logsdon v. Hains, 492 F.3d 334, 341 (6th Cir. 2007).

decisionmaker ultimately ignores any information before it."[38] The Static-99, both the original and revised versions, effectively throws out relevant information by lumping individuals into broad age categories.

One response to this argument is that the Static-99 is not the only piece of evidence considered, and the Constitution generally constrains the total package, not each constituent part. Decisionmakers are free to factor age into the equation notwithstanding its inclusion in the instrument. That may be true, but it is almost certainly not the case that decisionmakers' informal consideration of age always accurately reflects the true impact of age on recidivism.[39] Of course, the government need not wait for a perfect instrument,[40] but using one that throws away obviously relevant information seems irrational. A second response is that expert testimony based on the Static-99 may not be state action. The nuances of the state action doctrine are outside the scope of this article, but there is at least one state where use of the Static-99 is unequivocally state action: Virginia requires its use by statute.[41]

B. Norms

The recidivism rates reported in Table 2 above were derived from prisoners released from three penal institutions in Canada and one in the United Kingdom.[42] This should immediately give pause to those who would rely on Table 2 to estimate the likelihood of recidivism for a prisoner in the United States because "the rate of sexual assault in Canada . . . is more than twice that of the United States."[43] Moreover, the prisoners in the normative sample were released

[38] Mark W. Cordes, *Policing Bias and Conflicts of Interest in Zoning Decisionmaking*, 65 N.D. L. REV. 161, 217 (1989) (citing Martin H. Redish & Lawrence C. Marshall, *Adjudicatory Independence and the Values of Procedural Due Process*, 95 YALE L.J. 455, 476 (1986)).

[39] *See infra* notes 145-48 and accompanying text.

[40] *Cf.* Boccaccini et al., *supra* note 10, at 306.

[41] Va. Code Ann. § 37.2-903; *see also* West's Ann. Cal. Penal Code § 290.04(b)(1).

[42] Hanson & Thornton, *supra* note 12, at 123-24.

[43] John A. Fennel, *Punishment by Another Name: The Inherent Overreaching in Sexually Dangerous Person Commitments*, 35 NEW ENG. J. CRIM. & CIV. CONFINEMENT 37, 59 (2009).

between the late 1950s and early 1990s. Crime, including sexual offenses, peaked in the early 1990s and has been declining since then.[44]

Given these facts, it should not be surprising that studies have generally found recidivism rates below the Static-99 normative levels.[45] Hanson, Thornton, and Leslie Helmus, in more recent and diverse samples, found that "sexual recidivism was two-thirds (66%) the rate of the original sample."[46] There were significant differences among the samples. So rather than simply adjust downward the recidivism rates based on the overall results, the authors provided two estimates for each recidivism type and time period: one lower number for "routine" samples and a higher number for "preselected high risk" samples.[47] Evaluators are advised to report both the low- and high-end values, then to exercise judgment in opining which sample the individual more closely resembles.[48] However, the authors concede that the preselection factors that would place an individual into one of the two categories "are not fully known and would vary across samples."[49]

This "New Norms" article has called into serious doubt use of the Static-99. In *State v. Rosado*,[50] the sex offender respondent wanted to introduce his Static-99 score of 4 in the civil commitment proceedings. The court granted the state's motion *in limine* to exclude the

[44] Leslie Helmus, R. Karl Hanson, & David Thornton, *Reporting Static-99 in Light of New Research on Recidivism Norms*, 21 THE FORUM 38, at 38 (Winter 2009).

[45] Boccaccini et al., *supra* note 10, at 304; *see also* PATRICK A. LANGAN, ERICA L. SCHMITT, & MATTHEW R. DUROSE, RECIDIVISM OF SEX OFFENDERS RELEASED FROM PRISON IN 1994 (2003) at www.ojp.usdoj.gov/bjs/pub/pdf/rsorp94.pdf (5.3% of released sex offenders were rearrested within 3 years for a sex crime); *cf.* Shoba Sreenivasan et al., *Predicting the Likelihood of Future Sexual Recidivism: Pilot Findings from a California Sex Offender Risk Project and Cross-Validation of the Static-99*, 35 J. AM. ACAD. PSYCHIATRY L. 454, 465 (2007) (Static-99 underpredicted recidivism at scores of 2 and 3 and overpredicted at 4-6).

[46] Helmus, Hanson, & Thornton, *supra* note 44, at 39. *But see* Grant T. Harris & Marnie Rice, *Characterizing the Value of Actuarial Violence Risk Assessments*, 34 CRIM. JUSTICE & BEHAV. 1638, 1643 (2007) (finding no support for and one counter-example in five sources cited for the proposition that recidivism rates fell along with overall crime rates).

[47] Helmus, Hanson, & Thornton, *supra* note 44, at 41 tbls.1, 2. Presumably reflecting the larger sample size, the authors also reported values for each Static-99 score up to 10+. *Id.*

[48] *Id.* at 40.

[49] *Id.*

[50] 889 N.Y.S.2d 369 (2009).

evidence. One of the reasons was the article described above: "in view of the recent development of the new norms, and an entirely new and undeveloped methodology for applying those norms, it cannot be said that the new norms of the STATIC-99 (despite its past acceptance) are now sufficiently understood and accepted in the relevant scientific community under *Frye* . . . ."[51]

The Static-99 appears caught between a rock and a hard place. Fail to account for changes in recidivism rates and generate flawed estimates, or adjust and be excluded because the new adjustment is not yet generally accepted in the scientific community. "[T]he development of ARA [actuarial risk assessment], like all good science, is evolutionary."[52] One way for the Static-99 to evolve is suggested by a section of the "New Norms" article that the *Rosado* court did not mention. In it, the authors provide relative risk values for each Static-99 score up to 9+.[53] Applying those values to the actual recidivism rate for the relevant population would be one reasonable way to generate estimates, at least until there is time for a validation study in that population.[54]

This Article will outline a methodology that nearly automatically adjusts to changes not just in the overall recidivism rate, but in the relative predictive power of included variables. Because the methodology is constant, it would arguably not be subject to challenges like the one that succeeded in *Rosado*.

[51] *Id.* at 416. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923), states the standard for the admissibility of scientific evidence in New York.

[52] Eric S. Janus & Robert A. Prentky, *Forensic Use of Actuarial Risk Assessment With Sex Offenders: Accuracy, Admissibility and Accountability*, 40 AM. CRIM. L. REV. 1443, 1445 (2003).

[53] Helmus, Hanson, & Thornton, *supra* note 44, at tbl.3.

[54] *See* Calvin M. Langton et al., *Reliability and Validity of the Static-2002 Among Adult Sexual Offenders with Reference to Treatment Status*, 34 CRIMINAL JUSTICE & BEHAV. 616, 638 (2007) ("Clearly, likelihood ratios require examination before recidivism probabilities associated with any risk assessment instrument's scores for one population are assumed to apply to another population.").

C. Error and Standards of Proof

> If a risk scale is to be used in applied contexts, then it is important that the degree of predictive accuracy is sufficient to inform rather than mislead. Critics could suggest, for example, that a correlation in the 0.30 range is insufficient for decision making because it accounts for only 10% of the variance. Even if such an argument was [sic] correct . . . , most decision makers are not particularly concerned about "percent of variance accounted for." Instead, applied risk decisions typically hinge on whether offenders surpass a specified probability of recidivism (e.g., >50%).[55]

So wrote Karl Hanson and David Thornton in 2000 reporting the results of an early test of the Static-99, which did in fact show correlations around 0.30.[56] Their statement may be true for certain low-stakes decisions, but it is frighteningly flawed with respect to sex offender civil commitment, as explained below.

In the same article, Hanson and Thornton reported 15-year sexual recidivism above their hypothetical 50% threshold for the Static-99 "High" risk category (52%) (*see supra* Table 2).[57] The implication is that commitment would be proper for individuals in that risk category in jurisdictions applying a 50% threshold. That is false. The question is how sure are we that an individual in this risk category is more likely than not to reoffend. The percentage of variance accounted for (and hence prediction error) is absolutely critical in making that determination. Hanson and Thornton missed the distinction—explained more than 20 years earlier in the same journal by John Monahan and David Wexler—between the standard of commitment and the

[55] Hanson & Thornton, *supra* note 12, at 129-30. A correlation of 1.0 means a perfect positive fit between the predictor and outcome variables; 0 indicates no relationship. The square of the correlation coefficient is the percentage of variance (or spread) of the outcome variable accounted for by the predictor. Thus, as the quoted passage states, a 0.30 correlation coefficient corresponds to roughly 10% of variance accounted for (0.30^2 = 9%).
[56] Id. at 126 tbl.4.
[57] Id. at 129 tbl.5.

standard of proof: "one must prove to a given standard [of proof] only that a specified probability threshold [*viz.*, commitment standard] has been crossed."[58]

If the standard of proof in civil commitment were merely a "preponderance of the evidence" ("POE"; "more likely than not" or 50%), then, assuming symmetric error, the distinction would not be important.[59] But the United States Supreme Court in *Addington v. Texas* held that due process requires a higher standard for civil commitment.[60] The "clear and convincing evidence" ("CCE") standard was found to be sufficient, though perhaps not required. That standard has been interpreted as proof with greater than 75% confidence.[61] If the commitment standard is greater than 50%, then the question is whether it is 75% likely that an individual's risk of recidivism is above that threshold.[62] The 52% value in the Static-99 recidivism table may or may not be sufficient evidence of that fact, but given the modest correlation with recidivism (0.30), that would seem quite unlikely.

[58] John Monahan & David B. Wexler, *A Definite Maybe: Proof and Probability in Civil Commitment*, 2 L. & HUMAN BEHAV. 37, 38 (1978). *See also* M. Neil Browne & Ronda R. Harrison-Spoerl, *Putting Expert Testimony in Its Epistemological Place: What Predictions of Dangerousness in Court Can Teach Us*, 91 MARQ. L. REV. 1119, 1207-10 (2008) (recognizing the significance of the standard of proof). However, Browne & Harrison-Spoerl would simply multiply the commitment threshold by the standard of proof, *id.* at 1209 n.429, which is inappropriate as explained in the text below.

[59] "[S]tatistical decision theory" sometimes leads non-lawyer experts in this area to ignore the heightened standard of proof. D. Mossman & T. Sellke, *Avoiding errors about 'margins of error,'* 192 BRITISH J. PSYCHIATRY 561 (2007) (correspondence).

[60] 441 U.S. 418 (1979). This Article follows legislatures and courts in assuming that *Addington* applies to sex offender civil commitment.

[61] C.M.A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence or Constitutional Guarantees?*, 35 VAND. L. REV. 1293, 1328 tbl.5 (1982) (survey of 170 federal judges reported a mean, median, and mode of 0.75 for the clear and convincing standard); *see also* United States v. Fatico, 458 F. Supp. 388, 410 tbl. (E.D.N.Y. 1978) (reporting a range of 0.6 to 0.75 in survey of eight federal district judges); *see generally* Fredrick E. Vars, *Toward a General Theory of Standards of Proof*, 60 CATHOLIC UNIV. L. REV. 1 (2010). Quantification is resisted by many. *See, e.g.*, Kevin M. Clermont, *Procedure's Magical Number Three: Psychological Bases for Standards of Decision*, 72 CORNELL L. REV. 1115, 1147-48 (1987). That resistance should be somewhat attenuated in this context. For better or worse, commitment thresholds are framed in probabilistic terms, *see infra* Table 3, and most of the evidentiary work is performed by probabilistic or "actuarial" instruments.

[62] At least this is the way that sex offender civil commitment statutes are in fact structured: with separate commitment and proof standards. That is not the only possible reading of *Addington*. Arguably, setting the commitment standard below 75% violates *Addington*. Nicholas Scurich & Richard John, *The Normative Threshold for Psychiatric Civil Commitment*, 50 JURIMETRICS J. 425, 448-51 (2010).

Since 2000 Hanson and Thornton have not been entirely deaf to this criticism.[63] Their revised age-specific recidivism risk tables included, for the first time, 95% confidence intervals ("CIs").[64] In other contexts, however, they continue to omit critical error estimates. The Static-2002 is a refinement of the Static-99. In a leading multi-site study of the Static-2002, Hanson, Thornton, and a co-author once again report recidivism rates for each score, omitting CIs.[65]

Reporting CIs (when they do) is a substantial improvement over the original reports. But it falls short of answering the key question. The CIs reported by Hanson and Thornton are group intervals; the legally relevant statistic is the individual interval. In other words, the Static-99 creators are telling us how confident we can be that the recidivism rate (e.g., 52%) accurately reflects the rate for the group of individuals in this risk category (group). A civil commitment decisionmaker needs to know how likely it is the individual before it meets the commitment standard (individual).

Stephen Hart and colleagues have estimated an individual 95% CI on the 52% reported recidivism rate of between 6% and 95%.[66] In other words, if we had a large sample of individuals in the "High" Static-99 risk category, 95% of them would have a recidivism risk somewhere between 6 and 95 percent. Some have concluded from this statistic that actuarial risk assessment, and perhaps sex offender commitment generally, should be eliminated.[67] That may well be the correct conclusion, but the wide confidence interval alone does not decide the issue.

[63] *See, e.g.*, Janus & Prentky, *supra* note 52, at 1471 (pointing out "absence of information on standard errors").

[64] Waggoner, Wollert, & Cramer, *supra* note 11, at 310-11. As explained later in the text, a confidence interval is another, and more useful, measure of the precision of prediction.

[65] Hanson, Helmus, & Thornton, *supra* note 34, at 210 tbl.7.

[66] Hart, Michie, & Cooke, *supra* note 9, at s62 tbl.2; *see also* David J. Cooke & Christine Michie, *Limitations of Diagnostic Precision and Predictive Utility in the Individual Case: A Challenge for Forensic Practice*, 34 LAW & HUM. BEHAV. 259 (2010).

[67] Fennel, *supra* note 43, at 39, 56, 61. Some suggest that the imprecision of ARAIs may render them inadmissible under the standards for expert or scientific evidence. Hart, Michie, & Cooke, *supra* note 9, at s64. Others contend that ARAIs clear present admissibility hurdles. *E.g.*, Christopher Slobogin, *Dangerousness and Expertise Redux*, 56

First and fundamentally, there is arguably lack of equivalence between the statistical concept of a confidence interval and the legal concept of a standard of proof.[68] As Professor David Kaye observes, "the confidence interval is not the probability that [a parameter] lies within the lonely interval we observed. Rather, it is the long run frequency with which various and varied CIs would cover the unknown value for [the parameter]."[69] (This connection to frequency is why the approach used to generate CIs is called "frequentist.") Two additional facts arguably are needed to bridge the divide: (1) the probability prior to the subject evidence; and (2) the probability of the evidence under the alternative hypothesis.[70] (These facts are terms in Bayes's Theorem.) However, the large sample size in this study ($n \approx 9000$) suggests convergence between the frequentist logit CIs presented and the likely results of the alternative, Bayesian methodology.[71] Thus, although recognizing valid criticisms, I equate CIs with standards of proof,[72] at least as a heuristic device.[73]

EMORY L.J. 275 (2006); Janus & Prentky, *supra* note 52. Even if ARAIs are admissible, their imprecision obviously goes to weight and whether dangerousness has been established with the requisite certainty.

[68] Turpin v. Merrell Dow Pharmaceuticals, Inc., 959 F.2d 1349, 1353 n.1 (6th Cir. 1992) ("The confidence interval is not a 'burden of proof' in the legal sense; rather, it is a common sense mechanism upon which statisticians rely to confirm their findings and to lend persuasive power within their profession."); Nicholas Scurich & Richard S. John, *A Bayesian Approach to the Group Versus Individual Prediction Controversy in Actuarial Risk Assessment*, __ LAW & HUMAN BEHAV. ___ (2011).

[69] D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 CORNELL L. REV. 54, 62 (1987). *See generally* "Credible interval," at http://en.wikipedia.org/wiki/Credible_interval (visited Jan. 28, 2011).

[70] David H. Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS. 13, 23 (1983). *See generally* "Bayes' theorem," at http://en.wikipedia.org/wiki/Bayes'_theorem (visited Jan. 28, 2011).

[71] Kaye, *supra* note 69, at 69-70; "Credible interval," *supra* note 69. *See* M.J. Bayarri & J.O. Berger, *The Interplay of Bayesian and Frequentist Analysis*, 19 STATISTICAL SCIENCE 58, 71 (2004) ("Bayesian and frequentist asymptotic answers are often (but not always) the same."); Gauri Sankar Datta et al., *Bayesian Prediction with Approximate Frequentist Validity*, 28 ANNALS OF STATISTICS 1414, 1414 (2000) ("It is . . . shown that, for any given prior, it may be possible to choose an interval whose Bayesian predictive and frequentist coverage probabilities are asymptotically matched.").

For an example of the Bayesian approach in this context, see Andreas Mokros et al., *Assessment of Risk for Violent Recidivism Through Multivariate Bayesian Classification*, 16 PSYCHOL. PUB. POL'Y & L. 418 (2010).

[72] Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385 (1985). There are other non-Bayesian alternatives as well. *E.g.*, Majid Bani-Yaghoub et al., *A Time Series Modeling Approach in Risk Appraisal of Violent and Sexual Recidivism*, 34 LAW & HUM. BEHAV. 349 (2010).

[73] Neil B. Cohen, *Conceptualizing Proof and Calculating Probabilities: A Response to Professor Kaye*, 73 CORNELL L. REV. 78, 93 (1987) ("I believe that the confidence interval analogy performs well as a heuristic for the decisionmaking process."). It is worth noting that even Professor Cohen's most outspoken critic on this point,

Second, as Hart et al. conceded[74] and others further elaborated,[75] their methodology for estimating CIs had shortcomings. Rehearsing those criticisms here would serve little point, as none of them apply to the CIs calculated in this Article. Indeed, Hart et al. responded to their critics by explaining that they could have used better methodology if ARAIs were based on logistic regression rather than actuarial methods and Static-99 data were made publicly available.[76]

Finally, ninety-five is the conventional spread in social science, but the standard of proof in this context requires a different statistic. Jurisdictions are split between requiring proof beyond a reasonable doubt ("BRD") and proof by CCE.[77] As noted above, the latter standard corresponds roughly to 75% certainty. BRD is also generally not quantified by courts, but a survey of 171 judges yielded a mean, median, and mode of 90%.[78]

Shifting to a 75% or 90% CI, however, is not the right way to operationalize these standards of proof. Because the commitment question is whether an individual has a risk *greater* than a given threshold, only one tail of the error distribution matters. Assuming symmetric error, applying the CCE standard (75%) therefore requires calculation of the 50% CI: 25% of the error

Professor Kaye, favors the use of interval estimates, just not as equivalents to standards of proof. D.H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1363-64 (1986).

In footnote 138 *supra*, I also report results using false positive to false negative error ratios to determine cut-scores. *See* Cohen, *supra* note 72, at 417; Kaye, *supra* note 69, at 68-69.

[74] Hart, Michie, & Cooke, *supra* note 9.

[75] Harris & Rice, *supra* note 46, at 1648; Grant T. Harris, Marnie E. Rice, & Vernon L. Quinsey, *Shall evidence-based risk assessment be abandoned?*, 192 BRITISH J. PSYCHIATRY 154 (2008) (correspondence); Mossman & Sellke, *supra* note 59, at 561.

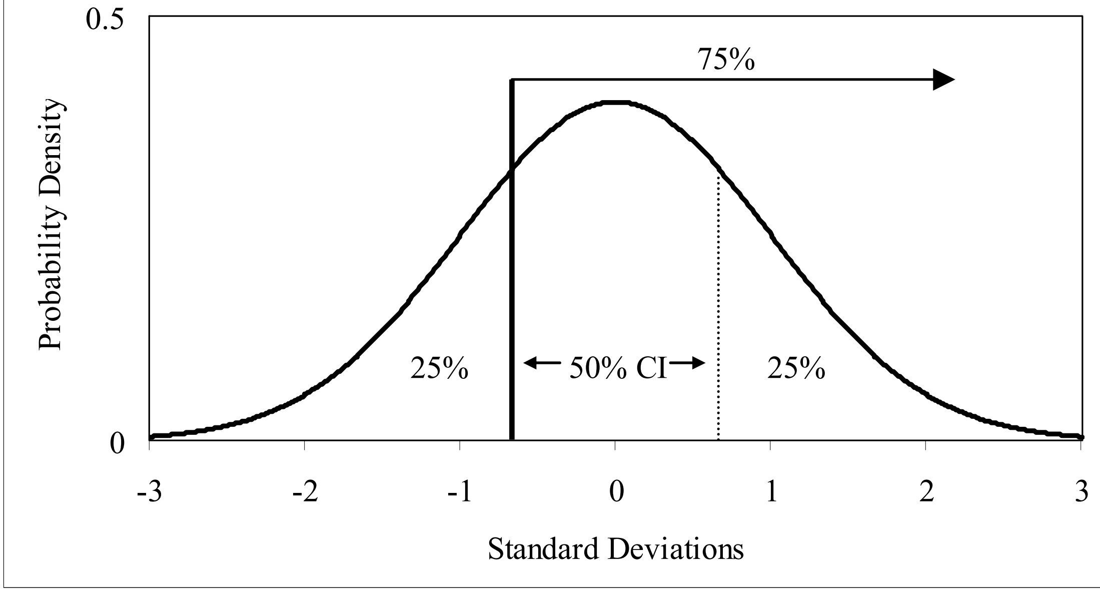
[76] S.D. Hart, C. Michie, & D.J. Cooke, *Avoiding errors about 'margins of error': Authors' reply*, 192 BRITISH J. PSYCHIATRY 561, 561-62 (2007) (correspondence)

[77] Beyond a reasonable doubt: Arizona, California, Illinois, Iowa, Kansas, Massachusetts, Missouri, South Carolina, Texas, Washington, and Wisconsin. Clear and convincing evidence: Florida, Minnesota, New Jersey, North Dakota, Virginia. Nat'l Center for Prosecution of Child Abuse, "Civil Commitment of Sexually Violent Predators," at www.ndaa.org/pdf/sexually_violent_predator_statutes.pdf (downloaded July 8, 2010). A complete classification is presented in Table 3.

[78] McCauliff, *supra* note 61, at 1325 tbl.2. On quantifying BRD, see generally Vars, *supra* note 61, at 22-23, and Peter Tillers & Jonathan Gottfried, *Case Comment--United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim that Proof Beyond a Reasonable Doubt Is Unquantifiable?*, 5 LAW, PROBABILITY & RISK 135, 141-51 (2006).

will be below the bottom end of the interval. If that lower bound is above the commitment standard, then commitment is appropriate.[79] Figure 1 illustrates, assuming a normal error distribution.

Figure 1. Lower Bound of 50% Confidence Interval ("CI") Is the Threshold for Clear & Convincing Evidence (75%)



This new approach can be applied to the 52% figure with its 6% to 95% range. Assuming a normal error distribution, the lower range of the 50% CI is around 37%. This means there is a 75% chance that the individual recidivism rate is above 37%. That is not enough to clear a "more likely than not" commitment standard, but it is substantially better than the 6% lower bound of the 95% CI would suggest. For BRD the relevant CI is 80% and the lower bound is around 23%.

Whether values in this range (23-37%) suffice for commitment depends on the commitment standards. Discerning these standards is not always easy. Writing in 1997, Eric

[79] This approach was suggested by Professor Cohen. *See* Cohen, *supra* note 72, at 421 ("The difference [between CCE and POE], however, also could be that the clear and convincing standard requires the factfinder to use a higher level of confidence in constructing the interval estimate.").

Janus and Paul Meehl found that no court or legislature had quantified its standard of commitment.[80] They assumed that "highly likely" meant 75% and "likely" meant 50%.[81] Others more recently claimed that all the state statutes were clear and, despite varying language, set a bar of "roughly 70%."[82] Some courts and legislatures have now provided one relatively clear benchmark: Washington, for example, by statute requires the likelihood of recidivism to be "more probably than not."[83] At least one commentator[84] and two courts[85] have suggested that "likely" is a lower bar.

Into this morass it is with some trepidation that I offer the Table 3, illustrating the diversity of both commitment and proof standards:

[80] Janus & Meehl, *supra* note 15, at 40, 60. Professor Janus essentially reiterated this point in 2006, calling the thresholds "poorly defined" and "vague." Prentky et al., *supra* note 8, at 360, 372.

[81] *Id.* at 41.

[82] George G. Woodworth & Joseph B. Kadane, *Expert Testimony Supporting Post-Sentence Civil Incarceration of Violent Sexual Offenders*, 3 L., PROBABILITY & RISK 221, 225, 227 (2004).

[83] Wash. Stat. Ann. § 71.09.020(7).

[84] Jackson & Hess, *supra* note 11, at 439; *see also* Thomas Grisso & Paul S. Appelbaum, *Is it Unethical to Offer Predictions of Future Violence?*, 16 L. & HUMAN BEHAV. 621 (1992) ("Given accurate and scientifically supportable predictive testimony about degree of risk, it is up to society (usually its representative on the bench) to determine whether 40%, 30%, or even 20% risk of future violence might reach a threshold justifying a particular legal intervention . . . .").

[85] Fennel, *supra* note 43, at 39 (reporting on California and Massachusetts court opinions)

Table 3. Standards of Commitment and Proof by State

Commitment Standard	*Proof Standard*	
Likelihood of recidivism	*Clear and convincing evidence (75%)*	*Beyond a reasonable doubt (90%)*
>50%	Minn.,[86] N.J.[87]	Ariz.,[88] Ill.[89]
50%	Fla.,[90] Mo.,[91] Neb.[92]	Iowa,[93] Wash.,[94] Wis.[95]
<50%		Cal.,[96] Fed.,[97] Mass.[98]
Unspecified	N.H.,[99] N.Y.,[100] N.D.,[101] Va.[102]	Kan.,[103] S.C.,[104] Tex.[105]

[86] In re Linehan, 594 N.W.2d 867, 876 (Minn. 1999) ("highly likely"); Minn. Stat. Ann. § 253B.09(1)(A) ("clear and convincing evidence"); Minn. Stat. Ann. § 253B.185.

[87] In re Commitment of W.Z., 801 A.2d 205, 218 (N.J. 2002) ("highly likely"); N.J. Stat. Ann. § 30:4-27.32(a) ("clear and convincing evidence").

[88] In re Leon G., 26 P.3d 481, 489 (Ariz. 2001) (en banc) ("highly probable"), vacated on other grounds sub nom., Glick v. Arizona, 535 U.S. 982 (2002); Ariz. Rev. Stat. § 36-3707(A) ("beyond a reasonable doubt").

[89] In re Detention of Hayes, 747 N.E.2d 444, 453 (Ill. App. 2001) ("much more likely than not"); 725 Ill. Comp. Stat. § 207/35(d)(1) ("beyond a reasonable doubt").

[90] Westerheide v. State, 831 So. 2d 93, 106 (Fla. 2002) ("having a better chance of existing or occurring than not"); Fla. Stat. § 394.917(1) ("clear and convincing evidence").

[91] Mo. Stat. Ann. § 632.480(5) ("more likely than not"); Mo. Stat. Ann. § 632.495(1) ("clear and convincing evidence").

[92] In re G.H., 781 N.W.2d 438, 445 (Neb. 2010) ("more likely than not"); Neb. Rev. Stat. § 71-1209(1) ("clear and convincing evidence").

[93] Iowa Code Ann. § 229A.2(4) ("more likely than not"); Iowa Code Ann. § 229A.7(5)(a) ("beyond a reasonable doubt").

[94] Wash. Stat. Ann. § 71.09.020(7) ("more probably than not"); Wash. Stat. Ann. § 71.09.060(1) ("beyond a reasonable doubt").

[95] Wis. Stat. Ann. § 980.01(1m) ("more likely than not"); Wis. Stat. Ann. § 980.05(3)(a) ("beyond a reasonable doubt").

[96] People v. Superior Court (Ghilotti), 44 P.3d 949, 968 (Cal. 2002) (stating that "likely" "does not mean the risk of reoffense must be higher than 50 percent," but instead means the person "presents a substantial danger--that is, a serious and well-founded risk--of reoffending"); Cal. Welfare & Inst. Code § 6604 ("beyond a reasonable doubt").

[97] 18 U.S.C.A. § 4247 ("serious difficulty refraining"); United States v. Hunt, 643 F. Supp. 2d 161, 180 (D. Mass 2009) ("this court does not construe the 'serious difficulty' criterion for commitment to require proof of any statistical probability of reoffense"); 18 U.S.C.A § 4248(d) ("clear and convincing evidence").

[98] Commonwealth v. Boucher, 780 N.E.2d 47, 53 (Mass. 2002) (defining "likely" not as "more likely than not," but rather as "would reasonably be expected"); Mass. Gen. L. Ann. Ch.123A, § 14(d) ("beyond a reasonable doubt").

[99] N.H. Rev. Stat. § 135-E:2(VI) ("potentially serious likelihood"); State v. Paradis, 455 A.2d 1070, 1072 (N.H. 1983) ("dangerous"); N.H. Rev. Stat. § 135-E:11(I) ("clear and convincing evidence").

[100] N.Y. Mental Hyg. Law §10.03(e) ("likely to be a danger to others"); N.Y. Mental Hyg. Law §10.07(d) ("clear and convincing evidence").

[101] In re B.V., 708 N.W.2d 877, 882 (N.D. 2006) (stating that defining "likely" as "of such a degree as to pose a threat to others . . . prevents a contest over percentage points and the results of other actuarial tools"); N.D. Stat. Ann. § 25-03.3-13 ("clear and convincing evidence").

[102] Shivaee v. Commonwealth, 613 S.E.2d 570, 577 (Va. 2005) ("a menace to the health and safety of others"); Va. Code Ann. § 37.2-908(C) ("clear and convincing evidence").

[103] Kan. Stat. Ann. § 59-29a02(c) ("menace"); Kan. Stat. Ann. § 59-29a07(a) ("beyond a reasonable doubt").

[104] S.C. Stat. § 44-48-30(9) ("pose a menace"); S.C. Stat. § 44-48-100(A) ("beyond a reasonable doubt").

Table 3 and the accompanying notes illustrate two important points: (1) there is great diversity *across* states in approaches on both standard of commitment and standard of proof; and (2) with the exception of the relatively precise "more likely than not" standard, the vague statements of commitment standards strongly suggest that there is no uniformity *within* most states either.[106] On the second point, commentators have recommended quantification "so that the true distribution of the risk of error in prediction can be seen."[107] This would reveal the otherwise hidden policy tradeoffs.[108] The debate about quantification is, as in other contexts, partly about who one wants to make these tradeoffs.[109] By adopting numerical standards, legislatures and appellate courts can shift discretion away from trial courts, juries, and testifying experts. Doing so would advance the goals of transparency and consistency, but at the price of case-specific flexibility.[110]

One argument for flexible standards in this context deserves special attention. The likelihood of recidivism is only one component in determining the gravity of the threat posed by a particular sex offender. Also relevant are the magnitude of future harms, their frequency, and their imminence (how soon they are likely to occur).[111] Sex offender commitment statutes generally do not capture these other elements, but case-specific adjustment of the required probability level might (e.g., "menace"). Current ARAIs may aggravate the problem: "existing

[105] Tex. Health & Safety Code Ann. § 841.003(a)(2) ("likely"); Beasley v. Molett, 95 S.W.3d 590, 600 (Tex.App.-Beaumont 2002) ("The term 'likely,' as ordinarily defined, means 'probable.' Something that is probable is beyond a mere possibility or potential for harm."). Tex. Health & Safety Code Ann. § 841.062(a) ("beyond a reasonable doubt").

[106] *See* Jason A. Cantone, *Rational Enough To Punish, But Too Irrational To Release: The Integrity Of Sex Offender Civil Commitment*, 57 DRAKE L. REV. 693, 713 (2009) (pointing out that vague legal standards may lead to lack of uniformity).

[107] Janus & Meehl, *supra* note 15, at 34.

[108] *Id. Cf.* Grisso & Appelbaum, *supra* note 84, at 627.

[109] Vars, *supra* note 61, at 21-22.

[110] *Id.*

[111] Janus & Prentky, *supra* note 52, at 1449.

actuarial methods are optimized to predict the most common but least severe sexual offenses."[112] Of course, ARAIs and statutes could be adjusted to address the problem without sacrificing transparency and consistency, but decisionmaker discretion with non-quantified risk thresholds is perhaps a next best solution.

The diversity of approaches shown in Table 3 underscores the importance of the fact that ARAIs like the Static-99 are meaningful only when confidence intervals are properly understood and reported. There are at least six possible combinations of standard of proof and standard of commitment. The state-specific combination must be factored into any ultimate opinion based on an ARAI. And experts are making such judgments as a matter of routine. Ninety-five percent of evaluators reported using the Static-99 "always or most of the time," and the same percentage "reported that it was either essential or recommended for an evaluator to state an ultimate opinion regarding whether a sex offender meets civil commitment criteria in their final report."[113]

Whether or not one favors quantification of the commitment and proof standards, one should favor quantification of the likely error associated with the Static-99 or other ARAIs. The recidivism tables and risk categories are at best misleading in the absence of confidence intervals. No one—not the expert, the trial court, the jury, an appellate court, or the legislature—can balance the costs and benefits of sex offender commitment without some sense of the error of actuarial prediction.[114] Experts are ethically bound to report the limitations of actuarial results.[115] Actuarial and other statistical methods have the potential to generate both risk

[112] Sreenivasan et al., *supra* note 45, at 466.

[113] Jackson & Hess, *supra* note 11, at 434, 435.

[114] *Id.* at 439.

[115] Randy K. Otto & John Petrila, *Admissibility of Testimony Based on Actuarial Scales in Sex Offender Commitments: A Reply to Doren*, 3 SEX OFFENDER L. REPORT 1, 15 (Dec./Jan. 2002) ("there is an obligation on the part of experts to be as precise as possible not only about their testimony, but about the limitations on the tests that underlie their testimony"); Stephen D. Hart, Christopher D. Webster, & Robert J. Menzies, *A Note on Portraying the*

estimates and confidence intervals on those estimates.[116] Parts II and III of this Article realize that potential for a new model and consider what the results mean for risk assessment generally.

This endeavor follows in the footsteps of Janus and Meehl.[117] Given certain assumptions about the meaning of commitment standards (e.g., "likely" = 50%; "highly likely" = 75%),[118] accuracy of prediction (0.75),[119] and the base rate of recidivism (20%-45%),[120] Janus and Meehl concluded that it was possible to meet the commitment standard. Notably, the standard of proof was simply folded into standard of commitment with no downward adjustment in the likelihood of recidivism.[121] This Article corrects that methodological error[122] and replicates for a particular prediction model and dataset what Janus and Meehl attempted as a matter of theory. I make no assumptions about accuracy or base rate,[123] but rather let the data set those values. Nor do I assume a single cut-score.[124] Finally, I go farther than Janus and Meehl by pointing the direction toward better ARAIs.

Accuracy of Violence Predictions, 17 L. & HUMAN BEHAV. 695, 696 (1993) ("In the context of psycholegal assessments, unwillingness to qualify one's confidence in violence predictions or failure to make probabilistic statements regarding the likelihood of future violence is, at best, poor practice; at worst, it is simply unethical . . . ."); Grisso & Appelbaum, *supra* note 84, at 630 (explaining that expert has ethical duty of "presenting reliable testimony and clearly explaining its limitations").

[116] Janus & Prentky, *supra* note 52, at 1493. Some have suggested that this is impossible. *See* Gina M. Vincent, Shannon M. Maney, & Stephen D. Hart, *The Use of Actuarial Risk Assessment Instruments in Sex Offenders*, *in* SEX OFFENDERS: INDENTIFICATION, RISK ASSESSMENT, TREATMENT, AND LEGAL ISSUES 71 (Fabian M. Saleh et al., eds., 2009) (estimates "cannot be done with known precision").

[117] Janus & Meehl, *supra* note 15.

[118] *Id.* at 41.

[119] *Id.* at 49.

[120] *Id.* at 51.

[121] *Id.* at 43.

[122] I use the term "error" descriptively, not normatively. Jurisdictions in fact decouple the commitment standard and standard of proof. Whether that bifurcated approach is defensible (or constitutional) is outside the scope of this Article. *See* CHRISTOPHER SLOBOGIN, MINDING JUSTICE: LAWS THAT DEPRIVE PEOPLE WITH MENTAL DISABILITY OF LIFE AND LIBERTY 144 (2006); Grant H. Morris, *Defining Dangerousness: Risking a Dangerous Definition*, 10 J. CONTEMP. LEGAL ISSUES 61, 87-88 (1999).

[123] *See* Dennis M. Doren & Douglas L. Epperson, *Great Analysis, But Problematic Assumptions: A Critique of Janus and Meehl (1997)*, 13 SEXUAL ABUSE: A JOURNAL OF RESEARCH & TREATMENT 45, 46-48 (2001) (arguing that the assumed base rate was too low).

[124] *Id.* at 49-51.

George Woodworth and Joseph Kadane also examined a particular ARAI—in their case another pre-existing instrument, the Mn-SOST-R.[125] That instrument shares the deficiencies of the Static-99 as described above. Furthermore, Woodworth and Kadane collapsed standards of commitment into a single, unsupported cut-off and ignored standards of proof entirely.[126] The present Article, however, agrees with and implements their suggestion that logistic regression can improve upon less sophisticated, "actuarial" methods.[127]

In perhaps the closest precursor to the present Article, Richard Wollert applied Bayesian techniques to evaluate several ARAIs, including the Static-99.[128] Wollert apparently followed Janus and Meehl in assuming a recidivism threshold of between 50% and 75%.[129] He not only found that the studied instruments failed to identify even one individual qualified for commitment, but concluded that the instruments would always fail unless base rate recidivism rose above 25%.[130] Wollert again used a single cut-off, but this Article confirms his basic results using corrected commitment criteria and much different methodology.

II. A New Model

A. Data

The data for this study are taken from the United States Department of Justice, Bureau of Justice Statistics (BJS), *Recidivism of Prisoners Released in 1994: [United States]* (ICPSR Study No. 3355). That database includes prior criminal history information and recidivism over a three-year follow-up period for 38,624 sampled prisoners released from prisons in 15 states in

[125] Woodworth & Kadane, *supra* note 82.
[126] *Id.* at 227 ("roughly 70%"); *id.* at 239 (treating recidivism percentage equal to or greater than commitment standard as sufficient to justify commitment).
[127] *Id.* at 239.
[128] Wollert, *supra* note 15.
[129] *Id.* at 58.
[130] *Id.* at 75, 79.

1994.[131] The data were chosen for several reasons: (1) the dataset is very large and therefore comes closest to representing the United States as a whole; (2) it covers a time period in which sex offender commitment was not yet prevalent—thus, it includes every sex offender, not just those deemed safe enough to release.[132] All violent sex offenders were included in the BJS study; non-violent sex offenders were sampled. Because the present study includes both types of sex offenders, probability weights were used to adjust for sampling.[133]

The present study is limited to the 10,400 men in the BJS study who were incarcerated for a sex offense immediately prior to their release in 1994. Table 4 sets forth some of their relevant characteristics.

[131] Codebook iii. The states are: Arizona, California, Delaware, Florida, Illinois, Maryland, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Texas, and Virginia. *Id.* iv.

[132] No state in the study had sex offender commitment in 1994, except for Minnesota late in that year. Since they had expressly not been selected for commitment, individuals released in Minnesota after the effective date of the sexual offender commitment law were omitted.

[133] All analyses were rerun without weights. There were only trivial changes in results. This was not unexpected as in most cases fewer than 50 individuals had probability weights not equal to one. Along the same lines, excluding non-violent sex offenders had no significant impact on the results. Such offenders made up about 3% of the total and were mostly serving time for statutory rape or incest. *See infra* Table 4.

Table 4. Summary Statistics (Unweighted)

	Number	*Percentage*
Offense		
Rape	2,407	23.1%
Statutory Rape	282	2.7%
Incest	59	0.6%
Sexual Abuse	3,856	37.1%
Child Molestation	3,278	31.5%
Sodomy	518	5.0%
Race/Ethnicity		
Black	3,238	31.1%
Hispanic	1,697	16.3%
Age*		
<25	1,340	12.9%
25-35	3,733	35.9%
35-50	4,141	39.9%
>50	1,177	11.3%

*Age is missing for 9 individuals.

Due to missing data, roughly 15% of the total sample was excluded: the key regression presented below in Table 5 is based on 8,881 instead of 10,400 observations.[134]

B. Methodology

The basic approach of this Article is to employ the BJS data to shed light on current sex offender commitment practice. This is *not* a direct test of the efficacy of the Static-99, because data including Static-99 scores have not been made publicly available. Rather, more sophisticated statistical tools are used to demonstrate the points made above regarding age, norms, and error. Most fundamentally, the data are used to estimate how many individuals met the legal standards for commitment, and how error of prediction affects that estimate. The

[134] By far the largest source of missing data is the lack of arrest data. Every convict must have been arrested at least once, so observations without any arrests were dropped. In contrast, I retained observations with one or more arrest charge and additional charges coded as “missing” or “unknown.” A strict reading of the Codebook would exclude such individuals because known negatives should have been coded “not applicable.” Codebook 13-14. Such a reading, however, would in almost every case contradict the recorded number of charges per arrest (*e.g.*, variable name = A001CNT).

primary data analysis tool is logistic regression. It is a commonly-used model in social science for true dichotomous outcomes, like recidivism. Technical details of the methodology are presented in the Appendix.

C. Results

1. Age

As described above, the original Static-99 converts the continuous variable age into a dichotomous variable Young equal to one if the offender is less than 25 years old at release. Even the creators of the Static-99 admit that this was a mistake. One obvious question is whether a more refined treatment of age, on its own, can predict recidivism as well as or better than the Static-99. The answer is mixed: age alone does as well as the Static-99 in predicting violent (both sexual and non-sexual) recidivism, but not as well in predicting sexual (both violent and non-violent) recidivism.

The present study includes two continuous age variables: age at first arrest and age at release. Squared and cubed versions of each are also included to allow non-linear effects.[135] Two logistic regression (or logit) models estimated the likelihood of recidivism using these six age variables alone. The mean predicted likelihood for the group that was arrested for a subsequent violent offense was significantly higher than the group that did not recidivate violently: 30.6% versus 24.6% (Cohen's *d* = 0.58 [95% CI = 0.53, 0.63]). This effect size—a measure of the strength of association between two variables—matches that of the Static-99. A recent meta-analysis of 35 studies of the Static-99's predictions of violent recidivism found a mean Cohen's *d* equal to 0.57, with a 95% CI of 0.52 to 0.62.

[135] This is in some ways a less constrained version of Prentky et al., *supra* note 8, at 376-77, who concluded that recidivism estimates should be reduced by two percent for every year after age forty. The results reported in Table 5 suggest a large negative effect of age throughout the range of released individuals.

Sexual recidivism is a different story. The same Static-99 meta-analysis found a mean Cohen's *d* of 0.67 in predicting sexual recidivism with a 95% confidence interval of 0.62 to 0.72. For the age-only logit model described above, the Cohen's *d* was only 0.30 (95% CI = 0.23, 0.38). Hence, the Static-99 was significantly better than age alone at predicting sexual recidivism.

But that is not really a fair comparison. The Static-99 includes nine variables other than age. Five assess prior involvement with the criminal justice system. Can the combination of the two continuous age variables, two criminal history variables, and superior methodology (logistic regression) compete with the Static-99 in predicting sexual recidivism? The new model adds two sets of variables closely mimicking two items on the Static-99: (1) a dummy variable equal to one if the individual has a prior conviction for a violent offense (along with another dummy equal to one if offense code was missing); and (2) a set of dummy variables based on the number of prior arrests for sexual offenses (0 or 1; 2 or 3; 4, 5, or 6; and 7 or more). Table 5 reports the results.

Table 5. Logit Regression Predicting Rearrest for Sexual Offense

Number of obs = 8881
Pseudo R2 = 0.0557
Area under ROC curve* = 0.6605
Log likelihood = -2591.4108
Cohen's *d* = 0.761

Variable	*Coefficent*	*Std. Error*	*z*	*P>|z|*	*[95% Conf. Interval]*	
Age at Release	-0.000384	0.000351	-1.09	0.275	-0.001073	0.000305
…squared	3.31E-08	2.53E-08	1.31	0.191	-1.65E-08	8.26E-08
…cubed	-9.33E-13	5.89E-13	-1.58	0.113	-2.09E-12	2.21E-13
Age at First Arrest	1.06E-04	1.52E-04	0.70	0.485	-1.92E-04	4.05E-04
…squared	-2.19E-08	1.36E-08	-1.61	0.108	-4.86E-08	4.81E-09
…cubed	7.86E-13	3.91E-13	2.01	0.045	1.92E-14	1.55E-12
Violent Prior Conviction?						
…yes	-0.220320	0.110261	-2.00	0.046	-0.436428	-0.004211
…missing	0.420117	0.201135	2.09	0.037	0.025900	0.814334
Sexual Prior Arrests						
…0 or 1	-1.494985	0.230609	-6.48	0.000	-1.946970	-1.043000
…2 or 3	-0.875472	0.227590	-3.85	0.000	-1.321540	-0.429404
…4, 5, or 6	-0.289424	0.270223	-1.07	0.284	-0.819052	0.240205
…6 or more	0.023862	0.561787	0.04	0.966	-1.077220	1.124945
Constant	0.719014	1.414187	0.51	0.611	-2.052742	3.490770

*Based on unweighted regression.

Perhaps not surprisingly, the most significant predictor of sexual recidivism was having many prior arrests for sexual offenses. But more important for the present Article is the result that better methodology and more sensitive treatment of age more than compensated for omitting seven of the ten Static-99 items: the Cohen's *d* for this expanded model was 0.76, which is above the 95% CI reported by the Static-99 meta-analysis. To be sure, the CIs overlap substantially, but it appears fairly safe to say that the new model does as well or better than the Static-99.

However, the analysis to this point has been unfair in at least one way to the Static-99. The meta-analysis reviewed applications of the Static-99 to populations other than the ones with which the instrument was developed. In contrast, the logit models have been evaluated with the same data that was used to construct them. This out-of-sample versus in-sample comparison is not apples to apples. A properly specified model will generally perform best in the construction sample. To correct this bias, the data were split into two parts: the model was constructed using data from every state other than California and its predictions tested in California. California was chosen because it was home to the largest number of released prisoners, roughly one-third of the total.

The results with respect to violent recidivism stand: the logit model with two sets of age variables does as good as or better than the Static-99 in predicting violent recidivism in new samples. Specifically, the Cohen's *d* for the out-of-sample logit model was 0.56 with a 95% confidence interval of 0.48 to 0.64. This Cohen's *d* is nearly identical to the Static-99.

In contrast, the Static-99 has an edge over the logit model for out-of-sample prediction of sexual recidivism. The Cohen's *d* for logit is 0.53 (95% CI = 0.41, 0.65), as compared with 0.67 (95% CI = 0.62, 0.72) for Static-99. Note that the confidence intervals overlap, so one cannot reject the hypothesis of equivalence. Recidivism rates vary considerably by race and ethnicity in this dataset,[136] so including these variables might increase predictive power, particularly since California's demographics may be unusual. However, after controlling for age and criminal history, race and ethnicity actually *reduce* effect size, strongly suggesting that these characteristics do not belong in the model.[137]

To summarize, the data support using age as a continuous variable with a technique like logistic regression. A logit model on age at first arrest and age at release, together with squared and cubed terms allowing for non-linearity, was better at predicting violent recidivism than the 10-item Static-99. Adding two sets of criminal history variables made the logit model nearly as

[136] Patrick A. Langan, Erica L. Schmitt, & Matthew R. Durose, Recidivism of Sex Offenders Released from Prison in 1994, at 18 tbls. 11, 12, U.S. Dep't of Justice: Bureau of Justice Statistics (Nov. 2003).

[137] Consistent with this conclusion, the coefficient on the black variable failed to achieve statistical significance. However, the Hispanic coefficient was negative and significant (*p* = 0.008).

good as the Static-99 in predicting sexual recidivism. In other words, more sophisticated use of age eliminated the need to collect additional criminal history information or to code the three Static-99 items based on victim characteristics (see *supra* Table 1). Of course, even greater predictive power would likely be achievable by including those items in a regression model.

2. Norms

Recall that recidivism rates vary significantly across time and jurisdiction. The present study is no exception. There is wide disparity in the recidivism rates of sex offenders among the 15 states included in this study (see Figure 6).

Table 6. Rearrest Rate by State and Offense Type

State	*Sexual*	*Violent*
Arizona	5.9%	22.8%
California	8.4%	24.6%
Delaware	10.4%	42.9%
Florida	9.7%	28.5%
Illinois	12.2%	39.6%
Maryland	13.5%	40.0%
Michigan	5.7%	13.6%
Minnesota	18.1%	27.1%
New Jersey	6.6%	23.7%
New York	10.0%	26.2%
North Carolina	7.3%	26.6%
Ohio	13.0%	31.2%
Oregon	8.6%	23.8%
Texas	6.7%	23.0%
Virginia	9.3%	27.5%

This suggests that state-specific norms are needed to evaluate an individual's risk of recidivism. Better evidence would be to find significant state effects after controlling for age and criminal history.

To the sexual recidivism model outlined in the previous section I added dummy variables for 14 of the 15 states (the last, Virginia, was omitted).[138] The coefficients on two states,

[138] For technical reasons, there must always be an omitted reference category for regression models to work. See "Dummy variables," at http://en.wikipedia.org/wiki/Dummy_variable_(statistics) (visited Jan. 28, 2011) (footnote omitted):

Maryland and Texas, were negative and statistically significant ($p < 0.05$).[139] A chi-square test rejected the null hypothesis that all the state dummy coefficients were equal ($chi^2(14) = 54.01$; $p < 0.0001$). I repeated the analysis for violent recidivism. Here again, the independent variables were based on age and criminal history, along with state dummy variables. Five state coefficients were statistically significant at the 5% level (Illinois, Maryland, Michigan, North Carolina, and Texas). Given this, it is not surprising that overall there is a highly significant difference among states ($chi^2(14) = 106.73$; $p < 0.0001$).

Many factors could explain the significant state effects. The important point is that significant differences among states persist even after controlling for factors like those included in the Static-99. Hence, those creating risk assessment instruments and those using them should consider seriously state-specific norms.[140]

3. Error and Standards of Proof

Again, the logit model described above predicted sexual recidivism within sample as well as or better than the Static-99. One great advantage of the logit model is that it is possible to measure the standard error associated with individual predictions. As a result, one can actually test whether each released sex offender exceeded the commitment standard with the requisite degree of confidence. Table 7 summarizes the results.

If dummy variables for all categories were included, their sum would equal 1 for all observations, which is identical to and hence perfectly correlated with the vector-of-ones variable whose coefficient is the constant term; if the vector-of-ones variable were also present, this would result in perfect multicollinearity, so that the matrix inversion in the estimation algorithm would be impossible. This is referred to as the dummy variable trap.

[139] North Carolina had a negative coefficient that came very close to statistical significance ($p = 0.054$).

[140] The Static-99 creators themselves have recently concluded that "variation in recidivism rates across samples cannot be ignored." Hanson, Helmus & Thornton, *supra* note 34, at 207.

Table 7. Individuals Who Met the Dangerousness Standards for Commitment (Out of 8881)*

Commitment Standard	*Proof Standard*	
Likelihood of recidivism	*Clear and convincing evidence (75%)*	*Beyond a reasonable doubt (90%)*
>75%	0	0
>50%	0	0
>25%	217 2.4%	201 2.3%

*Using predictions and standard errors of model summarized in Table 5.

The most striking result is that *not one* of the 8,881 released sex offenders was more likely than not to be rearrested for a sexual offense even at the lower clear and convincing evidence standard. This means that, using the instrument alone, no one met the dangerousness threshold used in half of the jurisdictions with sex offender commitment (*see supra* Table 3).[141]

However, at least three jurisdictions—California, Massachusetts, and the federal government—set the bar lower than that: a less than 50% chance of recidivism (which, again, I arbitrarily set at 25%) beyond a reasonable doubt. About 2.3% of individuals (201 of 8,881) met this standard. Among this most dangerous group, the actual recidivism rate was very close to 40%. In other words, if these individuals had been committed, three people who would not have reoffended would have been detained for every two recidivists. This analysis also showed that almost 90% of recidivists still would have been released.[142]

[141] At least one group of researchers objects to using logistic regression in this context. Wollert et al., *supra* note 33, at 483. A linear regression model achieves comparable results: no one qualified at the 50% or 75% levels; under 3% qualified at the 25% level.

Some have argued that violent recidivism is a better measure among sex offenders of the conduct civil commitment is designed to prevent. Marnie E. Rice et al., *Violent Sex Offenses: How are They Best Measured from Official Records?*, 30 L. & HUMAN BEHAV. 525 (2006). The analysis was repeated using the logit model described above predicting violent recidivism using six age variables. Despite a much higher base rate of violent recidivism (about 26%), no one qualified for commitment at the 75% threshold and around 0.2% at the 50% threshold (out of 9015 individuals, 23 at CCE and 19 at BRD). However, solid majorities cleared the 25% hurdle.

[142] Woodworth & Kadane, *supra* note 82, at 239 (reporting somewhat better numbers in direct test of the MnSOST-R).

The contribution of error can be quantified: how many individuals would have qualified for commitment if the error associated with predictions were ignored, as the creators of the Static-99 originally advocated?[143] Still none at the 50% and 75% commitment standards. At the lowest threshold (25%), however, 242 individuals would have qualified. In other words, properly factoring in error and applying a higher standard of proof can reduce commitments identified by the logit model by up to 17% (242 to 201). The error of individual prediction associated with an actuarial tool like the Static-99 is likely greater because it rounds variable effects and lumps individuals into rough categories.[144]

III. Discussion

A. Limitations

This is not a direct test of the Static-99. Nonetheless, its findings illuminate shortcomings of that instrument and other actuarial approaches. The most important conclusion is that an instrument as good as the Static-99 generally cannot identify individuals who satisfy the legal requirements for sex offender civil commitment. By achieving effect sizes comparable to the Static-99 with more sensitive treatment of age and fewer variables, this Article provides further support for the view that the Static-99 does not properly account for age. By showing significant state effects using a model comparable to the Static-99, this Article underscores the importance of tailoring predictions to the particular jurisdiction. And, finally, the Article

[143] Recall that this is equivalent to applying the preponderance of the evidence standard.

[144] As mentioned above, an alternative to this confidence interval approach to standards of proof is to set the cut-score in order to achieve a desired ratio of false positives (FP) to false negatives (FN). Normally, these values are unavailable, Cohen, *supra* note 72, at 417, but here no one was civilly committed and we have actual data on recidivism. (Notably, this approach is independent of the commitment threshold and therefore probably not a good fit in this context.) The three standards of proof can be equated to FP:FN error ratios as follows: POE (50%), 1:1; CCE (75%), 1:3; and BRD (90%), 1:9. Applying these standards, 479, 346, and 149 individuals, respectively, qualified for commitment based on the logit predictions and observed recidivism. Thus, the unequal weighting of errors implied by heightened standards of proof can reduce commitments by up to 69%.

demonstrates in two different ways the large impact of prediction error on how many individuals will qualify for commitment.

One criticism of the analysis above is that it is directed against a straw man: the Static-99 is not used in isolation. Rather, experts testify about the meaning of the score and offer opinions as to dangerousness that incorporate other factors. Existing data, however, suggest that adding clinical judgment to actuarial results does not improve predictive accuracy.[145] Indeed, to the extent there have been studies, they suggest that adjusting actuarial results actually decreases accuracy.[146]

Another limitation is that recidivism information in these data is available only for the first three years after release. This generates a relatively low base-rate. The observed sexual recidivism rate in the data is about 9.2%; in contrast, according to some researchers, "approximately 30% of sex offenders released from secure custody will have subsequent offenses recorded as sexual on police rap sheets."[147] On the one hand, the low base-rate makes the high effect sizes more impressive since prediction of low probability events is more difficult. This bolsters the present findings on age. But on the other hand, the low base-rate artificially reduces the number of individuals who qualified for commitment and perhaps exaggerates the impact of prediction error. One could respond by arguing that neither effect is "artificial." When spending limited resources, the imminence of harm is plainly relevant. To prevent one

[145] *See* Hamilton, *supra* note 16, at 44 ("there is no empirical evidence that modifying actuarial scores improves the accuracy of predictions"); Terence W. Campbell & Gregory DeClue, *Flying Blind with Naked Factors: Problems and Pitfalls in Adjusted-Actuarial Sex-Offender Risk Assessment*, 2 OPEN ACCESS J. FORENSIC PSYCH. 75, 96 (2010), at http://www.forensicpsychologyunbound.ws/ – 2010. 2: 75-101 ("Based on available data, at its best, [Adjusted Actuarial Assessment] neither increases nor decreases the accuracy of actuarial classification."); *see also* Harris & Rice, *supra* note 46, at 1640.

[146] *See* R.K. Hanson & K.E. Morton-Bourgon, *The Accuracy Of Recidivism Risk Assessments For Sexual Offenders: A Meta-Analysis Of 118 Prediction Studies*, 21 PSYCH. ASSESSMENT 1, 7 (2009) ("the adjusted scores showed lower predictive accuracy than did the unadjusted actuarial scores"); Campbell & DeClue, *supra* note 145, at 75 ("At its worst, [Adjusted Actuarial Assessment] dilutes actuarial accuracy.")

[147] Harris & Rice, *supra* note 46, at 1642. *But see* Hanson & Bussière, *supra* note 20, at 351 (reporting 13.4% sexual recidivism).

sexual reoffense by locking up an individual for more than three years is arguably not cost-benefit justified. That point is of course debatable, and it must be conceded that the short follow-up period covered by the data is a limitation.

B. Implications

The way we select sex offenders for civil commitment is inadequate: essentially no one meets the legal standards. The practice of sex offender commitment should be curtailed or eliminated, the selection criteria lowered, or the selection methodology improved. The first two options involve policy judgments outside the scope of this Article. This Article does shed light on a way forward for improved methodology. It should be emphasized that there is no guarantee adopting one or even all of the suggestions below will solve the bottom-line problem.

Specifically, due to the low base rate of recidivism and substantial prediction error, an instrument as good as the Static-99 identified *not one* individual who qualified for commitment at the 50% or 75% threshold.[148] As noted above (*see supra* Table 3), half of jurisdictions with sex offender commitment apply thresholds at or above 50%. No evidence shows that adding other evidence to actuarial results improves prediction accuracy.[149] The obvious implication is that no one in these jurisdictions deserved to be civilly committed as a sex offender.

There are several responses to this finding. First, it depends crucially on the short follow-up period and resulting low base rate.[150] Arrest reports listing sexual offenses may understate recidivism for other reasons as well—*e.g.*, failure of victims to report and failure of police to list the more difficult to prove sexual component of offenses like assault.[151] Second, the other half

[148] This finding stands in contrast to that of Janus and Meehl, who concluded as a theoretical matter that those standards could be met. Janus & Meehl, *supra* note 15, at 33.

[149] *See supra* notes 145-148 and accompanying text.

[150] Observed base rates vary as widely as 7.5% to 66.7%, with two large meta-analyses finding rates around 13-14% for 5-year recidivism. Prentky et al., *supra* note 8, at 373-74.

[151] Harris & Rice, *supra* note 46, at 1643-44.

of jurisdictions have lower or unspecified commitment thresholds. This Article finds that standard could have been met with requisite certainty for a significant fraction of the sample. The problem therefore could be described as setting the threshold too high, not failing to meet an impossible standard.

Still, this Article represents one of the first empirical tests of whether an instrument like the Static-99 can identify qualified individuals. The instrument failed to do so in half of jurisdictions. The base rate may be wrong or those jurisdictions may have the wrong standard, but it would seem the burden going forward should be on the developers of ARAIs like the Static-99 to show that the instruments as revised can identify individuals who meet the commitment threshold at the required standard of proof. If no such showing is forthcoming, the entire enterprise of sex offender commitment is justifiably in doubt.[152]

The Static-99 can be improved, at least in accounting for age, adjusting for jurisdiction-specific norms, and reporting error. The developers of the Static-99 have recognized some of this and offered updated alternatives. Recall that a revised version of the Static-99 includes four age categories instead of two.[153] This is certainly a step in the right direction, but why not go all the way: include age as a continuous variable? By similar token, new norms are necessary—as this Article confirms (*see supra* Section II.C.2)—but including dummy variables for each jurisdiction, updating data each year, and reestimating the logit model outlined above has the potential to seamlessly adjust predictions as observed behavior changes.[154] Because one can directly calculate individual errors using this approach, a successful challenge along the lines of

[152] Alternatives like longer prison sentences, Cantone, *supra* note 106, at 720-21, or supervision and treatment in the community, Eric S. Janus, *Minnesota's Sex Offender Commitment Program: Would An Empirically-Based Prevention Policy Be More Effective?*, 29 WM. MITCHELL L. REV. 1083, 1132-33 (2003), may be preferred.
[153] Helmus et al., *supra* note 17.
[154] *See* Woodworth & Kadane, *supra* note 82, at 238, 239 (advocating "a standardized data base" and that "prediction models will be developed and updated via logistic regression").

Rosado[155] would be much less likely. Reporting such error provides a critical link between risk assessment instruments and commitment decisions. In short, a logistic regression-based approach holds more promise than traditional actuarial methods.[156]

Although this Article focused on the Static-99 and sex offender civil commitment, the lessons apply more generally to other actuarial instruments used in other contexts. There are many such contexts. Take for example the following list of criminal applications:

> From the use of the IRS Discriminant Index Function to predict potential tax evasion and identify which tax returns to audit, to the use of drug-courier and racial profiles to identify suspects to search at airports, on the highways, and on city streets, to the use of risk-assessment instruments to determine pretrial detention, length of criminal sentences, prison classification, and parole eligibility, prediction instruments increasingly determine individual outcomes in our policing, law enforcement, and punishment practices.[157]

The IRS formula is apparently based on regression analysis.[158] In contrast, drug-courier profiles have never been empirically validated (at least as of 1985).[159]

The most commonly used risk assessment instrument in this country is the Level of Services Inventory Revised (LSI-R).[160] The LSI-R, used for parole and other purposes, is more extensive but closely analogous in structure to the Static-99.[161] The LSI-R includes a dummy based on age at first arrest, but not age at release.[162] If the goal is predicting recidivism, failing

[155] State v. Rosado, 889 N.Y.S.2d 369 (2009).
[156] *But see* Harris & Rice, *supra* note 46, at 1639 ("regression weights are unstable on replication").
[157] HARCOURT, *supra* note 16, at 2.
[158] Bernard E. Harcourt, *From The Ne'er-Do-Well To The Criminal History Category: The Refinement Of The Actuarial Model In Criminal Law*, 66 LAW & CONTEMP. PROBS. 99, 147 n.204 (Summer 2003).
[159] Morgan Cloud, *Search and Seizure by the Numbers: The Drug Courier Profile and Judicial Review of Investigative Formulas*, 65 B.U. L. REV. 843, 845 (1985).
[160] TRACY W. PETERS & ROGER K. WARREN, NAT'L CTR. FOR STATE COURTS, GETTING SMARTER ABOUT SENTENCING: NCSC'S SENTENCING REFORM SURVEY 17 (2006), at sentencing.nj.gov/downloads/pdf/articles/2006/Aug2006/document09.pdf (visited 11/19/10).
[161] JAMES AUSTIN ET AL., RELIABILITY AND VALIDITY STUDY OF THE LSI-R RISK ASSESSMENT INSTRUMENT: FINAL REPORT (2003), at www.portal.state.pa.us/portal/server.pt/document/.../lsi_r_final_report_pdf (visited 11/19/10).
[162] *Id.* at 14 tbl.7.

to include both as continuous variables is a mistake.[163] Failing to weight the items using regression analysis is another defect shared by the LSI-R. And, finally, reporting logit predictions along with errors could lead to better decision-making than, as the LSI-R does, merely lumping individuals into "low," "medium," and "high" risk groups.[164] The LSI-R demonstrates that the present examination of the Static-99 has potentially broad importance.

Conclusion

The Static-99 is the most thoroughly researched tool for predicting sexual recidivism.[165] Almost no one before this Article, however, empirically assessed the most critical question: can it predict recidivism well enough to meet the legal standard for sex offender commitment?[166] The answer is mixed and qualified, but largely negative. The limitations of this study preclude any strong conclusions, but my findings at least suggest that the goals and methods of sex offender civil commitment need to be reevaluated. In the meantime, this Article identifies several ways in which the Static-99 and like instruments are deficient and can and should be improved.

[163] Indeed, a validation study of the LSI-R found "arrested under age 16" to be significant in predicting recidivism. *Id.* at 18.

[164] A recent examination of the LSI-R found recidivism predictive power (AUC = 0.66 and 0.73) comparable to that achieved by this Article's main logit model (AUC = 0.66; Table 5 *supra*). Sarah M. Manchak, Jennifer Lynne Skeem, & Kevin S. Douglas, *Utility of the Revised Level of Service Inventory (LSI-R) in Predicting Recidivism after Long-Term Incarceration*, 32 L. & HUMAN BEHAV. 477, 482 (2008).

[165] *See, e.g.*, Hanson & Morton-Bourgon, *supra* note 146, at 17 tbl.A1 (listing 63 such studies).

[166] Again, Hart, Michie, & Cooke, *supra* note 9, and Janus & Meehl, *supra* note 15, come closest.

APPENDIX

Formally, the logit model is specified as follows:

Equation 1. $P_i = \frac{1}{1+e^{-\beta X_i}}$

where P_i is the probability of an individual hit or miss, *e* is the base of natural logarithms, *ß* is a matrix of coefficients, and X_i a matrix of individual-specific variable values.[167]

Two post-estimation calculations are complex enough to require explanation. In Table 7, I estimate the number of individuals who met the legal standards for commitment—for example, whose estimated likelihood of recidivism was above 50% ("more likely than not") with 75% confidence (CCE). This required calculating the lower CI for logit predictions, P_i. I substituted for $ßX_i$ in Equation 1 one side of the following formula:

Equation 2. $LB_p = LP_i - Z_p \times SE_{LP_i}$

where LB_p is the lower-bound of the linear CI for a given proof standard (75% or 90%), LP_i is the linear prediction of the logit model for an individual, Z_p is the inverse cumulative standard normal distribution for either 75% or 90%, and SE_{LPi} is the standard error of LP_i.[168]

The Cohen's *d* statistic is defined as $(M_1 - M_2)/S_w$, where M_1 is the mean of one group, M_2 is the mean of the comparison group, and S_w is the pooled-within standard deviation.[169] The complicated part of this equation is the last term:

Equation 3. $S_w = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$

[167] ROBERT S. PINDYCK & DANIEL L. RUBINFELD, ECONOMETRIC MODELS AND ECONOMIC FORECASTS 258 (3d ed. 1991).

[168] *See* Mark Inlow, Prediction Confidence Intervals After Logistic Regression (Apr. 1999, rev. July 2007), at http://www.stata.com/support/faqs/stat/prep.html (visited 11/10/2010). By deriving the confidence interval from the standard error, this methodology avoids one criticism leveled against Hart, Michie, & Cooke, *supra* note 9. *See* Harris, Rice, & Quinsey, *supra* note 75, at 154 ("The appropriate statistic is standard error of measurement . . . .").

[169] http://en.wikipedia.org/wiki/Effect_size (visited 11/5/2010).

Where *n* is group size and *s* is group standard deviation.[170] The CIs on Cohen's *d* statistics were calculated with the METAN downloadable add-on to Stata.[171] All computations in this Article were performed using Stata/SE 11.1.

[170] *Id.*
[171] http://ideas.repec.org/c/boc/bocode/s456798.html (visited 11/5/2010).